Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Reconocimiento de objetos y análisis del desempeño de conductores a partir del comportamiento y seguimiento de movimientos oculares

Tesis de Licenciatura en Ciencias de Datos

Jhosept Levyt Sanchez Goicochea

Director: Juan Kamienkowski
Codirector: Joaquin Gonzalez
Buenos Aires, 2025

RESUMEN

Los accidentes de tránsito son un problema que nos afecta a todos en nuestra vida diaria, de distintas maneras, algunas más personales, porque los sufrimos nosotros mismos o algún ser cercano o de formas más distantes pero igualmente presentes, porque los observamos en la calle o a través de algún medio de comunicación. Con el objetivo de encontrar estrategias más eficientes de exploración visual que permitan detectar con mayor rapidez y precisión situaciones potencialmente peligrosas en contextos de conducción, así como identificar aquellas estrategias menos eficaces, este trabajo busca aportar conocimientos aplicables a la formación y evaluación de nuevos conductores. Esto permitiría reducir la cantidad de siniestros viales, en especial aquellos protagonizados por conductores novatos, quienes representan un alto porcentaje de las víctimas en accidentes de tráfico. Por ello, se decidió respaldar y extender los resultados obtenidos en estudios previos, tales como los de Robbins [1] y Crundall [2].

Para esto, este trabajo tiene como objetivo, a partir de los videos experimentales que se muestran actualmente en el examen de conducción, medir el comportamiento visual de 24 participantes, con el fin de encontrar correlaciones entre dicho comportamiento ocular y su desempeño al detectar peligros en los videos. Con este propósito, el trabajo se estructuró en un enfoque progresivo compuesto por tres etapas, cada uno dependiente del éxito del anterior.

Se exploraron distintos modelos de detección de objetos para identificar peatones, vehículos y otros elementos relevantes en los videos del examen de conducción del Reino Unido, y se eligió *YOLOv12X* [3] por su rendimiento.

También se incorporó el módulo de seguimiento *ByteTrack* [4], que asigna identidades a los objetos detectados a lo largo del video. Esto permite realizar un seguimiento de la trayectoria de cada objeto a lo largo del video.

Por último, se integraron y procesaron los datos recolectados por el *eye-tracker*, y se realizó un análisis para investigar cómo las fijaciones y la amplitud de la búsqueda visual pueden explicar diferencias en el tiempo de reacción ante eventos críticos.

A partir del análisis realizado pudimos detectar la existencia de una fuerte correlación entre el tiempo de reacción y el tiempo que tardaban en fijar los peligros, lo que remarca la importancia de realizar una exploración visual eficiente para una detección temprana de estas situaciones. Asimismo, se observó que una mayor amplitud de búsqueda horizontal, combinada con una menor amplitud de búsqueda vertical, se asocia con tiempos de reacción ligeramente menores. En cambio, la duración y la cantidad de fijaciones no mostraron un efecto estadísticamente significativo en el tiempo de reacción.

A partir de estos resultados, se propone incorporar una instancia de evaluación de reacción, similar a la implementada en el Reino Unido, en el examen para la obtención de la licencia de conducir. Por otro lado, los efectos observados de la amplitud de búsqueda visual ofrecen pautas útiles para optimizar la enseñanza de la conducción, promoviendo estrategias visuales más eficaces.

Palabras clave: “Detección de objetos”, “comportamiento ocular”, “tiempo de reacción”, “seguridad vial”, “estrategia de búsqueda visual”

ABSTRACT

Traffic accidents are a problem that affects all of us in our daily lives in different ways. Sometimes the impact is more personal, when we or someone close to us is directly involved, and other times it is more distant but still present, as we witness these events on the street or through the media.

With the goal of identifying more efficient visual search strategies that enable faster and more accurate detection of potentially dangerous situations in driving contexts, as well as recognizing less effective strategies, this work seeks to contribute knowledge that can be applied to the training and evaluation of new drivers. This could help reduce the number of traffic accidents, especially those involving novice drivers, who represent a high percentage of traffic accident victims. For this reason, we decided to support and extend the findings of previous studies, such as those by Robbins [1] and Crundall [2].

To achieve this, the present work analyzes the visual behavior of 24 participants using experimental videos currently shown in the UK driving test, with the aim of finding correlations between their eye movements and their performance in hazard detection.

The study followed a progressive approach consisting of three stages, each dependent on the success of the previous one.

Different object detection models were explored to identify pedestrians, vehicles, and other relevant elements in the UK driving test videos. The model *YOLOv12X* [3] was selected for its performance.

The *ByteTrack* [4] tracking module was also incorporated. It assigns identities to the detected objects throughout the video, allowing each object's trajectory to be tracked over time.

Finally, the data collected by the *eye tracker* were integrated and processed, and an analysis was carried out to investigate how fixations and the extent of visual search may explain differences in reaction times to critical events.

The analysis revealed a strong correlation between reaction time and the time it took participants to fixate on hazards, emphasizing the importance of efficient visual search for early detection of these situations. Additionally, greater horizontal search amplitude combined with reduced vertical amplitude was associated with slightly shorter reaction times. In contrast, the duration and number of fixations did not show a statistically significant effect on reaction time.

Based on these results, we propose including a reaction evaluation component in the driving license exam, similar to the one implemented in the United Kingdom. Moreover, the observed effects of visual search amplitude offer valuable guidance for improving driver education by promoting more effective visual strategies.

Keywords: "Object detection", "eye movement behavior", "reaction time", "road safety", "visual search strategy"

AGRADECIMIENTOS

Primero que nada, quiero expresar mi más sincero agradecimiento a mi padre y a mi madre por apoyarme siempre en mis decisiones, por escuchar mis problemas y dudas, y por aconsejarme de la mejor manera posible. Gracias a ellos pude crecer, aprender y también equivocarme, como cuando elegí estudiar Medicina y luego decidí cambiar de carrera. Sin su confianza y comprensión, no habría podido llegar hasta donde estoy hoy.

También, quiero agradecer a mis amigos de la facultad, por acompañarme en tantas jornadas de estudio, almuerzos y momentos inolvidables. La carrera sin ustedes no hubiese estado tan llena de risas y buenos recuerdos. Gracias por acompañarme en este camino, a todo el grupo de *Mariposas al Día*.

Finalmente, quiero expresar mi profunda gratitud a mi director, Juan Kamienkowski, y a mi codirector, Joaquín González, por su acompañamiento, dedicación y guía a lo largo de este trabajo. También agradezco a la universidad pública, que me brindó la posibilidad de formarme y crecer académicamente.

Índice general

1. INTRODUCCIÓN

1.1. Motivación

Los accidentes de tránsito constituyen una problemática cotidiana que afecta significativamente nuestra vida y la de todas las personas. Ya sea porque los experimentemos en primera persona, los presenciemos en la vía pública o los conozcamos a través de los medios de comunicación, su presencia es constante y sus consecuencias son muy graves y muchas veces irreversibles. Sin embargo, a pesar de que todos conozcamos la existencia de esta problemática, no siempre somos plenamente conscientes de la magnitud del problema.

Para dimensionar esta realidad vamos a analizar el reporte de siniestralidad vial de la Dirección Nacional de Seguridad Vial [5]. Para entender correctamente es necesario comenzar por definir algunos conceptos clave que se tratan en el mismo. Un **siniestro vial** es cualquier hecho que ocurre en la vía pública (calles, rutas, autopistas, etc.) en el que interviene al menos un vehículo en movimiento y que produce daños a personas o bienes materiales. Por otro lado, una **víctima fatal** es aquella persona que fallece como consecuencia directa de un siniestro vial, dentro de los 30 días posteriores al incidente.

Considerando estas definiciones, a continuación se presenta la evolución histórica de víctimas fatales en siniestros viales en Argentina entre los años 2008 y 2023.

Fig. 1.1: Víctimas fatales en Argentina (2008–2023).
Fuente: Dirección Nacional de Seguridad Vial [5]

Tomemos como caso de análisis el año 2022 pues es ultimo año del que se tienen los números confirmados con certeza. En este año se registraron 4,711 víctimas fatales a causa de siniestros viales. Este número, aunque elevado por si mismo, cobra mayor relevancia cuando se analiza en términos relativos: representa una tasa de 10,2 muertes cada 100.000 habitantes.

Ese mismo año, el total de fallecimientos en Argentina llego a 397.115 personas, esto representa una tasa bruta de mortalidad de 860 fallecimientos por cada 100.000 habitantes. En este contexto, las muertes por siniestros viales representaron aproximadamente el 1,2 % del total de muertes, una proporción muy alarmante si consideramos que se trata de fallecimientos, en su mayoría, evitables.

Esta situación preocupante no se limita unicamente a la realidad argentina sino que se trata de un problema global. Debido a esto, se realizan diversas investigaciones en todo el mundo que buscan identificar patrones, características y tendencias en el comportamiento de los conductores que aumenten la probabilidad de ocurrencia de estos siniestros viales. Comprender estos factores resulta fundamental para el desarrollo de estrategias que contribuyan a reducir la siniestralidad y, en consecuencia, salvar vidas.

1.2. Estudios Previos

Antes de revisar los hallazgos principales en la literatura, es necesario definir varios de los conceptos comúnmente usados en estos estudios, con el objetivo de asegurar una comprensión adecuada y evitar ambigüedades que puedan surgir. En particular hay cinco definiciones que resultan muy relevantes: *peligros*, *percepción de peligro*, *fijaciones*, *sacadas* y *comportamiento ocular*.

- **Peligro**: se refiere a cualquier evento o situación en la vía con alta probabilidad de generar un siniestro vial. Ejemplos típicos incluyen un peatón cruzando inesperadamente, un vehículo que frena de manera abrupta, o un ciclista que irrumpe en la trayectoria del conductor.
- **Percepción de peligro**: es el proceso mediante el cual un conductor detecta, evalúa y reacciona frente a una situación potencialmente riesgosa. Este concepto es central en los estudios sobre seguridad vial, ya que constituye un indicador clave del desempeño y de la experiencia del conductor.
- **Fijación**: designa el intervalo durante el cual la mirada de una persona permanece relativamente estable y enfocada sobre un punto del entorno. Las fijaciones son fundamentales para el procesamiento visual y permiten inferir los elementos del entorno a los que un conductor presta atención.
- **Sacada**: es el movimiento rápido que realiza el ojo entre dos fijaciones. Durante las sacadas no se obtiene información visual significativa, pero su análisis proporciona datos importantes sobre la estrategia exploratoria y el comportamiento visual del conductor.
- **Comportamiento ocular**: hace referencia al conjunto de movimientos y patrones visuales que realiza una persona durante la observación de su entorno. Incluye fijaciones, sacadas y parpadeos. El análisis del comportamiento ocular permite inferir procesos cognitivos subyacentes, como la atención, la toma de decisiones y la carga mental.

Ahora definamos cuáles son los principales hallazgos de investigaciones previas relacionadas con el comportamiento ocular, ya que estos constituyen la base para las hipótesis que serán evaluadas en este trabajo. En particular, se tomarán como referencia dos estudios clave: el trabajo de Crundall et al. [2], que analiza en profundidad la percepción de peligros durante la conducción, y el estudio de Robbins et al. [1], que realiza una revisión sistemática y un meta análisis sobre las diferencias en la búsqueda visual entre conductores novatos y experimentados.

Comencemos analizando los resultados de Crundall et al. [2], una del las características principales de este estudio es la distinción entre peligro y precursor, un precursor es un

elemento que actúa como una señal contextual o visual que sugiere la posible aparición de un peligro. Por ejemplo, un peatón detenido en la vereda constituye un precursor, ya que podría ingresar a la calzada y convertirse en un peligro efectivo.

En dicho experimento participaron 49 conductores divididos en tres grupos: 14 aprendices, con un promedio de 7,5 meses de experiencia, 17 conductores experimentados, con 16,4 años de experiencia y una edad promedio de 33 años y 18 instructores, con 30 años de experiencia de conducción y 13 años como instructores. Todos los instructores estaban certificados oficialmente por el Reino Unido, habiendo aprobado el examen de la Driving Standards Agency.

En esta tesis nos centraremos en el análisis del comportamiento ocular ante los peligros, por lo que de este trabajo previo tomaremos exclusivamente los resultados vinculados a la atención visual frente a estos eventos.

Entre los hallazgos más relevantes se encuentra el siguiente:

- Los peligros fueron fijados visualmente con mayor frecuencia que los precursores, un 86 % para los peligros frente a un 61 % para los precursores. Este resultado sugiere que los conductores tienden a prestar más atención a los peligros una vez que estos se manifiestan de manera explícita, en lugar de anticiparlos a partir de señales contextuales previas.

- Los instructores y los conductores con experiencia miraron más peligros que los aprendices pero no hubo diferencias entre instructores y conductores experimentados, lo que indica que la experiencia mejora la velocidad de detección visual de peligros o señales al menos cuando comparamos a un conductor experimentado contra un novato.

- Los conductores experimentados e instructores fueron más rápidos para fijar la mirada en los estímulos críticos que los aprendices, lo que indica que la experiencia mejora la velocidad de detección visual de peligros o señales.

Por otro lado, en el estudio de Robbins et al. [1] se analizan diferentes medidas de búsqueda visual entre conductores novatos y experimentados. Entre los principales hallazgos se destacan los siguientes:

- **Amplitud de búsqueda horizontal:** Se observó una diferencia clara entre grupos, con los conductores novatos mostrando un rango más estrecho de exploración visual en comparación con los experimentados. Esta diferencia se mantuvo incluso al excluir grupos con niveles extremos de experiencia. Sin embargo, al segmentar los estudios por tipo de entorno, la diferencia fue significativa únicamente en metodologías inmersivas (simuladores avanzados o conducción real), y no en metodologías simples como la visualización de videos, la cual es el caso de estudio de este trabajo.

- **Amplitud de búsqueda vertical:** No se encontraron diferencias significativas entre conductores novatos y experimentados. Este resultado sugiere que la amplitud de búsqueda vertical no es una medida lo suficientemente sensible para distinguir niveles de experiencia, posiblemente debido a su menor relevancia en la detección de peligros.

- **Número de fijaciones:** No se observó una diferencia general entre grupos respecto a la cantidad de fijaciones realizadas. Este resultado se mantuvo incluso al analizar cada tipo de metodología utilizada por separado.

- **Duración media de las fijaciones:** En ambos tipos de metodologías de estudio no se hallaron diferencias significativas entre los distintos grupos de experiencia, aunque en contextos inmersivos por si solos sí se encontró una diferencia significativa, indicando que los conductores novatos tienden a mantener la vista fija por más tiempo que los experimentados. Esto sugiere que los entornos inmersivos pueden ser más adecuados para detectar diferencias sutiles en la atención visual vinculadas a la experiencia.

Estos estudios han buscado establecer correlaciones entre el nivel de experiencia del conductor y métricas asociadas al comportamiento ocular, tales como la amplitud de búsqueda visual horizontal y vertical, la duración promedio de las fijaciones y la cantidad total de fijaciones realizadas por participante. La razón detrás de este enfoque radica en que los conductores jóvenes y por lo tanto con menor experiencia están desproporcionadamente representados entre las víctimas fatales en siniestros viales.

Este fenómeno puede observarse claramente en el caso de Argentina. Tal como se muestra en la Figura 1.2, los grupos formados por las personas de 15 a 24 años y de 25 a 34 años representan, cada uno, un 21 % del total de víctimas viales registradas en 2022. En conjunto, estos dos grupos concentran el 42 % del total, lo que evidencia la vulnerabilidad de los conductores jóvenes y refuerza la necesidad de investigar cómo se manifiesta su comportamiento visual durante la conducción.

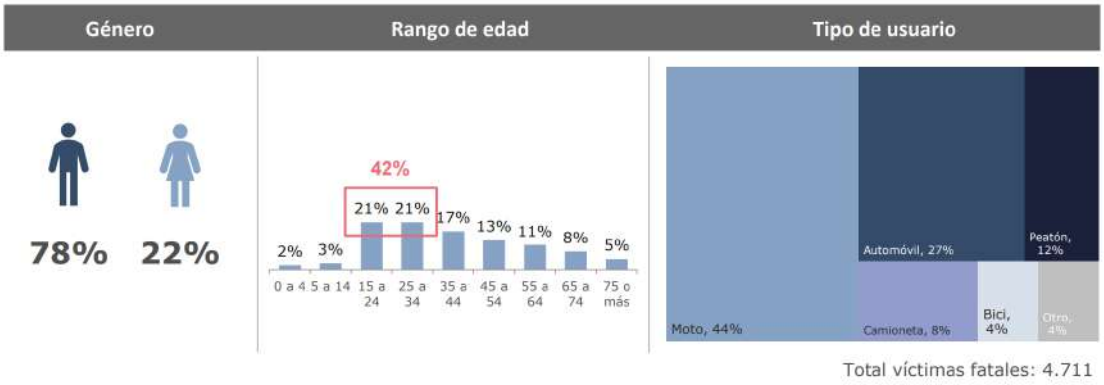

Fig. 1.2: Porcentaje de víctimas viales por edad, género y tipo de usuario en el año 2022.
Fuente: Dirección Nacional de Seguridad Vial [5]

1.3. Objetivos

A partir de los resultados reportados en las investigaciones previas, este trabajo se propuso avanzar en la comprensión de la relación entre el comportamiento visual y la capacidad para detectar y reaccionar ante situaciones de peligro durante la conducción. Mientras que estudios anteriores han explorado la relación entre la experiencia de los participantes y sus patrones de búsqueda visual, en esta tesis nos enfocamos en analizar directamente la relación entre las características del comportamiento ocular y el desempeño real frente a peligros concretos.

Para ello, se diseñó un experimento en el que los participantes observaron una serie de videos extraídos del examen de percepción de peligros del sistema de licencias de conducir del Reino Unido. Durante la visualización, se registraron sus comportamientos oculares

mediante un *eye tracker*. Ante la aparición de un peligro que requiriera una acción (como frenar, esquivar o reducir la velocidad), los participantes debían presionar un botón para indicar que realizaron alguna de las acciones.

El tiempo de respuesta ante cada peligro se utilizó como medida principal de desempeño, permitiendo así evaluar de manera cuantitativa la eficiencia con la que cada participante percibía y procesaba la situación. El objetivo general fue, entonces, estudiar cómo distintos indicadores del comportamiento visual (como la duración de las fijaciones, la amplitud de búsqueda o el tiempo hasta la primera fijación del peligro) se asocian con la capacidad de reacción frente a situaciones de peligro reales en contextos de conducción.

Para alcanzar este propósito, se desarrolló un enfoque progresivo compuesto por tres objetivos específicos, cada uno de los cuales abordó una etapa clave del análisis. La estructura de estos objetivos permitió avanzar de manera ordenada y sistemática.

1. **Reconocimiento de objetos en video.** El objetivo es identificar de forma automática los objetos relevantes, como peatones, vehículos y señales de tránsito, en los videos del examen de manejo del Reino Unido. Esto permite contar con la localización espacial y temporal de los elementos presentes en escena, posibilitando la incorporación de los tiempos de fijación de los peligros al análisis del comportamiento visual de los conductores frente a situaciones potencialmente peligrosas.

2. **Incorporación de continuidad en el seguimiento de objetos.** La mayoría de los modelos de detección de objetos operan de forma cuadro a cuadro, sin considerar la continuidad temporal de los objetos en la escena. Esta limitación impide una interpretación coherente del movimiento de los elementos del video y dificulta la estimación precisa del tiempo de fijación sobre un objeto especifico. Por ello, el objetivo es integrar un mecanismo de seguimiento que permita mantener la identidad de los objetos detectados a lo largo del tiempo en los videos. Esto posibilita una interpretación coherente del movimiento de los objetos en la escena y una estimación más precisa del tiempo de fijación visual sobre ellos, incluso en presencia de encubrimientos parciales o fallos temporales en la detección.

3. **Análisis del comportamiento ocular en relación con la percepción del peligro.** El objetivo es estudiar cómo indicadores del comportamiento visual, tales como la duración de las fijaciones, la amplitud de la búsqueda visual y el tiempo hasta la primera fijación en un peligro, se relacionan con la capacidad para detectar y reaccionar correctamente ante situaciones de peligro en contextos de conducción. Este análisis se propuso responder preguntas clave como:

   - ¿Cuál es la relación entre el tiempo de fijación del peligro y la velocidad de reacción?
   - ¿Qué patrones de exploración visual se asocian con un mejor desempeño?
   - ¿Cuáles son las variables oculares más predictivas del tiempo de reacción?

1.4. Estructura de la tesis

Con el fin de alcanzar los objetivos planteados, esta tesis se estructura en tres capítulos principales, cada uno orientado a abordar uno de los objetivos definidos.

- **Reconocimiento de objetos en video:** este objetivo se desarrolla en el capítulo de **Metodología**, donde se realiza un análisis comparativo (benchmark) de distintos modelos de detección de objetos en 2D. Se evaluó el desempeño de estos modelos sobre conjuntos de datos específicos para entornos de conducción, con el fin de seleccionar el detector más adecuado para la tarea. También se integraron y procesaron los datos recolectados por el eye-tracker para dejarlos listos para el análisis posterior.

- **Incorporación de continuidad en el seguimiento de objetos:** este objetivo se desarrolla en el capítulo de **Trackeo de identidad de objetos**, donde se implementa un sistema de seguimiento basado en el módulo *ByteTrack* [4] para mantener la identidad de los objetos detectados a lo largo de los cuadros en el video. Adicionalmente, se implementó una estrategia de interpolación de trayectorias para los objetos que desaparecen momentáneamente debido a encubrimientos o fallos en la detección. Esta interpolación se basó en supuestos como trayectorias suaves y predecibles para peatones y vehículos. De esta forma, se logró una reconstrucción más estable de las trayectorias y una medición más precisa del comportamiento visual sobre los objetos relevantes.

- **Análisis del comportamiento ocular en relación con la percepción del peligro:** este objetivo se desarrolla en el capítulo de **Análisis exploratorio y correlacional de los datos**, donde primero se realiza un análisis descriptivo de los tiempos de reacción para explorar su distribución y variabilidad entre participantes y videos. Luego, se lleva a cabo un análisis estadístico orientado a identificar cómo las principales características del comportamiento visual se correlacionan con las diferencias en el tiempo de reacción frente a eventos críticos. Finalmente, se construye un modelo predictivo que busca estimar el tiempo de reacción utilizando únicamente variables del comportamiento ocular, con el fin de analizar cómo interactúan estas variables y en qué medida pueden anticipar la respuesta ante situaciones de peligro.

2. METODOLOGIA

2.1. Descripción del dataset utilizado

El dataset empleado en este trabajo proviene de un experimento basado en videos del examen de manejo del Reino Unido. En total, se utilizaron **33 videos** generados mediante *CGI* (Computer-Generated Imagery), los cuales simulan entornos de conducción. Cada video incluye una variedad de objetos típicos del entorno vial, tales como *personas, vehículos, motocicletas, bicicletas, perros, caballos* y otros elementos como señales de vialidad, con el objetivo de representar situaciones de manejo realistas.

Los videos fueron diseñados para contener **situaciones cotidianas de conducción**, incorporando en cada uno **una o dos situaciones de peligro**. Estas situaciones peligrosas pueden incluir, por ejemplo, *un perro que cruza inesperadamente la calle, un vehículo que cambia de carril, o una persona que cruza sin mirar*. Es importante destacar que cada evento peligroso cuenta con una **marca temporal precisa** que indica el *inicio del peligro*: este no necesariamente coincide con la aparición del objeto en pantalla, sino que comienza en el momento en que el conductor debería percibir la necesidad de reaccionar ante la situación.

El experimento fue llevado a cabo con la participación de **26 sujetos**, a quienes se les mostraron todos los videos mientras se registraba su **comportamiento ocular** mediante un sistema de *eye-tracking*. Finalmente, se logró procesar la información completa de **24 sujetos**. Este sistema recolectó datos cada milisegundo, capturando:

- La **posición de la mirada** en la pantalla (coordenadas X e Y).
- El **estado de fijación visual**, indicando si el participante estaba fijando la mirada en un punto o realizando una sacada.

Además, los participantes contaron con un **botón de respuesta**, que podían presionar libremente cuando consideraran necesario *frenar, reducir la velocidad u otro tipo de reacción* ante un posible peligro. No se les indicó cuándo ocurrirían los eventos peligrosos ni se les impuso un límite en la cantidad de respuestas que podían emitir.

Este conjunto de datos ofrece una rica fuente de información tanto visual como conductual, permitiendo analizar la relación entre la atención visual de los conductores y su capacidad de respuesta frente a peligros en entornos de conducción simulada.

Fig. 2.1: Fotogramas de los videos

2.2. Selección de modelo

Como primer paso para realizar el análisis, fue necesario ubicar espacial y temporalmente los peligros que aparecían en los videos mostrados durante el experimento. Dado que entrenar un modelo de detección 2D de objetos desde cero resultaba muy costoso computacionalmente y requiere múltiples iteraciones de prueba y error, se optó en primera instancia por investigar cuáles son los modelos de detección de vanguardia en la actualidad. Como los peligros pueden corresponder a objetos muy diversos como perros, bicicletas, autos, camiones, caballos, personas, entre otros, se decidió buscar los modelos más adecuados en un entorno más amplio, como el que proporciona el dataset COCO [6].

2.2.1. COCO dataset

COCO (Common Objects in Context) es un dataset ampliamente utilizado en visión por computadora, compuesto por aproximadamente 330.000 imágenes, de las cuales unas 200.000 están etiquetadas para tareas como detección de objetos, segmentación de instancias, segmentación semántica, anotaciones de poses humanas, y descripción de imágenes. Este dataset incluye un total de 80 categorías, entre las que se encuentran muchos de los objetos relevantes para nuestro análisis como autos, bicicletas, peatones, camiones y animales, así como otras categorías, tales como paraguas, bolsos de mano o equipamiento deportivo. Dado que en este trabajo solo se aborda la tarea de detección de objetos, se emplearon exclusivamente las anotaciones correspondientes a esa tarea.

2.2.2. Métrica elegida: mAP@0.5:0.95

Dado que lo que vamos a realizar es una detección 2D de múltiples objetos, es fundamental seleccionar adecuadamente la métrica que permita evaluar el desempeño de los modelos de manera objetiva. La métrica más aceptada en la literatura de visión por computadora es la *mean Average Precision* **mAP@0.5:0.95**. Para definirla correctamente, es necesario introducir previamente algunos conceptos.

- **Verdaderos Positivos (TP)**: casos en los que el modelo predice correctamente la clase positiva.
- **Falsos Positivos (FP)**: casos en los que el modelo predice la clase positiva, pero en realidad pertenece a la clase negativa.
- **Falsos Negativos (FN)**: casos en los que el modelo predice la clase negativa, pero en realidad pertenece a la clase positiva.
- **Verdaderos Negativos (TN)**: casos en los que el modelo predice correctamente la clase negativa.

Con estas definiciones, se pueden expresar las métricas de *Recall* y *Precisión* de la siguiente forma:

- **Recall**: en el contexto de detección de objetos en 2D, mide cuántos de los objetos reales presentes en la imagen fueron identificados correctamente por el modelo. Un valor alto de *Recall* indica que el modelo detecta la mayoría de los objetos relevantes (es decir, pocos falsos negativos).

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Precisión**: en el contexto de detección de objetos en imágenes 2D, mide cuántos de los objetos identificados por el modelo eran realmente correctos. Una precisión alta implica que se realizaron pocas detecciones falsas positivas.

$$\text{Precisión} = \frac{TP}{TP + FP}$$

A partir de estas métricas se construyen la *Average Precision* (AP) y *mean Average Precision* (mAP):

- **Average Precision (AP)**: se calcula a partir de la curva de *Precisión-Recall* obtenida para cada clase. El AP representa el área bajo dicha curva, lo que proporciona un valor único que resume la precisión del modelo en la detección de una clase específica, considerando todos los posibles umbrales de confianza. Un valor alto de AP indica que el modelo mantiene tanto una alta precisión como un buen recall a lo largo de distintos umbrales.
- **mean Average Precision (mAP)**: es el promedio de las métricas AP obtenidas para todas las clases presentes en el conjunto de datos. Esta métrica permite evaluar el rendimiento global del modelo en tareas de detección de objetos.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{2.1}$$

  Donde:

  - N es el número total de clases.
  - AP_i es la *Average Precision* correspondiente a la clase i.

Ahora bien, la mAP tiene en cuenta la confianza en la clasificación de las detecciones, pero ¿cómo controlamos que la ubicación espacial de los objetos sea también correcta? Para ello, se utilizan variantes de la métrica mAP que incorporan el umbral de superposición espacial mediante el IoU (Intersection over Union).

Intersection over Union (IoU): es una métrica utilizada para evaluar la precisión espacial de las detecciones realizadas por un modelo. Se define como la división entre el área de superposición y el área de unión entre la caja delimitadora predicha por el modelo y la caja delimitadora real.

$$\text{IoU} = \frac{\text{Área de Intersección}}{\text{Área de Unión}}$$

Un valor de IoU igual a 1 indica una coincidencia perfecta entre la predicción y la anotación real, mientras que un valor cercano a 0 indica una superposición mínima.

Con todo lo anterior aclarado, definiremos ahora los conceptos de **mAP@0.5** y **mAP@0.5:0.95**:

- **mAP@0.5**: se calcula utilizando un umbral fijo de IoU igual a 0,5. Esto significa que una detección se considera correcta únicamente si la superposición entre la caja predicha y la caja real alcanza al menos un 50 %. Esta métrica evalúa el desempeño del modelo bajo un criterio de localización moderadamente estricto.
- **mAP@0.5:0.95**: se calcula promediando el *Average Precision* obtenido en múltiples umbrales de IoU, que van desde 0,5 hasta 0,95 en incrementos de 0,05. Esta métrica proporciona una evaluación más completa y exigente del modelo, considerando distintos niveles de precisión en la localización.

$$\text{mAP@0.5:0.95} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\text{AP}_i(\text{IoU} \geq \tau_t)\right)$$

donde

$$\tau_t \in \{0{,}5, 0{,}55, 0{,}60, \ldots, 0{,}95\} \quad \text{y} \quad T = 10$$

$$N = \text{número de clases}$$

2.2.3. Desempeño de los modelos

En esta sección se analiza el rendimiento de distintos modelos de detección de objetos sobre el dataset COCO, utilizando como métrica principal el mAP@0.5:0.95, descrita anteriormente. Los resultados generales de desempeño se resumen en la Tabla 2.1.

Se consideraron dos familias principales de modelos que representan el estado del arte en detección de objetos:

Por un lado, la línea **YOLO** (*You Only Look Once*), reconocida por su eficiencia y precisión en tiempo real, de Redmon et al. [7].. Dentro de esta familia, se evaluaron las variantes *M*, *L* y *X*, que corresponden a diferentes tamaños disponibles, en orden creciente: M, L y X. La arquitectura original de YOLO se basaba en redes neuronales convolucionales, pero sus últimas versiones integran bloques basados en *Transformers* desde YOLOv12.

Por otro lado, la familia **DETR** (*DEtection TRansformer*), propuesta por Meta, que utiliza un enfoque basado exclusivamente en *Transformers* para la detección de objetos, de Carion et al. [8]. Dentro de esta familia, se evaluaron variantes con distintas arquitecturas de *backbone*, como *ResNet-50 (R50)* y *ResNet-101 (R101)*, que representan redes convolucionales más profundas y potentes para extraer características visuales. Esta eleccion impacta directamente en la capacidad del modelo para detectar objetos con precisión, siendo R101 más robusto, pero también más costoso computacionalmente que R50. Cabe aclarar que, para las implementaciones de RT-DETR hechas por Ultralytics, se utilizan las denominaciones *L* y *X* para indicar nuevamente el tamaño del modelo, siguiendo una convención similar a la usada en YOLO.

Modelo	#Parámetros (M)	GFLOPs	mAP@0.5:0.95
DETR-DC5 R50	41	187	43,3
DETR-DC5 R101	60	253	44,9
Conditional-DETR-DC5 R50	44	195	45,1
Conditional-DETR-DC5 R101	63	262	45,9
YOLOv11M	20,1	68,0	51,5
YOLOv12M	20,2	68,0	52,5
RT-DETR-L	33	108,3	53,0
RT-DETR R50	42	136	53,1
YOLOv11L	25,3	86,9	53,4
YOLOv12L	26,4	86,9	53,7
RT-DETR R101	76	259	54,3
YOLOv11X	56,9	194,9	54,7
RT-DETR-X	67,5	232,7	54,8
YOLOv12X	59,1	194,9	55,2

Tab. 2.1: Comparación del rendimiento de distintos modelos de detección de objetos en COCO. Datos tomados de [3, 8, 9].

Dado que nuestro objetivo consiste en procesar videos de forma asincrónica, sin requerimientos de tiempo real, no es necesario buscar un compromiso entre precisión y velocidad de inferencia. Por lo tanto, se prioriza seleccionar el modelo con mejor rendimiento global en la métrica mAP@0.5:0.95 sin buscar un equilibrio entre el rendimiento con su costo computacional. Teniendo esto en cuenta, y como se observa en la Tabla 2.1, el modelo con mejor desempeño general es *YOLOv12X*, al obtener la mayor puntuación en la métrica mAP@0.5:0.95.

2.2.4. Analisis de performance por clase

Para profundizar el análisis, nos enfocamos en los tres modelos con mejor desempeño general: *YOLOv11X*, *YOLOv12X* y *RT-DETR-X*. Evaluamos su rendimiento sobre un subconjunto de clases relevantes para nuestro caso de estudio, ya que estas aparecen con mayor frecuencia en los videos del experimento.

En la Tabla 2.2 se presenta la métrica de precisión (P), recall (R), mAP@0.5 y mAP@0.5:0.95 para cada clase evaluada, indicando el modelo con mejor resultado para cada una.

Clase	Modelo	Precisión (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
bicycle	YOLOv11X	0,732	0,627	0,685	0,450
bus	YOLOv11X	0,859	0,841	0,906	0,802
car	YOLOv11X	0,771	0,702	0,772	0,555
dog	YOLOv12X	0,867	0,858	0,893	0,785
horse	YOLOv12X	0,864	0,893	0,934	0,764
motorcycle	YOLOv12X	0,798	0,757	0,820	0,593
person	YOLOv12X	0,836	0,766	0,857	0,658
stop sign	YOLOv12X	0,844	0,723	0,807	0,745
traffic light	YOLOv11X	0,750	0,494	0,605	0,347
train	YOLOv12X	0,905	0,905	0,954	0,798
truck	RT-DETR-X	0,717	0,575	0,671	0,503

Tab. 2.2: Rendimiento por clase de los modelos seleccionados.

Como se observa, no hay un único modelo que se destaque en todas las clases. Sin embargo, existe una clara predominancia de *YOLOv11X* y *YOLOv12X* sobre *RT-DETR-X*, que sólo supera a sus competidores en la clase `truck`.

Esto nos lleva a considerar una posible estrategia de combinación de modelos especializados por clase para realizar las predicciones. No obstante, para poder tomar una decisión informada, es necesario analizar la magnitud de las diferencias entre estos modelos. Si las diferencias entre modelos son pequeñas, podría no justificar la complejidad adicional que implica usar un modelo distinto por clase. En cambio, si las diferencias son significativas, una estrategia híbrida podría mejorar sensiblemente la precisión del sistema general de detección.

Clase	RT-DETR-X	YOLOv11X	YOLOv12X
bicycle	-0,040	0,000	-0,003
bus	-0,059	0,000	-0,006
car	-0,039	0,000	-0,003
dog	-0,002	-0,007	0,000
horse	-0,072	-0,039	0,000
motorcycle	-0,029	-0,004	0,000
person	-0,043	-0,001	0,000
stop sign	-0,061	-0,011	0,000
traffic light	-0,019	0,000	-0,007
train	-0,037	-0,015	0,000
truck	0,000	-0,012	-0,003

Tab. 2.3: Diferencia de desempeño por clase entre modelos, respecto al mejor modelo por clase (valor de referencia = 0), medido con mAP@0.5:0.95.

Como podemos observar en la Tabla 2.3, las diferencias de desempeño entre los modelos son, en general, pequeñas, presentando una variación en mAP@0.5:0.95 menor a 0,1. Por ello, se decidió utilizar un único modelo para todas las clases. En particular, se elige *YOLOv12X*, ya que presenta el mejor rendimiento en general y en la mayoría de las clases y es la versión más actual de YOLO.

2.2.5. YOLOv12X en distintos Datasets

Dado que nos interesaba que las clases previamente mencionadas sean correctamente identificadas por el modelo elegido, extendimos el análisis evaluando su rendimiento en datasets distintos al utilizado durante el entrenamiento. Evaluar únicamente en COCO, al ser un dataset tan general, implicaba que las clases de mayor interés no estén debidamente representadas.

Por este motivo, utilizamos dos conjuntos de datos enfocados en entornos de conducción autónoma para realizar un *benchmark* más representativo y cercano a nuestro caso: **BDD100K** [10] y **Udacity Self-Driving Car Dataset** [11]. Para ello, primero identificamos las clases de interés que estuvieran presentes tanto en COCO como en estos nuevos datasets. Luego, realizamos un mapeo de etiquetas desde estos nuevos conjuntos de datos hacia las etiquetas del dataset COCO. Por ejemplo, si en el dataset de COCO la clase «Car» tiene el ID 2, y en otro conjunto de datos se encuentra una clase denominada «automóvil» con el ID 1, se realiza una etapa de preprocesamiento en la que todas las etiquetas «automóvil:1» son reemplazadas por «Car:2». Este procedimiento se repite para todas las clases relevantes, utilizando un mapeo previamente definido que traduce las etiquetas de cada dataset externo a las clases e IDs establecidos por COCO.

Esto nos permitió evaluar el rendimiento de YOLOv12X sin necesidad de reentrenar una nueva capa de clasificación por cada dataset.

- **BDD100K** [10]: Contiene 100,000 videos y cubre 10 tareas distintas, centradas en visión por computadora aplicada a entornos de conducción. Se caracteriza por ofrecer una gran diversidad geográfica, ambiental y climática, lo cual permite entrenar y evaluar modelos más robustos ante distintas condiciones del mundo real.
- **Udacity Self-Driving Car Dataset** [11]: Contiene 15.000 imágenes, fue creado para entrenar y evaluar modelos de conducción autónoma, enfocado principalmente en tareas como detección de carriles, detección de objetos.

Clase	Images	Instances	Box(P)	R	mAP@0.5	mAP@0.5:0.95
all	1960	24327	0,531	0,426	0,428	0,261
person	639	2626	0,702	0,379	0,451	0,221
bicycle	121	221	0,627	0,367	0,408	0,198
car	1956	20327	0,802	0,488	0,624	0,375
bus	227	309	0,311	0,434	0,357	0,278
truck	526	844	0,214	0,464	0,301	0,234

Tab. 2.4: Resultados de evaluación por clase de YOLOv12X en BDD100K.

Clase	Images	Instances	Box(P)	R	mAP@0.5	mAP@0.5:0.95
all	5960	38760	0,452	0,361	0,332	0,155
person	1389	4192	0,715	0,335	0,395	0,166
bicycle	457	701	0,159	0,041	0,046	0,011
car	5142	25628	0,697	0,599	0,646	0,337
truck	984	1421	0,174	0,491	0,223	0,132
traffic light	1956	6818	0,513	0,338	0,351	0,128

Tab. 2.5: Resultados de evaluación YOLOv12X en Udacity

Como podemos observar en las Tablas 2.4 y 2.5, los resultados obtenidos no son tan buenos como los alcanzados en COCO, registrando un promedio de **0,261 mAP@0.5:0.95** en BDD100K y **0,155 mAP@0.5:0.95** en el dataset de Udacity. Esta disminución en el desempeño puede atribuirse a que ambos conjuntos, diseñados específicamente para tareas de conducción autónoma, presentan escenas particularmente caóticas y complejas, con una alta densidad de objetos, condiciones climáticas variables y entornos urbanos muy dinámicos.

Sin embargo, estas condiciones no reflejan fielmente la naturaleza de los videos utilizados en nuestro experimento, los cuales provienen de entornos mucho más controlados. En particular, dichos videos fueron generados mediante CGI (imágenes generadas por computadora), lo que implica una menor cantidad de objetos por escena, menor variabilidad visual y menor complejidad en los propios objetos.

Además, identificamos una diferencia en la manera en que COCO y estos datasets etiquetan ciertas clases, particularmente las relacionadas con vehículos tripulados. Mientras que el primero tiene por separado la *bicicleta* y la *persona* que la conduce, BDD100K y Udacity, agrupan ambos elementos bajo una única clase denominada *cyclist*. Esta diferencia en la definición de clases introduce inconsistencias en la evaluación, lo que se traduce en un rendimiento artificialmente bajo para dichas categorías. Por esta razón, se decidió excluir las clases **bicicleta** y **motocicleta** de los análisis posteriores de desempeño.

Por estos motivos, evaluamos el rendimiento utilizando un conjunto de datos generado con **CARLA** [12], un simulador urbano de conducción que permite crear entornos virtuales controlados mediante CGI, similares a los observados en nuestros videos experimentales.

Clase	Images	Instances	Box(P)	R	mAP@0.5	mAP@0.5:0.95
all	1600	2595	0,643	0,657	0,753	0,5035
person	333	367	0,560	0,973	0,795	0,577
car	638	1040	0,748	0,814	0,852	0,613
traffic light	721	721	0,927	0,782	0,877	0,448
stop sign	467	467	0,339	0,060	0,488	0,376

Tab. 2.6: Performance de YOLOv12X en el Dataset generado por CARLA.

Como puede observarse en la Tabla 2.6, el rendimiento en CARLA fue mucho más satisfactorio y cercano al obtenido en COCO. Por lo tanto, consideramos que el rendimiento del modelo es lo suficientemente bueno como para ser utilizado en las siguientes etapas del trabajo.

2.3. Procesamiento de datos del eye tracker

Una vez procesados todos los videos, comenzamos el procesamiento de los registros de eye-tracking y a la definición de las variables que se utilizarán para el análisis.

En primer lugar, para cada participante y para cada uno de los peligros presentes en los videos del experimento, se consideraron dos medidas temporales:

- **Tiempo de respuesta**: intervalo entre el momento en que un objeto es declarado como un peligro en el video y el instante en que el participante presiona el botón, indicando que ha detectado el peligro y tomaría una acción al respecto (como frenar, reducir la velocidad o girar).
- **Tiempo hasta la fijación**: intervalo entre el momento en que un objeto es declarado como un peligro en el video y el momento en que el participante fija dicho peligro. En los casos en que el objeto ya se encontraba fijado cuando fue declarado peligroso, se tomó como tiempo de inicio el instante inicial de esa fijación, permitiendo por lo tanto tener un tiempo hasta la fijación negativo para estos casos.

Adicionalmente, se procesaron métricas relevantes del comportamiento ocular, como la duración promedio de cada fijación y la cantidad total de fijaciones registradas por video. Para asegurar la validez de las mediciones, se excluyeron del análisis aquellas fijaciones realizadas fuera del área visible de la pantalla, ya que no correspondían al entorno simulado de conducción. Incluir estas fijaciones podría introducir ruido en los resultados y afectar negativamente la interpretación de las relaciones entre las variables oculares y el desempeño en la detección de peligros.

Finalmente, para cuantificar la amplitud de búsqueda, se calculó la desviación estándar de las coordenadas promedio de cada fijación registrada en la pantalla. Esta elección se basa en el hecho de que las distribuciones de las fijaciones, tanto en el eje horizontal como en el vertical, presentan una distribución aproximadamente normal para todos los participantes como se puede observar en las Figuras 2.2 y 2.3. Cabe destacar que las fijaciones fuera del área visible de la pantalla y por lo tanto del video de interés tampoco fueron consideradas en este análisis.

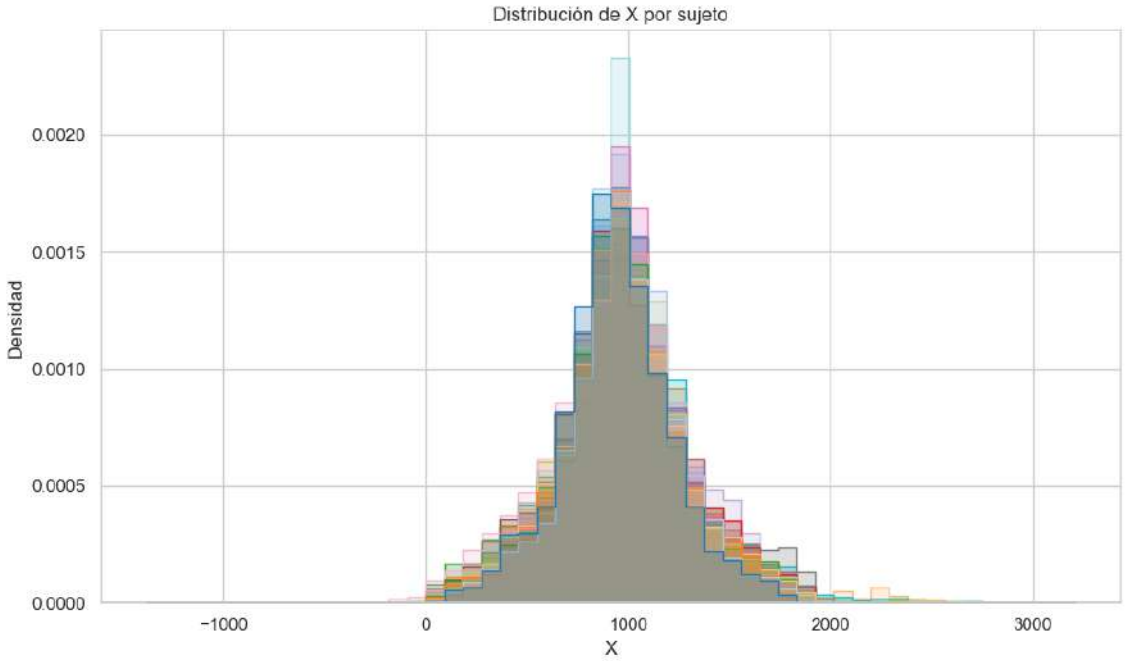

Fig. 2.2: Distribución horizontal de las fijaciones por participante.

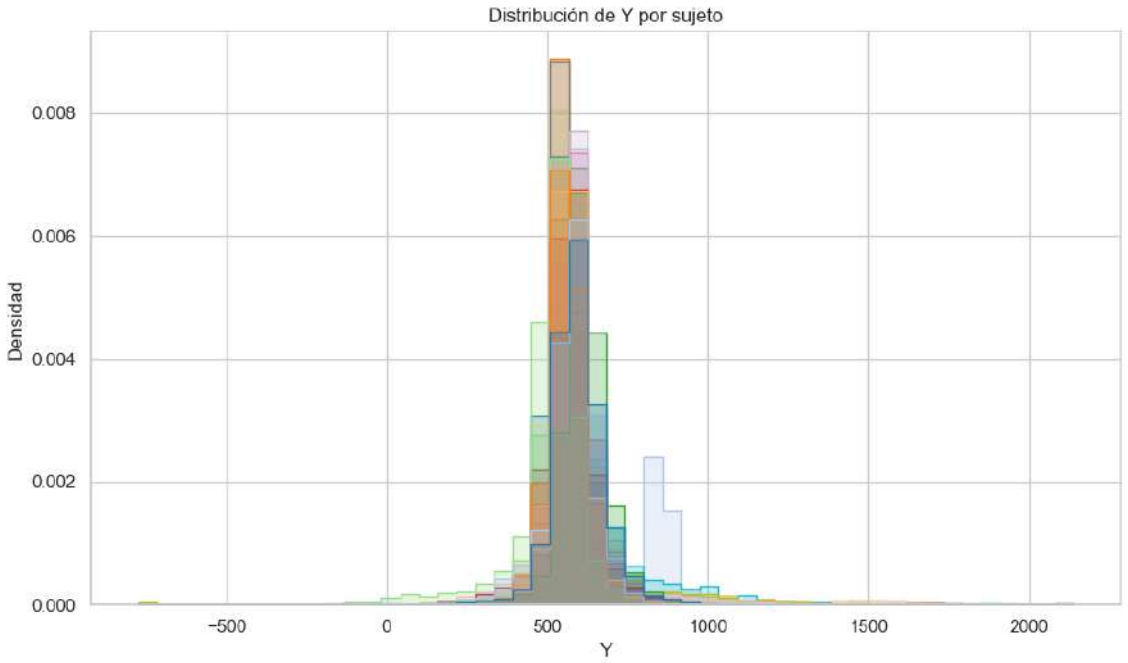

Fig. 2.3: Distribución vertical de las fijaciones por participante.

2.4. Detalles técnicos de librerías usadas

En este trabajo se utilizaron múltiples librerías de Python para realizar el procesamiento de datos y videos, predicciones con modelos, benchmark, etc. A continuación, se

describen brevemente las principales herramientas utilizadas, su propósito dentro del proyecto y las referencias bibliográficas correspondientes.

2.4.1. Detección y seguimiento de objetos en video

Con el objetivo de identificar y realizar un seguimiento de los objetos relevantes en los videos del experimento, se evaluaron distintos modelos ya entrenados de detección y seguimiento que permiten localizar entidades como vehículos, peatones o señales, y mantener la identidad de cada uno de ellos a lo largo del tiempo.

Para realizar la comparación de los modelos se utilizaron las versiones entrenadas disponibles en la biblioteca *Transformers* de HuggingFace y en `Ultralytics` [3]. La implementación del algoritmo de seguimiento `ByteTrack` también fue la provista por `Ultralytics` [3].

2.4.2. Procesamiento y visualización de los videos

Con el objetivo de facilitar el análisis del progreso del experimento, se realizaron distintos procesamientos de video mediante la biblioteca `OpenCV` [13]. Estos incluyeron la superposición de las predicciones generadas por el modelo y las trayectorias interpoladas, sobre los videos originales. Además, se realizó la inversión de la secuencia temporal de los videos, extracción de fotogramas para examinar momentos específicos con mayor detalle, y la generación de nuevos videos integrando todos estos elementos.

Estas herramientas permitieron verificar visualmente la correcta ejecución del procesamiento a lo largo del experimento.

2.4.3. Análisis de correlaciones y pruebas estadísticas

Para estudiar el efecto que tienen de las variables clave del experimento como los tiempos de fijación sobre el peligro, la amplitud de búsqueda visual vertical y horizontal , y las características de las fijaciones sobre el tiempo de reacción, se calcularon las correlaciones entre todos los pares de variables, con el objetivo de identificar qué tan relacionadas estaban entre sí.

Además, con el objetivo de determinar si los datos seguían alguna distribución teórica conocida, se estimaron los parámetros mediante ajustes por máxima verosimilitud y se aplicaron pruebas estadísticas de hipótesis. Esto se hizo con el fin de establecer con qué supuestos se contaba al momento de elegir el tipo de correlación a calcular o los modelos estadísticos a utilizar, considerando estos como una posible vía de explicabilidad del efecto de las variables sobre el tiempo de reacción. Dado que no se cumplían los supuestos de normalidad requeridos para aplicar el coeficiente de correlación de Pearson, se optó por utilizar el coeficiente de correlación de **Spearman**, también conocido como *rho de Spearman* (ρ). Este es una medida no paramétrica que evalúa la fuerza y dirección de la asociación monótona entre dos variables. A diferencia del coeficiente de Pearson, que mide relaciones lineales y asume normalidad en los datos, Spearman se basa en los rangos de los valores en lugar de sus valores absolutos. Esto lo hace especialmente útil cuando los datos no cumplen con los supuestos de normalidad o cuando se sospechan relaciones no lineales pero monótonas como es el caso de nuestro estudio.

El coeficiente de Spearman se define como:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre los rangos de cada observación, y n es el número total de pares de datos. los rangos de cada observación en las dos variables, y n es el número de observaciones. El valor de ρ oscila entre -1 y 1, donde 1 indica una correlación monótona positiva perfecta, -1 una correlación monótona negativa perfecta, y 0 la ausencia de correlación monótona.

Todo este análisis fue realizado utilizando las implementaciones provista por la biblioteca `SciPy` [14].

2.4.4. Importancia de características

Se realizó un análisis de importancia de características con el objetivo de comprender el impacto de cada variable en el modelo de predicción. Para ello, se entrenó un *Random Forest* compuesto por 1000 árboles, utilizando el conjunto total de datos. El entrenamiento se llevó a cabo mediante la implementación provista por la biblioteca `Scikit-learn` [15].

Con el fin de analizar la interpretabilidad del modelo, se empleó el método SHAP (SHapley Additive exPlanations) [16], que permite estimar cómo influye cada variable en las predicciones del modelo, tanto a nivel global como individual.

SHAP proporciona un gráfico resumen en el que se visualiza el impacto de cada variable predictora en las predicciones del modelo. En el eje **X** se representan los **valores SHAP**, estos son los que indican cuánto y en qué dirección contribuye cada variable a la predicción del modelo en cada una de las predicciones, en nuestro caso en particular valores negativos disminuyen el tiempo de respuesta y valores positivos aumentan el tiempo de respuesta. el **color de los puntos** refleja el **valor real de la variable predictora correspondiente** para esa observación: los valores bajos se representan en azul y los valores altos en rojo. Por ultimo en el eje **Y** se muestran las variables ordenadas según la **importancia que tienen en la predicción del modelo**, medida por el valor absoluto medio de los valores SHAP.

2.4.5. Visualización y análisis gráfico

Con el objetivo de generar resultados más interpretables, se realizaron visualizaciones que permiten observar de forma clara la distribución de los datos, las matrices de correlación, así como integrar los resultados provenientes del procesamiento de los videos (predicciones, interpolaciones, etc.) con los datos del *eye-tracker*.

Para la generación y personalización de estos gráficos se utilizaron las bibliotecas `Matplotlib` [17] y `Seaborn` [18] que permitieron adaptar las visualizaciones a las necesidades específicas del análisis.

3. TRACKEO DE IDENTIDAD DE OBJETOS

Para el algoritmo de trackeo de múltiples objetos (MOT) decidimos utilizar un algoritmo ya implemetando previamente. Para decidir cual íbamos a usar nos basamos en el leaderboard del dataset BDD100K, al ser un dataset enfocado en entornos de conducción presentan dificultades muy similares a las nuestras permitiendo así manejar adecuadamente los encubrimientos y desapariciones temporales. Dado que ByteTrack [4] ocupó el primer lugar en este benchmark de MOT, decidimos utilizarlo.

3.1. ByteTrack:

ByteTrack es un método de asociación de identidad de objetos simple pero efectivo. Cuenta con una estrategia clave, utilizar casi todas las cajas de detección en lugar de descartar aquellas con puntuaciones que no superan el umbral predefinido para la detección de objetos. Esta incorporación de las detecciones de baja puntuación es fundamental para recuperar objetos que podrían estar parcialmente ocultos o cuya confianza se vea afectada por factores climáticos, como la luz, y que a menudo se pierden en métodos mas simples, lo que resulta en trayectorias fragmentadas.

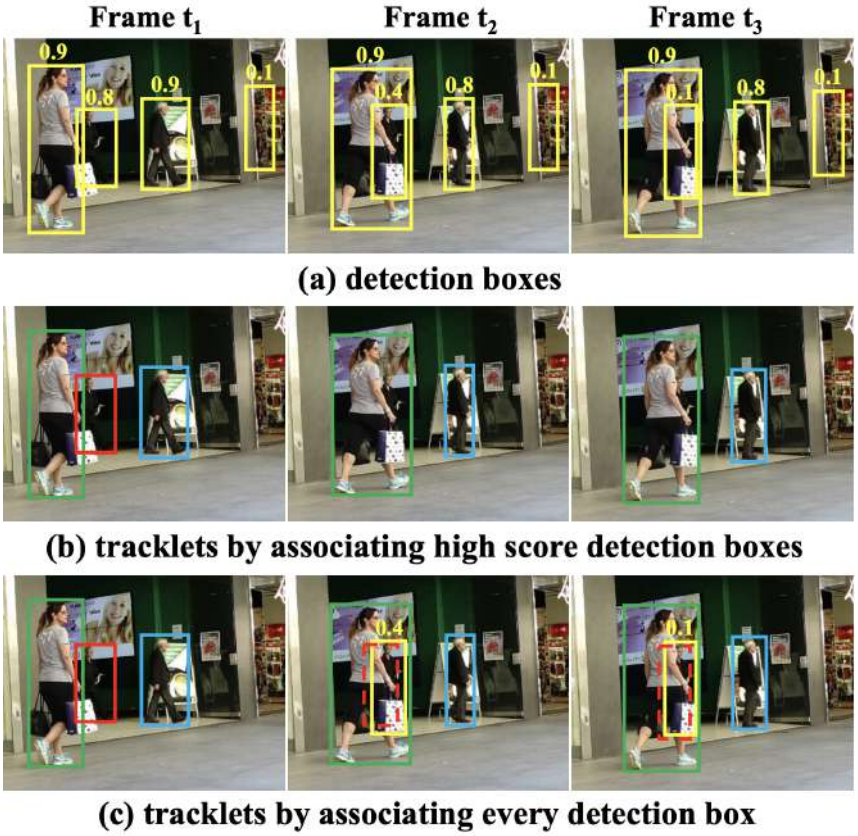

Fig. 3.1: Funcionamiento de ByteTrack

En la Figura 3.1 podemos observar cómo funciona la incorporación de detecciones de más baja confianza con ByteTrack.

Los parámetros seleccionados para el procesamiento de los videos con *ByteTrack* fueron los siguientes:

- **track_high_thresh:** `0,5`

  Umbral de confianza para la primera asociación entre detecciones y trayectorias existentes. Detecciones con confianza superior a este valor se consideran para el emparejamiento principal.

- **track_low_thresh:** `0,002`

  Umbral inferior para una segunda ronda de asociación, generalmente con detecciones de menor confianza.

- **new_track_thresh:** `0,3`

  Umbral mínimo para iniciar una nueva trayectoria si una detección no coincide con ninguna trayectoria existente.

- **track_buffer:** `90`

  Número de fotogramas que se mantiene una trayectoria en memoria sin recibir nuevas asociaciones antes de ser eliminada.

- **match_thresh:** `0,8`

  Umbral de coincidencia de pistas. Los valores más altos hacen que la coincidencia sea más indulgente.

3.2. Interpolación de trayectorias y resolución de problemas de identidad

Pese al buen desempeño de ByteTrack, se detectaron dificultades en el seguimiento de objetos en algunos momentos, ocasionadas principalmente por la complejidad de la escena y las limitaciones de YOLO. Esto generó que se realicen predicciones incorrectas, ya sea porque se detectó un objeto incorrectamente durante algunos fotogramas, o por la desaparición de otro objeto durante ciertos fotogramas, aumentando así la probabilidad de que, al reaparecer, se le asigne un nuevo `track_id`. Estos errores son problemáticos ya que reducen la confianza al momento de realizar un análisis utilizando estos datos.

Para abordar esta problemática, se implementó un procedimiento de interpolación lineal de trayectorias y resolución de problemas de identidad, que consta de las siguientes reglas aplicadas en el siguiente orden:

- **Eliminación de Tracks Cortos:** Para evitar el impacto de objetos erróneamente asignados a un `track_id` durante un número muy reducido de fotogramas (lo que puede ocurrir con detecciones falsas o un error en la asignacion de ID), se introduce una regla de eliminación. Cualquier `track_id` que aparezca en menos de un número mínimo de fotogramas (cinco) es eliminado del conjunto de datos. Esto garantiza que solo se mantengan objetos que hayan sido suficientemente detectados y rastreados a lo largo del tiempo.

- **Corrección de cambios de identidad:** En situaciones donde un mismo objeto es asignado a múltiples `track_id` debido a errores de detección, se aplica un procedimiento para unificar estas identidades. Si se observa que un objeto de una clase reaparece en los 60 fotogramas posteriores con un `track_id` diferente, y su distancia con el `track_id` anterior es menor a (`200px`), se fusionan ambos `track_id`.
- **Interpolación de cuadros delimitadores:** Cuando se detecta que un objeto desaparece temporalmente (debido a un encubrimiento o errores en la detección) y reaparece en un fotograma posterior, se realiza una interpolación de las coordenadas de su detección. Se realizó una interpolación lineal para estimar las posiciones de los objetos en los fotogramas intermedios. La interpolación se realiza únicamente si la distancia entre los fotogramas inicial y final no supera un umbral de distancia (`200px`) y si la diferencia entre los fotogramas está dentro de un rango aceptable de tiempo (`60 frames maximos`).

3.3. Motivo y resultados

Estas estrategias de procesamiento fueron implementadas para mejorar la coherencia y precisión del seguimiento de objetos de interés detectados en los videos del experimento, donde las condiciones de visibilidad cambian rápidamente debido a factores como los encubrimientos, cambios de dirección y la naturaleza del tráfico. Al aplicar estas reglas, se logró mejorar la estabilidad en el seguimiento de vehículos y peatones, lo que resultó en una reducción de cambios innecesarios de `track_id`, disminuyendo las inconsistencias en el seguimiento de trayectorias.

Adicionalmente, se realizó un procesamiento basado en la inversión temporal de los videos. Cada video fue procesado de la siguiente manera: primero se invirtió su secuencia temporal, luego se realizaron las predicciones sobre esta versión invertida, y finalmente se aplicaron interpolaciones para suavizar las trayectorias antes de reinvertir los resultados a su orden original. Esta técnica se aplicó debido a que, en contextos como la detección de peligros, es habitual que un objeto comience parcialmente oculto y luego se vuelva visible progresivamente. Dado que ByteTrack asocia detecciones de alta confianza en las primeras etapas del seguimiento, invertir el video permite que los momentos en los que el objeto está completamente visible (y por tanto tiene mayor confianza) sean utilizados para iniciar correctamente el seguimiento. Posteriormente, cuando el objeto comienza a desaparecer y este parcialmente oculto, las detecciones de menor confianza pueden seguir siendo asociadas a una trayectoria ya establecida, mejorando así la continuidad del seguimiento y permitiendo detectar objetos mas temprano.

Fig. 3.2: Fotogramas de los videos con el procesamiento

Como se puede apreciar en la Figura 3.2. los resultados obtenidos, después de todo el procesamiento, fueron muy satisfactorios. Por un lado se observó que YOLOv12X realizó muy buenas detecciones en los videos mientras ByteTrack, junto con la interpolación de predicciones, presentó una buena continuidad en el seguimiento de identidades de objetos. Ademas la eliminación de tracks muy cortos redujo el ruido generado por detecciones erróneas, lo que disminuyo la cantidad de falsas detecciones presentes.

Una vez finalizado el procesamiento tanto de los videos como de los datos provenientes del *eye tracker*, se comenzó con el análisis general de las variables obtenidas con la integración de ambas fuentes, exploramos sus distribuciones, correlaciones y características más relevantes para comprender la relación entre el comportamiento ocular y la capacidad de accionar frente a situaciones de peligro.

4. ANÁLISIS EXPLORATORIO Y CORRELACIONAL DE LOS DATOS

El primer paso consistió en realizar un análisis exploratorio con el objetivo de verificar la calidad de los datos, detectar posibles errores durante el procesamiento o anomalías inherentes a los videos, así como comprender mejor las características de las variables, en particular sus distribuciones para poder saber con que condiciones contamos al momento de buscar correlaciones. Por ejemplo, si una variable presenta una distribución aproximadamente normal, resulta que estamos en condiciones de utilizar el coeficiente de correlación de Pearson; en caso de que siga otra distribuciones, es preferible emplear el coeficiente de Spearman.

En particular, comenzamos analizando la distribución de los tiempos de reacción, con el objetivo de determinar si se ajustan a alguna distribución teórica conocida.

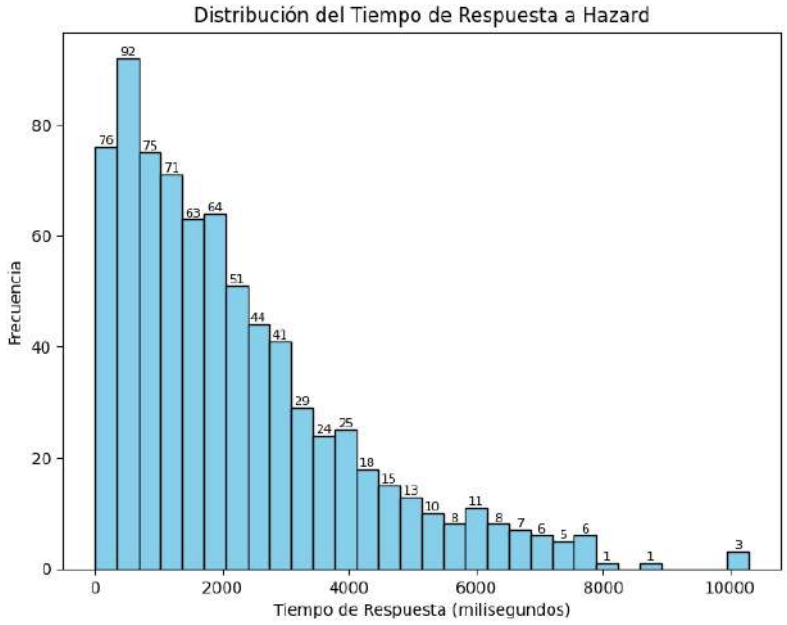

Fig. 4.1: Distribución de los tiempos de respuesta registrados con botón.

A simple vista, los tiempos de respuesta presentan una distribución asimétrica y positiva, con una alta concentración de valores por debajo de los 2 segundos (2000 ms), lo que sugiere que no presenta una distribucion normal aunque podrían seguir una distribución asimetrica positiva como una Gamma. Para evaluar esta hipótesis, se ajustó una distribución Gamma a los datos mediante el método de máxima verosimilitud, utilizando la implementación provista por scipy.

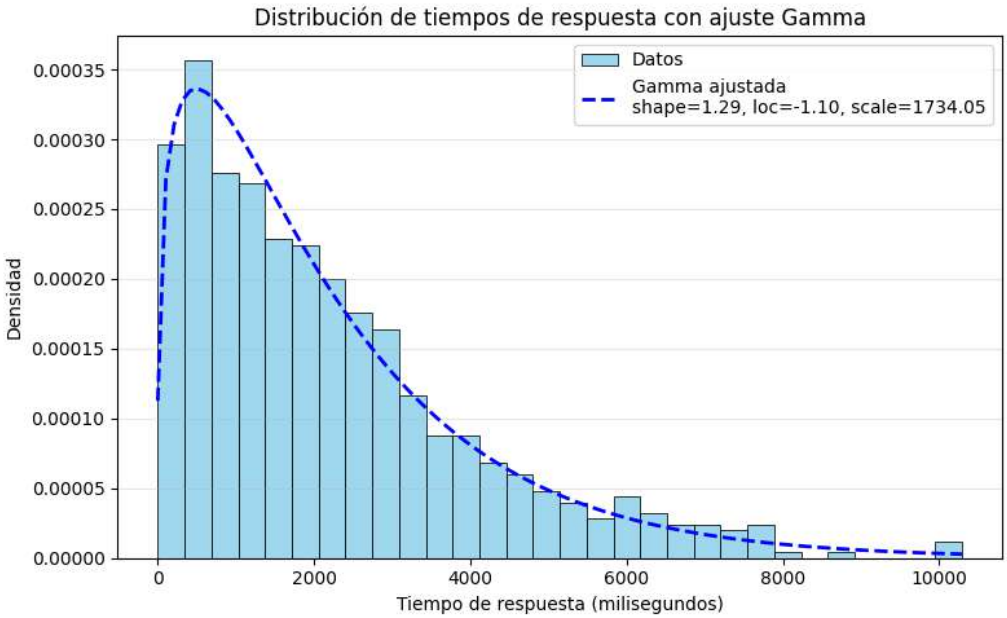

Fig. 4.2: Distribución Gamma ajustada sobre los tiempos de respuesta empíricos.

Realizamos un ajuste de máxima verosimilitud para conseguir los parámetros de una distribución Gamma que mejor se ajusten a la distribución de los tiempos de reacción, y graficamos esta distribución junto con los datos en la Figura 4.2, en la que se observa que los datos se ajustan muy bien a dicha distribución.

Para comprobar si efectivamente siguen esta distribucion, aplicamos el test de Kolmogorov-Smirnov; la implementación utilizada fue la de `scipy`. Las hipótesis evaluadas son: H_0: los tiempos de respuesta provienen de una distribución Gamma con los parámetros estimados; H_1: los tiempos de respuesta no provienen de dicha distribución.

El resultado del test arrojó un p-valor de 0,5861, lo que indica que, al considerar un nivel de significancia del 5 % (o un nivel de confianza del 95 %), no existen evidencias suficientes para rechazar la hipótesis nula. En consecuencia, podemos asumir que los tiempos de respuesta observados siguen una distribución Gamma con parámetros shape = 1,29, loc = -1,10 y scale = 1734,05.

Si bien los tiempos de respuesta de los conductores ante los peligros no presentan una distribución normal, sino que se ajustan correctamente a una distribución Gamma, el análisis continuó con los tiempos de fijación sobre los peligros en cada uno de los videos.

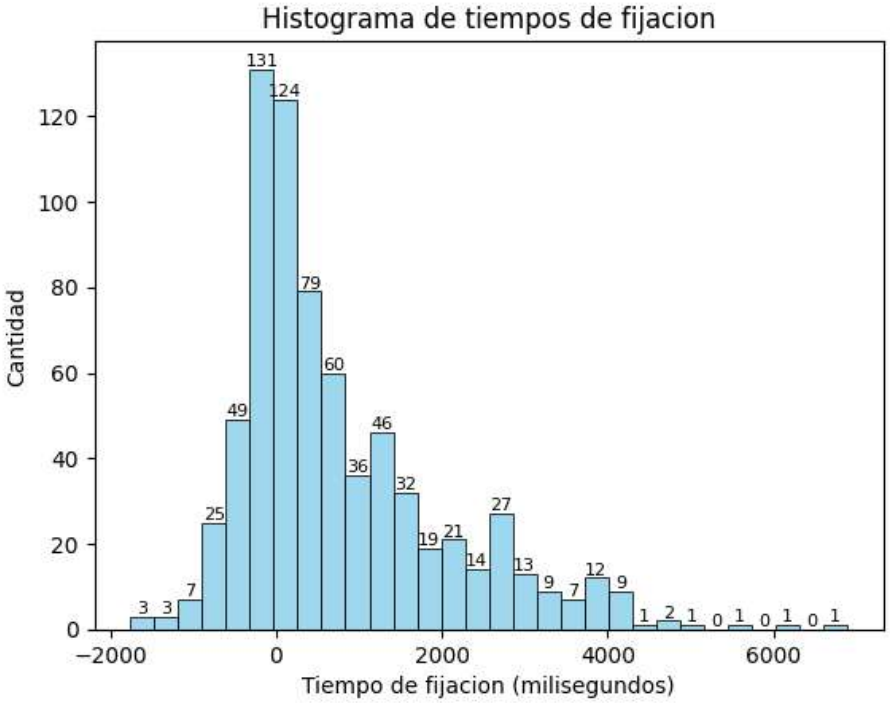

Fig. 4.3: Distribución de los tiempos de fijación sobre los peligros.

En este caso, los tiempos de fijación pueden ser negativos, lo que descarta distribuciones que solo admiten valores positivos, como la Gamma o la Chi-cuadrado. Se observa una alta concentración de fijaciones entre los 0 a 1 segundos, aunque la distribución no es completamente simétrica. Por esto, se realizó un ajuste de los datos con las distribuciones que permiten valores negativos, como la Normal, la t de Student y la Laplace.

Los resultados de los tests de bondad de ajuste fueron los siguientes:

- Shapiro-Wilk para distribución normal: estadístico = 0,8889, p-valor = 0,0000
- Kolmogorov-Smirnov para normal: estadístico = 0,1528, p-valor = 0,0000
- Kolmogorov-Smirnov para t de Student: estadístico = 0,1056, p-valor = 0,0000
- Kolmogorov-Smirnov para Laplace: estadístico = 0,1320, p-valor = 0,0000

Tanto el test de *Shapiro-Wilk* como los de *Kolmogorov-Smirnov* mostraron p-valores estadísticamente significativos (menores a 0,05), lo que indica que, considerando un nivel de significancia del 5 % (95 % de confianza), se tiene evidencia suficiente para rechazar la hipótesis nula en todos los casos. Esto nos indica que ninguna de las distribuciones evaluadas se ajusta correctamente a los tiempos de fijación, incluso tras estimar sus parámetros mediante el método de máxima verosimilitud.

Al analizar las distribuciones de los tiempos de reacción y de fijación, dos variables muy relevantes en nuestro trabajo, observamos que ninguna cumple con el supuesto de normalidad por lo que no podemos utilizar el coeficiente de correlación de Pearson con estas variables. Además, dado que nuestro interés se centra en detectar relaciones monotónicas y no exclusivamente lineales entre las variables, optamos por utilizar el coeficiente de correlación de Spearman para el iniciar análisis de correlaciones con todas ellas.

Por último, antes de iniciar el análisis de correlaciones, se analiza la distribución de los tiempos de respuesta agrupados por video y por usuario, con el objetivo de detectar posibles errores en el procesamiento de datos del *eye-tracking*, así como también anomalías asociadas a las características intrínsecas del experimento.

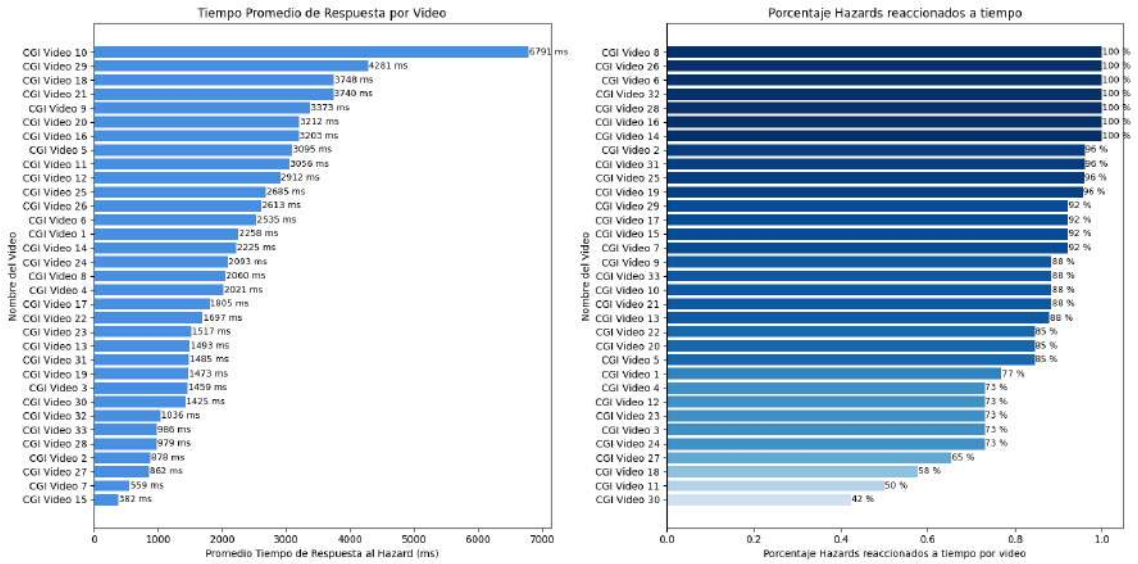

Fig. 4.4: Distribución del tiempo de respuesta y porcentaje de respuestas por video

Como se observa en la Figura 4.4, si excluimos el Video 10, la mayoría de los videos presentan tiempos de respuesta razonables, con un valor máximo aproximado de 4,2 segundos. Sin embargo, el Video 10 destaca por tener un tiempo de respuesta medio mucho mayora los otros videos de alrededor de 6,8 segundos.

Para investigar este caso particular, analizamos con mayor detalle el contenido del Video 10. Comprobamos que no se trataba de un error de procesamiento, sino de una situación compleja de interpretar, ya que el peligro aparece con bastante antelación, y aunque se debe comenzar a disminuir la velocidad en ese momento, existe una ventana temporal muy amplia para reaccionar.

Esto se confirma al analizar el porcentaje de respuestas a tiempo mediante los botones de reacción de la Figura 4.4, observamos que en el 88 % de los participantes logró reaccionar dentro del tiempo marcado, lo que sugiere que, a pesar del mayor tiempo promedio, la mayoría reconoció el peligro y logro actuar en consecuencia a pesar de detectarlo bastante después de su aparición.

En el caso del porcentaje de peligros reaccionados a tiempo, lo único que llama la atención es el bajo rendimiento observado en los videos 11, 18 y 30. Esto se debe a que presentan situaciones más complejas, ya sea por la dificultad para identificar el momento exacto en el que se debe reaccionar o por contar con ventanas temporales muy reducidas. Dado que estas diferencias no se deben a errores de procesamiento, se continúa con el análisis de los tiempos por usuario, con el objetivo de verificar que todos los datos hayan sido procesados correctamente.

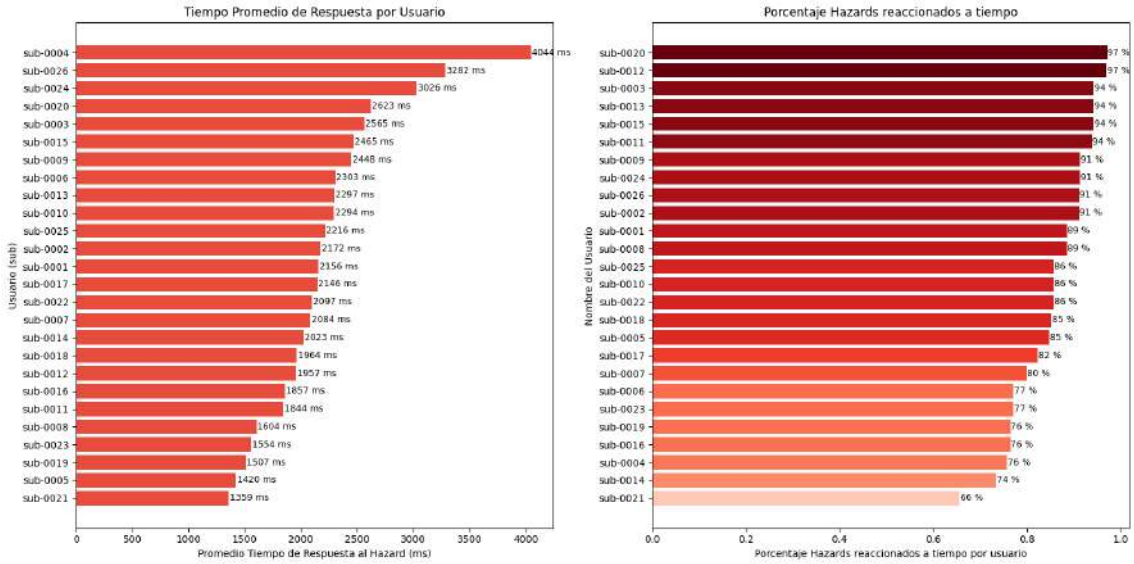

Fig. 4.5: Distribución del tiempo de respuesta y porcentaje de respuestas por usuario

Al observar las respuestas según el usuario, no se detecta ningún error relacionado con el procesamiento de los datos. El sujeto 4 parece ser el más lento en reaccionar, siendo en promedio 0,8 segundos más lento respecto al segundo más lento. Sin embargo, logró reaccionar correctamente en el 76 % de los videos, por lo que no se evidencia ningún error en el procesamiento de sus datos.

4.1. Análisis de correlaciones

Ahora, con el análisis exploratorio realizado, comenzaremos a buscar relaciones entre las variables de interés.

Primero, ¿por qué la variable que queremos explicar es el **tiempo de reacción** y no el **tiempo de fijación**? Esto se debe a que nuestra mejor aproximación a la capacidad de reaccionar correctamente ante un peligro es la velocidad con la que los participantes presionaron los botones durante el experimento. Mientras que el tiempo de fijación refleja la habilidad de prestar atención en el lugar correcto en el momento adecuado, el tiempo de reacción no solo requiere esta misma habilidad (ya que para reaccionar a tiempo es necesario haber detectado el peligro), sino que también implica que, independientemente de la complejidad de la escena o de la situación, el participante fue capaz de realizar todo el procesamiento mental necesario para concluir que debía actuar para evitar un accidente. Esta capacidad de reacción es el foco principal de este trabajo.

Por lo tanto, nuestro principal propósito es intentar encontrar relaciones entre el tiempo de reacción y las métricas características del comportamiento ocular: tales como la amplitud de búsqueda horizontal y vertical, la duración promedio de las fijaciones, la cantidad total de fijaciones y el tiempo que tarda en fijar la mirada sobre el peligro.

Como solo en el 85 % de los casos se reaccionó al peligro en el tiempo indicado, tomamos la decisión de agregar una nueva métrica en la cual penalizamos estas no reacciones con el tiempo máximo permitido para reaccionar. Esto nos permite conservar dichos casos en el análisis, en lugar de descartarlos. Asimismo, decidimos mantener la columna original con los tiempos reales de reacción, en caso de que esta decisión afectara significativamente las

correlaciones observadas.

Para obtener una primera visión general de los datos, decidimos construir una matriz de correlación con el objetivo de identificar posibles relaciones entre las variables. El cálculo de la correlación se realizó utilizando el coeficiente de **Spearman**. Ver Sección 2.4.3.

4.2. Análisis agrupado por usuario

Dado que cada participante posee un patrón de búsqueda visual característico, influenciado por su experiencia y nivel de desempeño en general, y considerando que todos observaron el mismo conjunto de videos, nos proponemos analizar la relación entre el comportamiento ocular de cada individuo y su desempeño general a lo largo de todo el experimento.

Para ello, optamos por promediar las métricas de comportamiento ocular a lo largo de todos los videos para cada usuario. Este se hizo para reducir la variabilidad asociada al contenido específico de cada video y centrarse en las diferencias entre individuos. Esto nos permite realizar un análisis más robusto de los patrones visuales y su relación con el desempeño general, evitando que las características particulares de los videos, como la complejidad de la situación en particular, las condiciones lumínicas o la presencia de distractores que terminen enmascarando las verdaderas relaciones entre el comportamiento ocular y el rendimiento de los participantes.

Al observar la matriz de la Figura 4.6 y concentrarnos particularmente en las dos últimas filas, que presentan las correlaciones con los tiempos de reacción, no se encontró ninguna con un nivel de significancia superior al 95 %. Esto sugiere que, si bien algunos coeficientes de correlación aparentan indicar la presencia de una relación entre el comportamiento ocular y la capacidad de reaccionar ante situaciones de peligro, no pueden considerarse válidos debido a que no son estadísticamente confiables. La razón detrás de esta baja confiabilidad radica en que, al agrupar los datos por usuario, la cantidad total de observaciones se redujo considerablemente, resultando en una sola fila por participante. Fue esta reducción en el volumen de datos la que impactó negativamente en la robustez estadística de los coeficientes de Spearman.

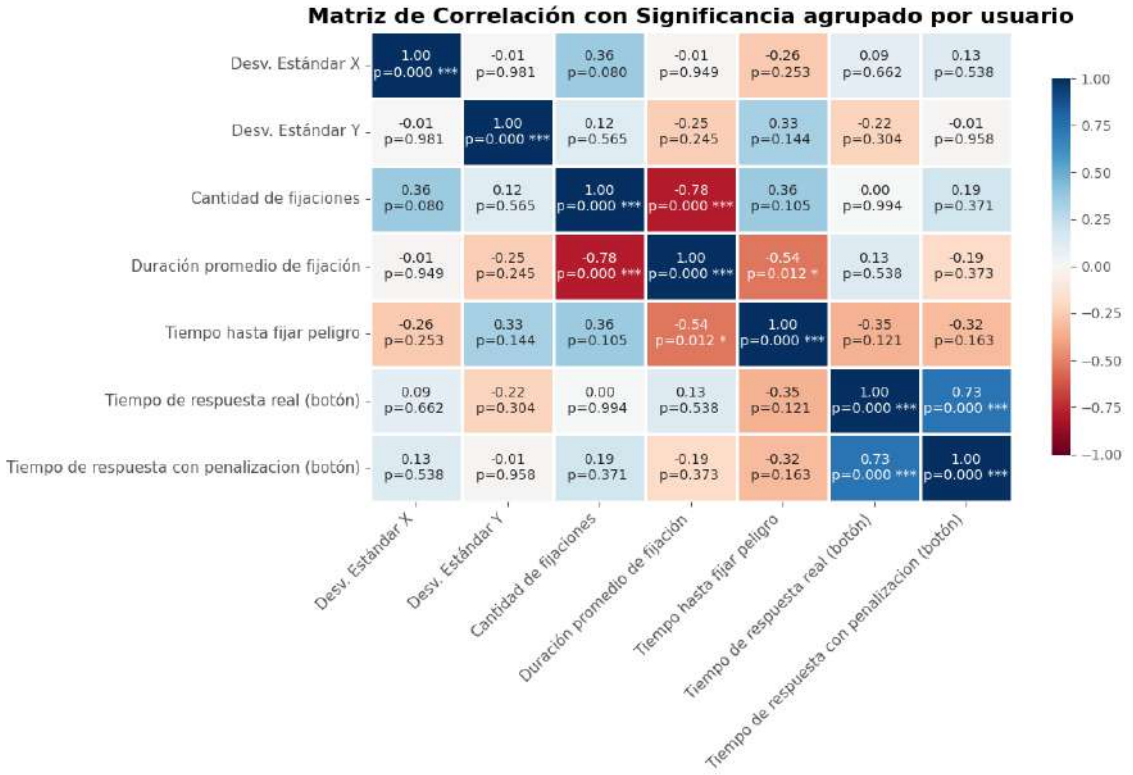

Fig. 4.6: Matriz de correlación entre el tiempo de reacción y las variables del comportamiento ocular agrupado por usuario.

Por lo tanto, con el fin de continuar el análisis desde otra perspectiva, decidimos implementar una nueva estrategia: entrenar múltiples modelos de Random Forest e interpretar la importancia de las variables mediante SHAP. El objetivo de este enfoque es investigar la contribución conjunta de cada variable en la predicción del tiempo de reacción asignado por el modelo entrenado.

Se entrenó un *Random Forest* con 1000 árboles utilizando como variables predictoras el desvío estándar en los ejes horizontal y vertical, la duración promedio de las fijaciones y el tiempo hasta la primera fijación. La variable a predecir fue el tiempo de respuesta con penalización.

Elegimos esta métrica como variable objetivo porque consideramos importante penalizar las situaciones en las que no se reaccionó correctamente. Si hubiésemos utilizado únicamente el tiempo real de respuesta, los casos sin respuesta simplemente no serían considerados, lo cual impediría reflejar adecuadamente la falta de detección de los peligros. Esto podría generar problemas al momento de buscar relaciones entre estas no respuestas y el comportamiento ocular. Esta decisión nos permitió entrenar un modelo más robusto y obtener estimaciones más confiables sobre la importancia de cada variable sobre la capacidad de prevenir accidentes.

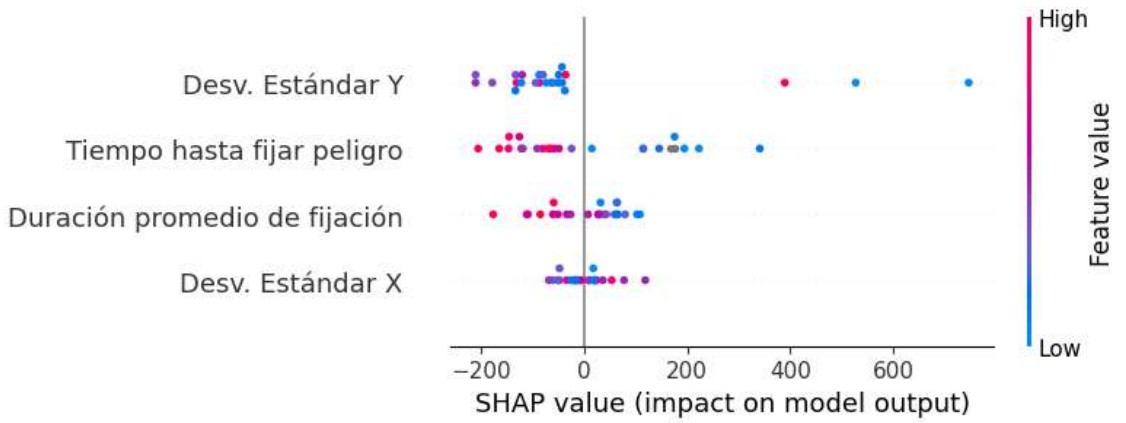

Fig. 4.7: Importancia de variables según SHAP - Resultados agrupados por usuario

Para interpretar correctamente la Figura 4.7, recordemos que SHAP estima la contribución de cada variable a la predicción del modelo, evaluando cuánto cambia la salida al incluir o excluir dicha variable. Para una explicación más detallada, puede consultarse la Sección 2.4.4.

En la Figura 4.7 se observa que la variable con mayor relevancia según SHAP fue el *desvío estándar en el eje Y*. Sin embargo, no parece haber una relación clara entre esta y el *tiempo de reacción*, ya que la concentración de puntos con valores SHAP negativos, que, por lo tanto, disminuyen el tiempo de reacción predicho por el modelo, incluye tanto valores altos como bajos de dicha variable.

Por otro lado, en el caso del *tiempo de fijación*, se observa que un mayor tiempo de fijación (indicado por cúmulos de puntos rojos) se asocia con una disminución en el tiempo de reacción, mientras que un menor tiempo de fijación (puntos azules) se asocia con un mayor tiempo de reacción. Esta relación resulta, a priori, contradictoria y difiere de lo que uno esperaría en un principio.

Una posible explicación es que, al agrupar los datos por usuario, se reduce considerablemente la cantidad de observaciones, lo que amplifica la acumulación del error. Esta situación ademas se ve agravada por el hecho de que cada usuario posee un nivel distinto de calibración del *eye-tracker*, lo cual afecta directamente la calidad de las mediciones de fijación. Al analizar de forma separada cómo influye la calidad de la calibración sobre el tiempo de fijación, se obtiene el siguiente resultado:

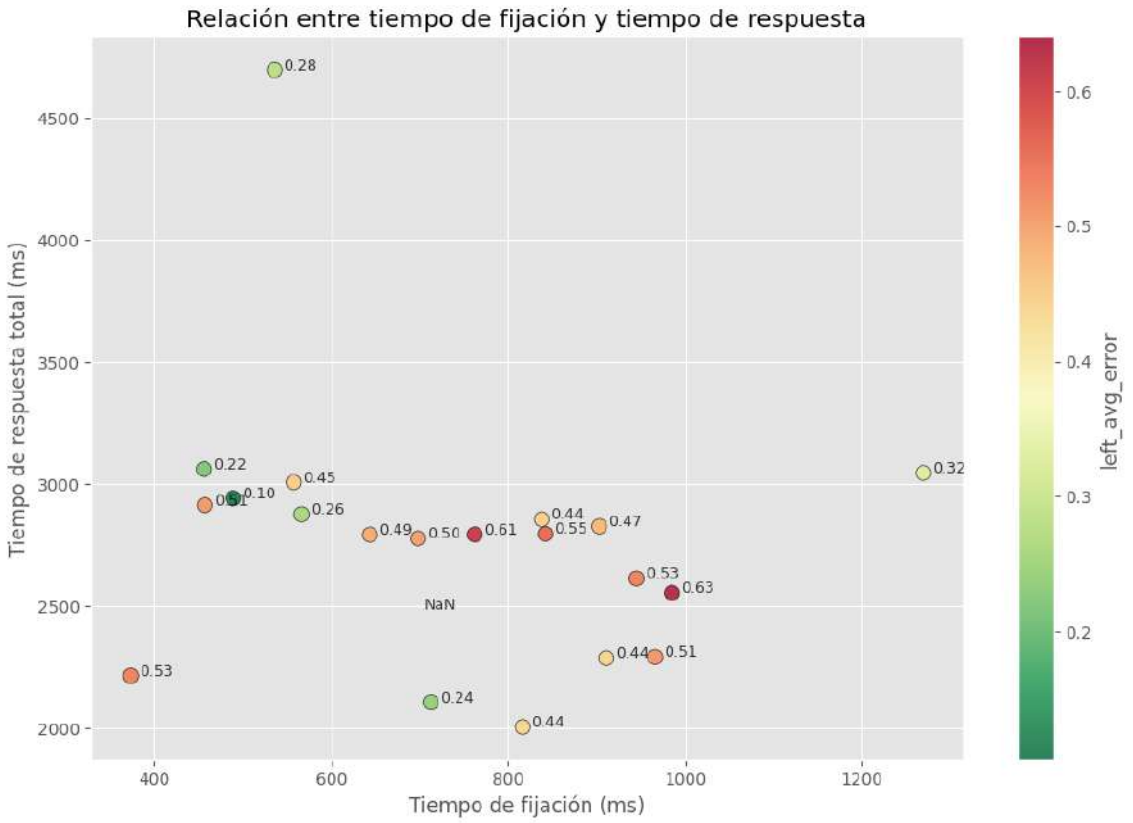

Fig. 4.8: Relación entre tiempos de reacción y fijación con la calidad de calibración por usuario.

En la Figura 4.8 el eje X representa el tiempo de fijación promedio, mientras que el eje Y corresponde al tiempo de reacción promedio. El color y valor de cada punto indican su error de calibración, siendo los puntos en rojo aquellos con mayor error, y los puntos en verde los que presentan el menor error de calibración.

A partir de estos datos, se observa una tendencia clara: los participantes con peor calidad de calibración tienden a presentar tiempos de fijación promedio más altos. Es decir, a mayor error de calibración del *eye-tracker*, mayor es el tiempo registrado hasta que se detecta la fijación sobre un peligro. Esta relación es esperable, ya que una calibración deficiente compromete la precisión espacial del registro ocular, lo cual dificulta su correcta integración con las detecciones de objetos realizadas por YOLO. Como consecuencia, se retrasa o incluso se imposibilita la asignación precisa de una fijación sobre los peligros simulados, lo que influye directamente en el cálculo del tiempo de fijación.

Cabe destacar que esta dependencia se evidencia en los tiempos de fijación, pero no en los tiempos de reacción. Esto se debe a que el tiempo de reacción está determinado por la interacción consciente del usuario, presionar un boton cuando siente que debe reaccionar y no depende directamente del mapeo espacial del punto de mirada proporcionado por el *eye-tracker*, por lo tanto no se ve afectado por la calidad de la calibración.

Por lo tanto, con el objetivo de evitar la acumulación de errores derivados de la calibración al agrupar los resultados por usuario y al mismo tiempo, contar con una mayor cantidad de observaciones que permita obtener correlaciones más robustas y confiables, decidimos desagregar el análisis. Específicamente, el estudio se realizará a nivel de cada usuario por video, permitiendo así evaluar con mayor precisión la relación entre el **com-**

portamiento ocular y el **desempeño individual** en términos de velocidad de reacción ante peligros en contextos particulares.

4.3. Análisis de tiempos y correlaciones con datos desglosados

Retomamos el análisis de la matriz de correlación, esta vez considerando los datos desglosados por usuario y por video.

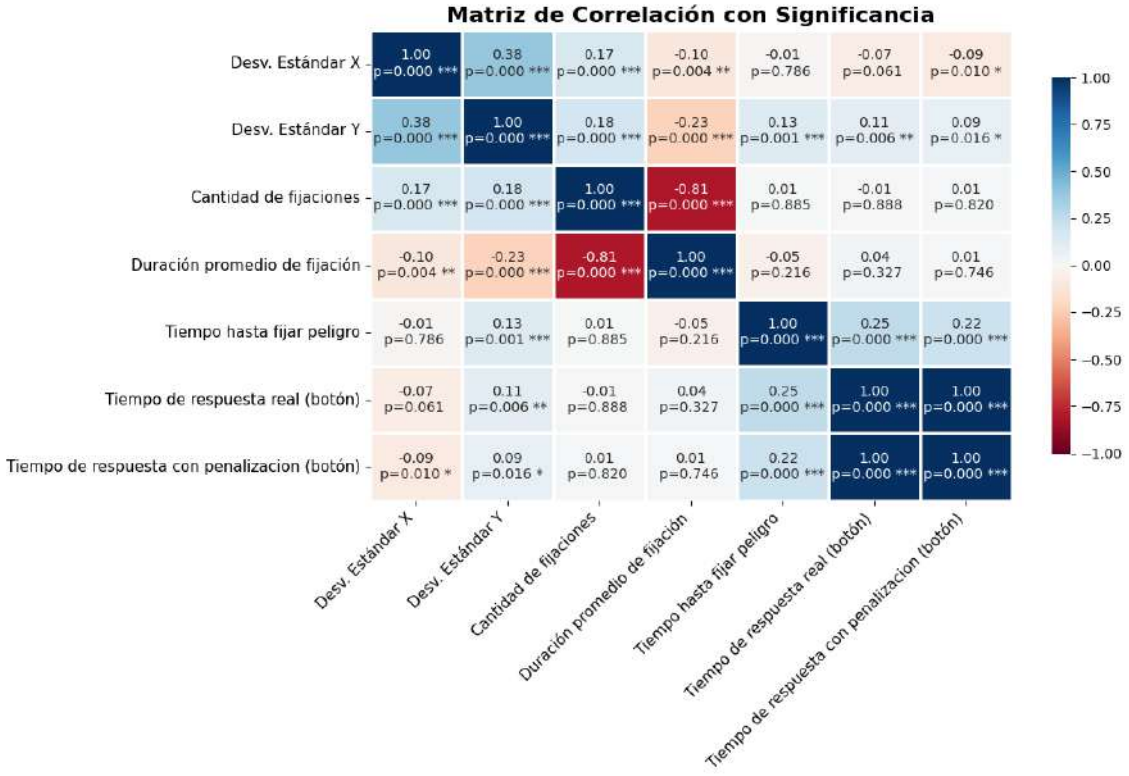

Fig. 4.9: Matriz de correlación entre el tiempo de reacción y las variables del comportamiento ocular.

En la Figura 4.9 nos centramos en la última fila de la matriz, correspondiente a la correlación entre el **tiempo de reacción** y las demás variables de comportamiento ocular. Si bien ninguna variable presenta una correlación elevada, se identifican patrones sutiles acompañados de un alto nivel de significancia estadística. Esto sugiere que, aunque las relaciones no son fuertes, sí existen asociaciones sistemáticas entre el comportamiento visual y el tiempo de respuesta, lo cual refuerza la validez del análisis y motiva un estudio más detallado de las dinámicas involucradas.

En particular, se observa que una mayor amplitud de búsqueda en el eje horizontal tiende, aunque de forma leve con un coeficiente de -0,09, a asociarse con tiempos de reacción más bajos. Por el contrario, una mayor amplitud de búsqueda en el eje vertical parece relacionarse, también de manera tenue con un coeficiente de 0,09, con un aumento en el tiempo de reacción, ambas con un nivel de signifancia mayor al 95 %.

En cuanto a la duración promedio de las fijaciones y la cantidad de las mismas, los valores de correlación son igualmente bajos o incluso menores, y los niveles de significancia

estadística son insuficientes: con un p-valor de 0,820 para la cantidad de fijaciones distintas y 0,746 para la duración promedio de las fijaciones, muy por arriba del umbral comúnmente aceptado de 95 %.

Por otro lado, al analizar la correlación entre el tiempo de reacción y el tiempo de fijación en las situaciones detectadas como peligrosas, se obtuvieron los valores más elevados y niveles de significancia estadística muy buenos. En particular, se observó una correlación de 0,22 para los tiempos con penalización, lo cual resulta consistente con lo que uno espera en la vida real. Ya que aunque no todas las situaciones fijadas correctamente y a tiempo son necesariamente procesadas y reaccionadas debidamente, el hecho de detectar y fijar la mirada en un peligro con antelación otorga al conductor/participante un mayor margen temporal para procesar la situación y reaccionar debidamente. Por lo tanto, una correlación positiva entre el tiempo de fijación y el tiempo de reacción en contextos de conducción es esperable y conceptualmente consistente.

Sin embargo, es importante remarcar que estas correlaciones, aunque estadísticamente significativas, tienen coeficientes relativamente bajos. Este resultado es esperable, ya que la detección y el procesamiento de una situación de peligro no dependen exclusivamente del comportamiento visual sino que también influyen factores cognitivos, de atención, la experiencia previa y la capacidad de toma de decisiones. Por este motivo, decidimos repetir el análisis previamente realizado mediante el entrenamiento de múltiples modelos *Random Forest* y la posterior interpretación de la importancia de las variables utilizando valores SHAP, con el objetivo de obtener una comprensión más completa y robusta de los factores que influyen en el tiempo de reacción.

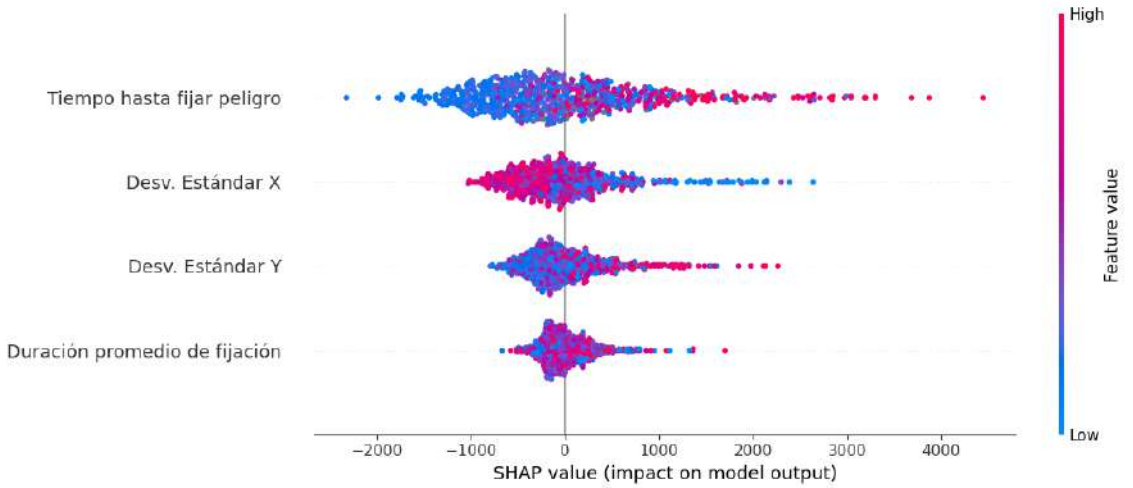

Fig. 4.10: Importancia de variables por SHAP - resultados.

La variable **Tiempo de fijación** aparece como la más influyente sobre el **tiempo de reacción**, al ocupar la primera posición en el gráfico. Además, se observa una concentración de puntos azules con valores SHAP negativos, lo que indica que, en general, **valores bajos de tiempo de fijación tienden a reducir el tiempo de reacción**. También se nota una mayor presencia de puntos rojos con valores SHAP positivos, lo que sugiere que **valores altos de tiempo de fijación aumentan el tiempo de reacción**.

En el caso de la variable **Desvío estándar en el eje X**, se observa una concentración de puntos rojos con valores SHAP ligeramente negativos. Esto sugiere que, aunque esta

variable tiene una influencia menor que el tiempo de fijación, **un alto desvío estándar en el eje horizontal también contribuye a una disminución del tiempo de reacción**.

Para la variable **Desvío estándar en el eje Y**, se observa una leve concentración de puntos azules con valores SHAP negativos y puntos rojos con valores SHAP positivos. Esto sugiere que una menor amplitud de búsqueda vertical tiende a asociarse con tiempos de respuesta más bajos, mientras que una mayor amplitud de búsqueda en el eje Y se asocia con tiempos de respuesta más altos.

En el caso de la variable **Duración promedio de fijación**, no se aprecia un patrón tan marcado como en los casos anteriores al analizar el gráfico de resumen generado por SHAP. Por lo tanto, no es posible concluir, a partir de este análisis, una relación clara entre esta variable y el tiempo de reacción predicho por el modelo.

4.4. Comparación con Trabajos Previos

Al comparar nuestros resultados con los resultados del estudio de Robbins et al. [1], se evidencian algunas diferencias relevantes. Si bien no trabajamos exactamente con las mismas variables, consideramos que las conclusiones de Robbins sobre las diferencias comportamentales entre grupos segmentados según la experiencia de conducción deberían, en principio, reflejarse en la capacidad de reacción ante situaciones de peligro. Esto es debido a que a un mayor nivel de experiencia, la respuesta frente a accidentes deberia ser más rápida.

En cuanto a la duración y la cantidad de fijaciones visuales, Robbins no encontró una diferencia clara entre los conductores segmentados según su nivel de experiencia. De la misma forma, en este trabajo tampoco se encontro que estas variables influyan en la velocidad de reacción ante peligros.

Por otro lado, encontramos una diferencia notable en lo que respecta a la amplitud de búsqueda visual horizontal y vertical. Mientras que Robbins no encontró diferencias significativas en la amplitud de búsqueda visual en entornos no inmersivos según la experiencia de los conductores, en nuestro estudio encontramos que una mayor amplitud de búsqueda horizontal y una menor amplitud de búsqueda vertical tienden, a disminuir el tiempo de reacción de los conductores en entornos no inmersivos.

4.5. Limitaciones

Una de las limitaciones más importantes que afectó el desarrollo del trabajo fue la poca cantidad de participantes en el experimento. Esta imposibilitó realizar un análisis estadístico más robusto y limitó la exploración de diferencias interindividuales en el comportamiento ocular, así como el control del efecto de la complejidad de los distintos videos de forma independiente. Contar con una muestra más amplia hubiese permitido identificar patrones más generalizables y relaciones más sólidas entre desempeño y comportamiento visual.

Por otro lado, si bien la performance del modelo *YOLOv12X* resultó suficiente para el contexto del experimento, consideramos que podría mejorarse sustancialmente mediante un proceso de *fine-tuning* enfocado exclusivamente en los objetos de interés definidos en este trabajo. Esto permitiría aumentar la precisión y confianza de las predicciones, mejorando así la estimación de los tiempos de detección de peligros y permitiendo realizar un análisis mas profundo sobre estas cuestiones.

Finalmente, otro limitante muy importante a la hora de realizar el análisis fue la calidad de las calibraciónes del *eye-tracker*, ya que una mala calibración impacta directamente en la precisión de las métricas derivadas del Eye Tracker, principalmente en el tiempo de fijación aunque también de menor manera en la amplitud de búsqueda vertical y horizontal. Una mejora en los procedimientos de calibración, contribuiría significativamente a la fiabilidad de los datos y aumentar la validez de los análisis posteriores.

4.6. Trabajos futuros

Una de las maneras de ampliar este trabajo es la incorporación de modelos lineales mixtos en el análisis estadístico de los datos. A diferencia de los modelos lineales simples, los modelos lineales mixtos permiten incluir efectos aleatorios, lo cual resulta especialmente útil en nuestro caso, donde tanto el sujeto del experimento como el video observado influyen en las métricas medidas. Específicamente, se podrían modelar como variables aleatorias el *ID del participante* y el *ID del video*, y como efectos fijos las métricas de comportamiento ocular (tiempo de fijación, amplitud de búsqueda, etc.).

Esto permitiría captar de manera más precisa la influencia del comportamiento ocular sobre el tiempo de reacción, sin que los resultados se vean sesgados por las diferencias individuales entre participantes o las variaciones en la complejidad de los videos.

5. CONCLUSIONES

En este trabajo nos propusimos como objetivo principal replicar, validar y ampliar estudios previos sobre la relación entre el comportamiento ocular de los conductores y su desempeño para detectar y reaccionar ante situaciones de peligro.

Para alcanzar este objetivo, fue necesario en primer lugar elegir un modelo de detección y un algoritmo de seguimiento de objetos para permitirnos incorporar la variable del tiempo tardado en fijar la mirada sobre los peligros del video en el análisis de relaciones. Tras realizar múltiples comparaciones entre distintos modelos de detección de objetos, concluimos que el modelo **YOLOv12X**, en combinación con **ByteTrack** como tracker y la incorporación de reglas de interpolación de cuadros para mejorar la continuidad del seguimiento, cumplía satisfactoriamente con los requisitos del estudio. Por lo tanto, se decidió emplear esta configuración para procesar todos los videos del experimento.

Paralelamente, fue necesario desarrollar un procesamiento de los datos provistos por el *eye-tracker*, con el objetivo de calcular todas las métricas relevantes para este estudio. Entre ellas se encuentran: el *tiempo de respuesta* de los participantes, el *porcentaje de detección efectiva de peligros*, la *amplitud de búsqueda visual*, la *duración y cantidad de fijaciones*, entre otras. Estas métricas fueron posteriormente integradas con los resultados del modelo de detección de objetos, permitiendo una evaluación completa de los efectos que tienen estas características del comportamiento ocular sobre la capacidad de los participantes para reaccionar correctamente a situaciones altamente probables de desencadenar un accidente de trafico en caso de no ser respondidas adecuadamente.

En el análisis de la relación entre el comportamiento ocular y la capacidad para detectar y procesar correctamente los peligros, se observó una fuerte correlación entre el *tiempo de reacción* y el *tiempo de fijación* sobre los peligros, lo que resalta la importancia de una exploración visual eficiente para lograr una detección temprana y oportuna de posibles situaciones que puedan desencadenar un accidente.

Además, se identificó que una mayor amplitud de búsqueda en el eje horizontal, combinada con una menor amplitud en el eje vertical, se asocia con tiempos de reacción ligeramente menores. Por otro lado, la *duración* y la *cantidad* de fijaciones no mostraron una relación estadísticamente significativa con el tiempo de reacción.

A partir de los resultados obtenidos, y considerando la presencia de una correlación entre el tiempo de fijación visual ante un peligro y el tiempo de reacción, consideramos pertinente incorporar una etapa de evaluación perceptiva y de reacción similar a la implementada en el Reino Unido en el examen de obtención de licencias de conducir. La misma podría desarrollarse mediante plataformas no inmersivas, como la visualización de videos, o bien a través del uso de simuladores inmersivos, que ofrecen un entorno más realista y controlado para la medición de estos indicadores.

Por otro lado, el efecto de las variables vinculadas a la amplitud de búsqueda visual sobre el tiempo de reacción ofrecen pautas para optimizar la enseñanza de la conducción, promoviendo estrategias visuales más efectivas y rápidas.

Bibliografía

[1] Chloe Robbins and Peter Chapman. How does drivers' visual search change as a function of experience? a systematic review and meta-analysis. *Accident Analysis & Prevention*, 132:105266, 2019.

[2] David Crundall. Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard. *Accident Analysis & Prevention*, 45:600–609, 2012.

[3] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics yolo (version 8.0.0) [computer software], 2023.

[4] Yifu Zhang, Zhiqiang Wei, Qingxiong Yang, Shaojie Bai, Hao Yang, and Zicheng Liu. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2105.01808*, 2021.

[5] Agencia Nacional de Seguridad Vial. Estadísticas del observatorio vial nacional, 2023. Consultado el 23 de mayo de 2025.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[9] Xiaokang Meng, Zhenbo Chen, Yongxin Wang, Hongyang Wang, Xingang Yang, and Li Yuan. Conditional detr for end-to-end object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[10] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Roboflow. Self-driving car dataset. https://public.roboflow.com/object-detection/self-driving-car, n.d. Accessed: 2025-05-31.

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, pages 1–16, 2017.

[13] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[14] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[17] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[18] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.