



# FACULTAD DE INGENIERIA

Universidad de Buenos Aires

## CARRERA DE ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL

MEMORIA DEL TRABAJO FINAL

### People Behavior Tracking (PBT)

**Autor:**

**Hernán Contigiani**

Director:

PhD Pasquinell Urbani (UCh - Globant S.A)

Jurados:

Juan Vicente Montilla Cabrera (FIUBA)  
Facundo Lucianna (UNT)  
Roberto Compañy (UTN)

*Este trabajo fue realizado en la Ciudad Autónoma de Buenos Aires,  
entre diciembre de 2020 y agosto de 2021.*



## *Resumen*

La presente memoria describe el diseño de un sistema de monitoreo de personas desarrollado para Globant S.A. Tiene como principal objetivo estudiar los movimientos que realiza una persona al ingresar y transitar un espacio a fin de obtener métricas sobre las zonas que visitó.

Para poder realizar este trabajo se aplicaron conceptos de visión por computadora y aprendizaje profundo para detectar y seguir objetos en imágenes. Asimismo, se utilizaron técnicas de aprendizaje automático y segmentación para poder extraer características de dichos objetos con el propósito de aplicar técnicas de re-identificación y manejo de occlusiones que mejoran el seguimiento.



## *Agradecimientos*

Esta sección es para agradecimientos personales y es totalmente **OPCIONAL**.



# Índice general

<b>Resumen</b>	I
<b>1. Introducción general</b>	1
1.1. Sistemas de monitoreo por visión . . . . .	1
1.2. Motivación . . . . .	1
1.3. Estado del arte . . . . .	1
1.4. Objetivos y alcance . . . . .	1
<b>2. Introducción específica</b>	3
2.1. Requerimientos . . . . .	3
2.1.1. Tablas . . . . .	3
2.2. Modelos de inteligencia artificial utilizados . . . . .	3
2.2.1. Detector Yolo . . . . .	3
2.2.2. Seguidor Deepsort . . . . .	4
2.2.3. Extractor de características Osnet . . . . .	4
2.3. Desafíos en el seguimiento de personas . . . . .	4
2.4. Zonas de interés . . . . .	4
<b>3. Diseño e implementación</b>	5
3.1. Cadena de procesamiento de inteligencia artificial . . . . .	5
3.1.1. Arquitectura . . . . .	5
3.1.2. Integración de los modelos . . . . .	5
3.2. Entrenamiento de un modelo basado en Osnet . . . . .	5
3.2.1. Preparación de datos . . . . .	5
3.2.2. Entrenamiento grueso y fino . . . . .	5
3.2.3. Comparativa de los diferentes modelos de extracción de características . . . . .	5
3.3. Manejo de occlusiones y re-identificación . . . . .	5
3.3.1. Segmentación de personas . . . . .	5
3.3.2. Re-identificación por vectores . . . . .	5
3.3.3. Mejora del seguimiento por vectores . . . . .	5
3.3.4. Validación del manejo de occlusiones y re-identificación . . . . .	5
3.4. Motor de seguimiento y monitoreo (Engine) . . . . .	5
3.4.1. Arquitectura . . . . .	5
3.4.2. Máquina de estados . . . . .	5
3.5. Interfaz de usuario . . . . .	5
3.5.1. Definición de zonas de interés . . . . .	5
3.5.2. Visualización del seguimiento . . . . .	5
3.5.3. Visualización de las métricas . . . . .	5
<b>4. Ensayos y resultados</b>	7
4.1. Descripción del banco de pruebas . . . . .	7
4.2. Validación de los modelos de detección y seguimiento . . . . .	8

4.3. Validación del modelo de extracción de características . . . . .	9
4.4. Precisión de cada módulo del sistema ensayado . . . . .	12
4.5. Simulaciones . . . . .	14
4.6. Resultados . . . . .	16
<b>5. Conclusiones</b>	<b>19</b>
5.1. Resultados obtenidos . . . . .	19
5.2. Trabajo futuro . . . . .	20
<b>Bibliografía</b>	<b>21</b>

# Índice de figuras

2.1. Esquema de funcionamiento de Yolo. . . . .	4
4.1. Videos de ensayo generados en cada etapa. . . . .	7
4.2. Capturas tomadas de una personas. . . . .	8
4.3. Fallas en la detección de personas. . . . .	9
4.4. Programa de seguimiento por características. . . . .	9
4.5. Análisis de los clusters generados. . . . .	10
4.6. Sistema de seguimiento por características. . . . .	11
4.7. Precisión de seguimiento por personas. . . . .	12
4.8. Precisión de seguimiento por sistemas. . . . .	13
4.9. Intercambios de identificador por sistemas. . . . .	14
4.10. Imagen del video de tienda utilizado al comienzo el trabajo. . . . .	15
4.11. Imagen del escenario creado en el video juego "Los Sims 3". . . . .	15
4.12. Imagen del monitoreo de personas en la aplicación web. . . . .	16
4.13. Interacción de las personas con cada zona dibujada de interés. . . . .	16
4.14. Métricas de las zonas de interés del recinto. . . . .	17
4.15. Mapa de calor del recinto. . . . .	17



# Índice de tablas



*Dedicado a... [OPCIONAL]*



# **Capítulo 1**

## **Introducción general**

En este capítulo

- 1.1. Sistemas de monitoreo por visión**
- 1.2. Motivación**
- 1.3. Estado del arte**
- 1.4. Objetivos y alcance**



## Capítulo 2

# Introducción específica

En este capítulo

### 2.1. Requerimientos

Es esta sección

#### 2.1.1. Tablas

Para las tablas utilizar el mismo formato que para las figuras, sólo que el epígrafe se debe colocar arriba de la tabla, como se ilustra en la tabla [??](#). Observar que sólo algunas filas van con líneas visibles y notar el uso de las negritas para los encabezados. La referencia se logra utilizando el comando `\ref{<label>}` donde `label` debe estar definida dentro del entorno de la tabla.

```
\begin{table} [h]
\centering
\caption[caption corto]{caption largo más descriptivo}
\begin{tabular}{l c c}
\toprule
\textbf{Especie} & \textbf{Tamaño} & \textbf{Valor} \\
\midrule
Amphiprion Ocellaris & 10 cm & \$ 6.000 \\
Hepatus Blue Tang & 15 cm & \$ 7.000 \\
Zebrasoma Xanthurus & 12 cm & \$ 6.800 \\
\bottomrule
\hline
\end{tabular}
\label{tab:peces}
\end{table}
```

### 2.2. Modelos de inteligencia artificial utilizados

#### 2.2.1. Detector Yolo

*Yolo (You Only Look once)<sup>1</sup>* es un detector...

---

<sup>1</sup>[https://en.wikipedia.org/wiki/Raster\\_graphics](https://en.wikipedia.org/wiki/Raster_graphics)



FIGURA 2.1. Esquema de funcionamiento de Yolo.

- 2.2.2. Seguidor Deepsort**
- 2.2.3. Extractor de características Osnet**
- 2.3. Desafíos en el seguimiento de personas**
- 2.4. Zonas de interés**

## Capítulo 3

# Diseño e implementación

En este capítulo

- 3.1. Cadena de procesamiento de inteligencia artificial**
  - 3.1.1. Arquitectura
  - 3.1.2. Integración de los modelos
- 3.2. Entrenamiento de un modelo basado en Osnet**
  - 3.2.1. Preparación de datos
  - 3.2.2. Entrenamiento grueso y fino
  - 3.2.3. Comparativa de los diferentes modelos de extracción de características
- 3.3. Manejo de oclusiones y re-identificación**
  - 3.3.1. Segmentación de personas
  - 3.3.2. Re-identificación por vectores
  - 3.3.3. Mejora del seguimiento por vectores
  - 3.3.4. Validación del manejo de oclusiones y re-identificación
- 3.4. Motor de seguimiento y monitoreo (Engine)**
  - 3.4.1. Arquitectura
  - 3.4.2. Máquina de estados
- 3.5. Interfaz de usuario**
  - 3.5.1. Definición de zonas de interés
  - 3.5.2. Visualización del seguimiento
  - 3.5.3. Visualización de las métricas



## Capítulo 4

# Ensayos y resultados

En este capítulo se detallan los resultados esperados y obtenidos sobre cada etapa del trabajo. A su vez, se indican las herramientas y metodologías empleadas en cada caso. Finalmente se expone el caso de uso completo integrando todos los componentes que integran al sistema.

### 4.1. Descripción del banco de pruebas

En esta sección se detallan las herramientas y metodologías empleadas para evaluar cada etapa del sistema. El sistema se evalúa en las siguientes etapas:

- Mdelos de detección y seguimientos de personas.
- Extractor de características.
- Manejo de occlusiones y re-identificación de personas.

Para poder evaluar la detección, el seguimiento y la re-identificación de personas, se somete al sistema a una validación visual mediante los videos procesados por cada parte del sistema, como se puede observar en la figura 4.1. En el sección 4.2 se detallan los criterios de validación empleados para evaluar estos videos.

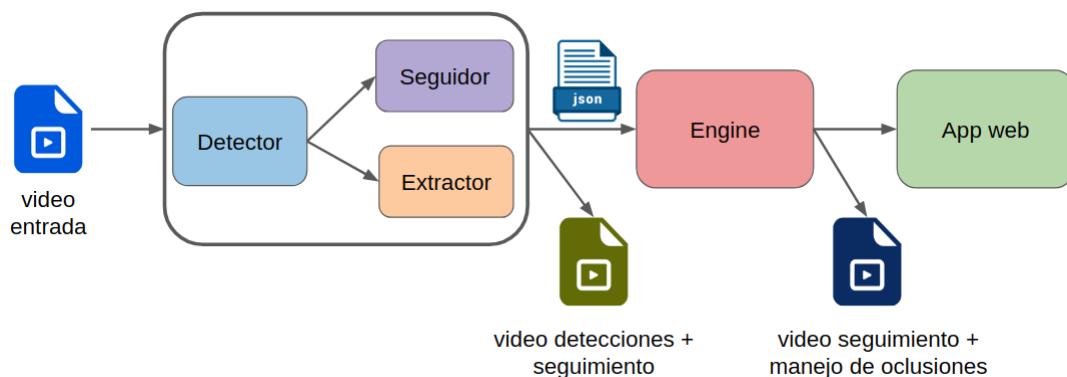


FIGURA 4.1. Videos de ensayo generados en cada etapa.

Para poder evaluar la calidad del extractor de características, se creó un programa que facilita la captura de imágenes de personas en un video. En la figura 4.2 se pueden observar ejemplos de capturas realizadas de una misma persona en diferentes poses y ubicaciones a fin de generar un pequeño banco de imágenes de cada individuo. En el sección 4.3 se detalla los criterios de validación empleados para evaluar el extractor con estas imágenes.



FIGURA 4.2. Capturas tomadas de una personas.

De la experiencia obtenida de evaluar videos y el extracto de características, se desarrollo un programa capaz de automatizar los ensayos de validación del sistema. En la sección 4.4 se detalla como se utilizó este programa para obtener métricas automáticas de la precisión de cada etapa y del sistema global.

## 4.2. Validación de los modelos de detección y seguimiento

Validar sistemas de monitoreo por visión o procesamiento de imágenes normalmente implica una inspección manual por parte de un experto (conocedor del sistema y los objetivos a lograr) ya que es un proceso muy complejo de automatizar. El proceso empleado consiste en:

- Analizar si hay una zona de la cámara en donde las personas no estén siendo detectadas.
- Analizar si hay problemas de iluminación o de foco en los videos capturados.
- Evaluar si las detecciones son continuas, o si las personas están constantemente cambiando de estado de detectada a no detectada.
- Evaluar si hay falsos positivos, es decir, objetos que no son personas detectadas en el video.
- Evaluar si el seguidor mantiene el mismo identificador para cada persona a medida que transitan por el recinto.
- Detectar posibles occlusiones estáticas en el recinto que perjudiquen al sistema (por ejemplo un cartel que tape a las personas).
- Evaluar con que frecuencia ocurre alguno de los desafíos conocidos en el seguimiento de cada persona (occlusiones, pérdidas o cambios de identificadores, personas que salen y vuelven a ingresar a cámara).

En la figura 4.3 se pueden observar ejemplos de problemas que ocurren con el detector que precisan examinación manual.

Es fundamental esta etapa de análisis porque permite entender las debilidades y fortalezas del sistema en una primera etapa de desarrollo.

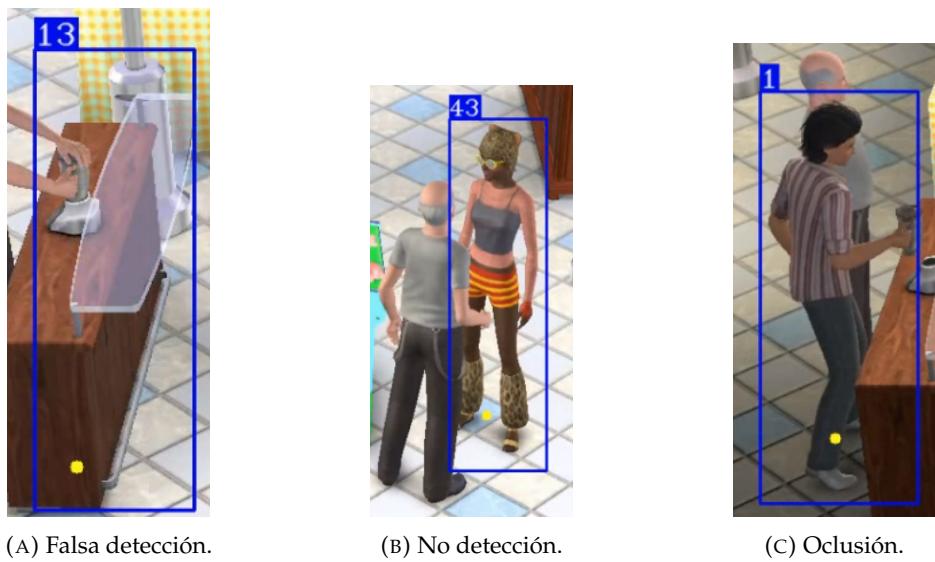


FIGURA 4.3. Fallas en la detección de personas.

### 4.3. Validación del modelo de extracción de características

Para evaluar la calidad del extractor de características no alcanzan las métricas de segmentación obtenidas en la sección 3.2.3. Es necesario llevar el concepto de re-identificación por vectores de características a algo tangible. Para ello, se creó un programa que permite realizar el seguimiento de personas únicamente por sus características, el cual se detalla en el diagrama de la figura 4.4.

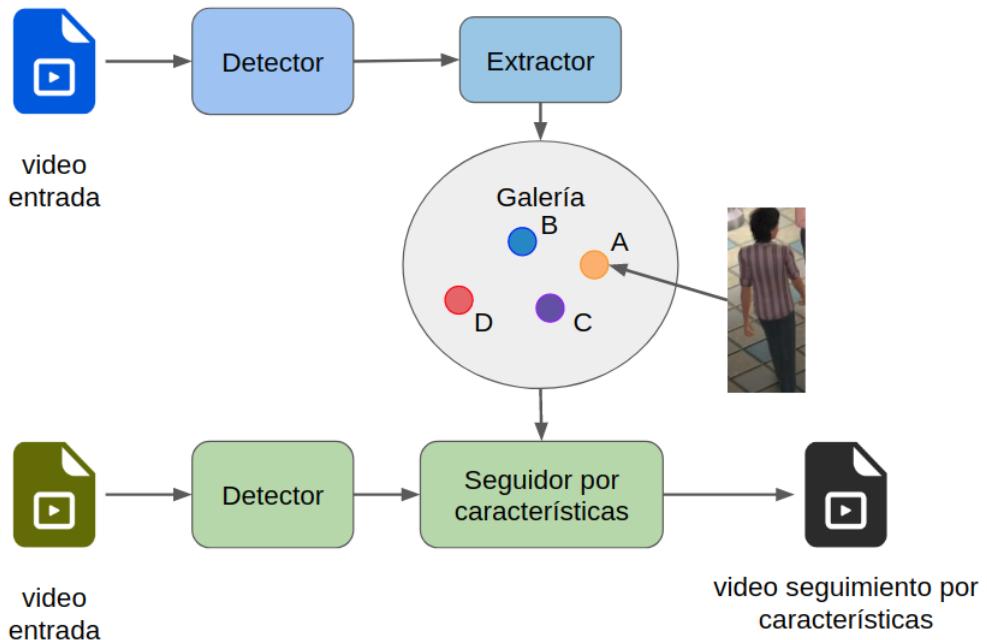


FIGURA 4.4. Programa de seguimiento por características.

El procedimiento llevado a cabo requiere dos pasados del video de entrada, en la primera se obtiene como salida la galería de características y en la segunda se obtiene el video de seguimiento por características.

Primera pasada de video:

- El sistema levanta las capturas de la imágenes de las personas y calcula los vectores de características de cada una.
- Arma una galería de clusters de personas utilizando los vectores calculados.
- Se analizan los clusters de la galería utilizando una herramienta que permite evaluarlos y graficarlos (ver figura 4.5).

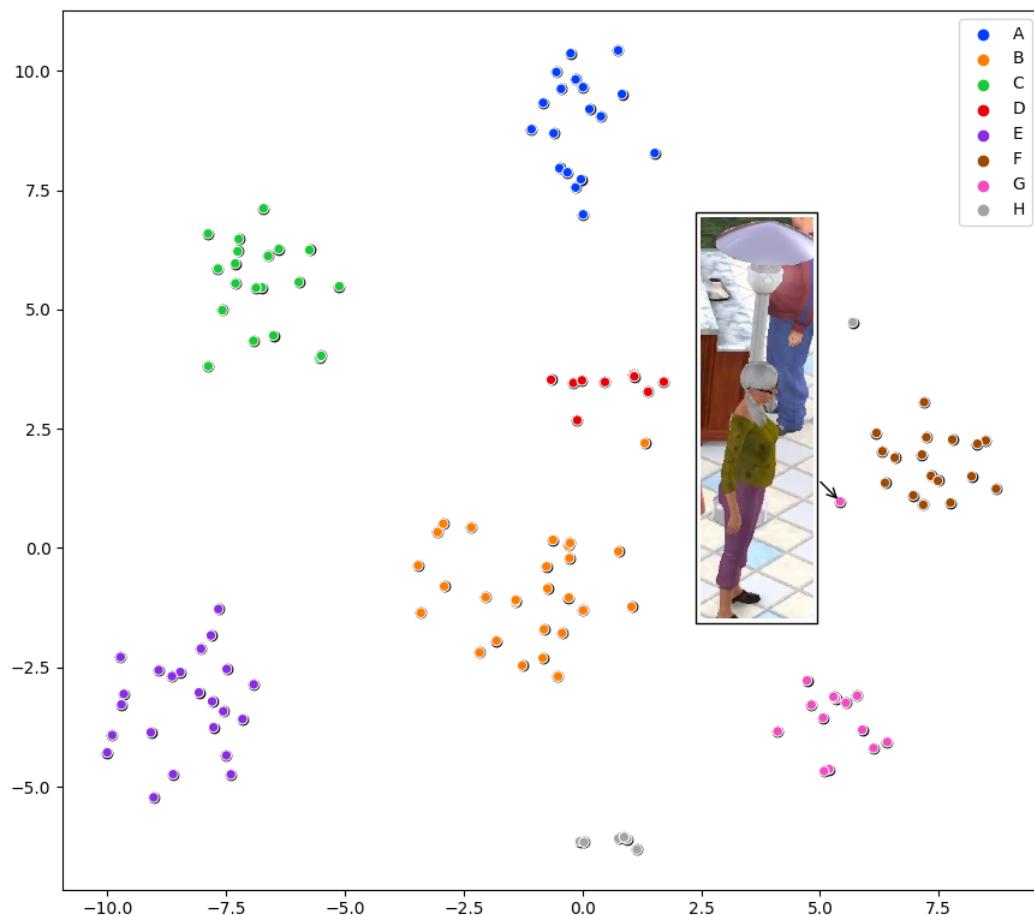


FIGURA 4.5. Análisis de los clusters generados.

La herramienta para analizar clusters permite sacar conclusiones como las siguientes:

- No hay solapamiento de clusters en la galería y los clusters se encuentran bien separados.
- Se puede observar que el cluster "G" tiene un vector muy cercano al cluster "F", debido a que dicho vector se calcula sobre una captura en donde ambas personas aparecen en la imagen.

Segunda pasada de video:

- De cada detección reportada se extrae su vector de características.
- El sistema toma cada vector de características y busca en la galería de clusters de personas a cual pertenece.
- En caso de encontrar una persona candidata para esa detección, le asigna la letra de ese cluster.
- Se genera un video de salida en donde las personas detectadas son representadas por las letras de los clusters candidatos.

La salida del sistema es un nuevo video con las detecciones de las personas. En vez de utilizar el seguidor para asignarles un identificador único a cada persona, se utiliza el procedimiento explicado en la figura 4.4 para asignar una letra a cada detección. En la figura 4.6 se observa un ejemplo del video resultante.

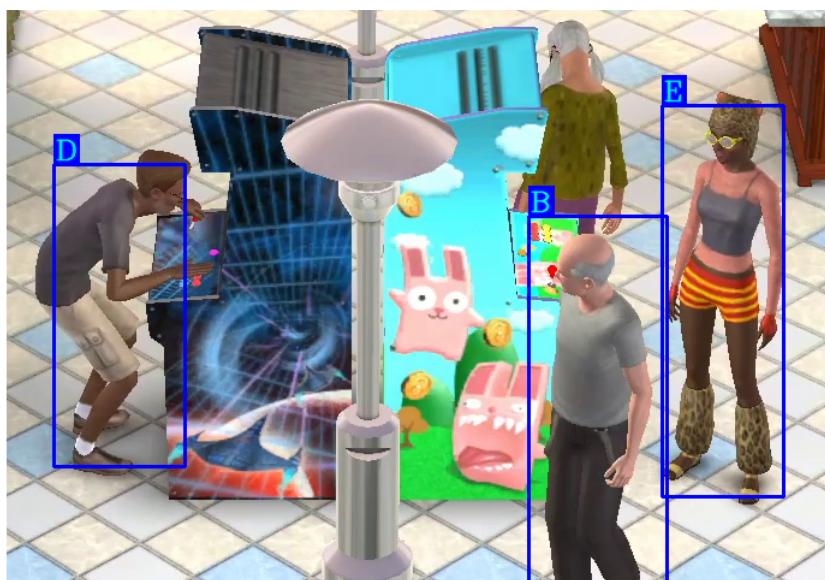


FIGURA 4.6. Sistema de seguimiento por características.

Este proceso es muy robusto porque el sistema en la segunda pasada analiza el video conociendo de ante mano las diferentes poses de cada persona, dado que posee las capturas de ellas que se encuentran transformadas en clusters dentro de la galería, es decir, inicia el proceso con información futura. Conclusión, si la galería está correctamente armada se puede realizar seguimiento de personas por características sin errores. Utilizando este sistema se pueden obtener métricas automáticas de los videos obtenidos a la salida del Engine:

- Cuánto tiempo fue monitoreada cada persona de forma correcta.
- Cuántos identificadores distintos se asignaron a una misma persona.
- Cuántos intercambios de identificadores hay en todo el video.

#### 4.4. Precisión de cada módulo del sistema ensayado

Utilizando las métricas automáticas obtenidas en el capítulo anterior se puede validar por cada video procesado los siguientes requerimientos:

- Se considerará que una persona es correctamente monitoreada si al menos se mantuvo su seguimiento el 80 % del tiempo que circuló en el recinto.
- Se considerará que el sistema funciona dentro de los parámetros aceptables si entre el 80 % y 100 % de las personas en el video fueron correctamente monitoreadas.

En la figura 4.7 se observa el porcentaje de correcto seguimiento correspondiente a cada persona en el video luego de toda la cadena de procesamiento.

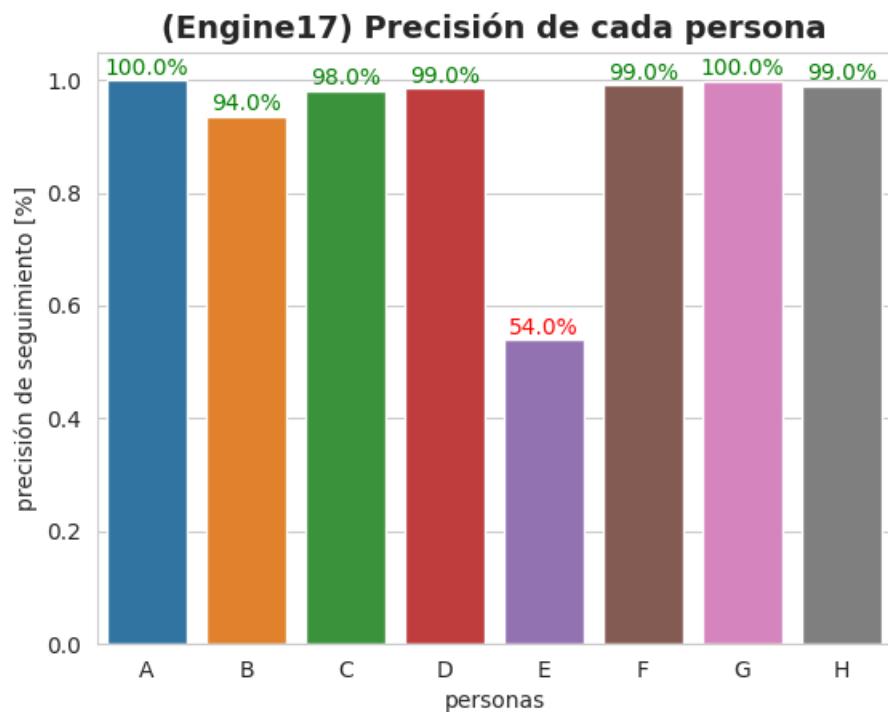


FIGURA 4.7. Precisión de seguimiento por personas.

En la 4.7 se observa que siete de ocho personas tienen una precisión de seguimiento por arriba del 80 %, por lo tanto, la precisión total del sistema para ese video es de 87.5 %.

En la figura 4.8 se aplica la misma lógica para calcular la precisión de cada etapa del sistema y sus diferentes versiones para un mismo video de entrada.

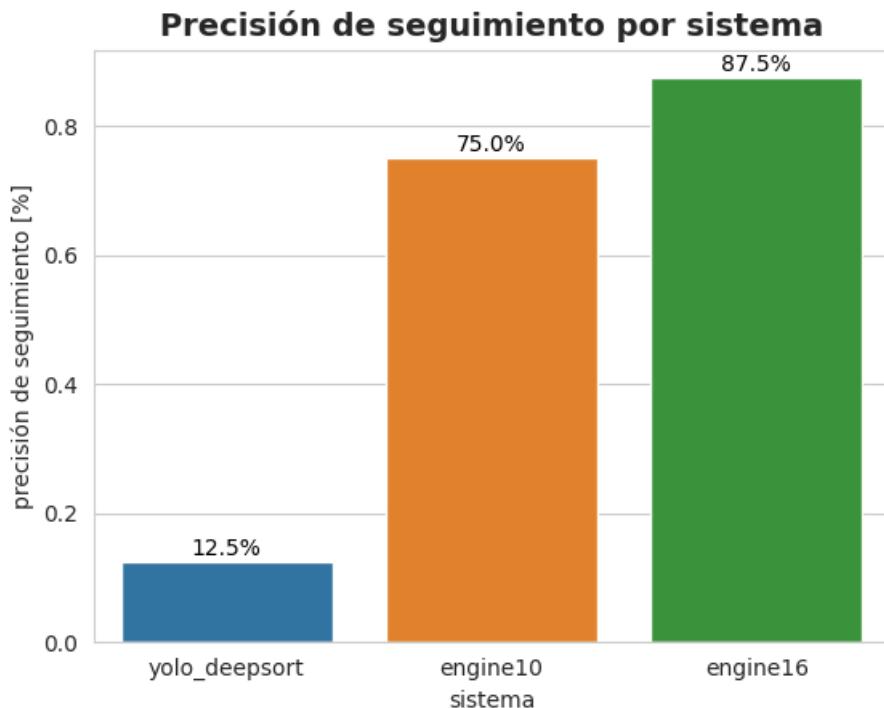


FIGURA 4.8. Precisión de seguimiento por sistemas.

Descripción de la evolución de los sistemas de la figura 4.8:

- Yolo y Deepsort: Precisión que se alcanza si se utiliza el detector Yolo y el seguidor Deepsort tal cual se encuentra disponible en la web.
- Engine10: Precisión que se alcanza si al detector Yolo y el seguidor Deepsort se agrega el postprocesado del Engine v1.0 utilizando los vectores de características para el manejo de occlusiones y re-identificación de personas.
- Engine16: Precisión que se alcanza si al detector Yolo y el seguidor Deepsort se agrega el postprocesado del Engine v1.6 utilizando los vectores de características para el manejo de occlusiones y re-identificación de personas junto con la información agregada de las zonas de interés.

En conclusión, la precisión del sistema aumenta con cada nueva funcionalidad que se incorpora. Por otro lado, en la figura 4.9 se observa que a medida que evolucionan los sistemas también se reducen los intercambios de identificador entre personas.

Con la última versión del Engine se logra alcanzar la precisión deseada de seguimiento, ya que con el agregado de las zonas se redujeron las falsas detecciones e intercambios de identificadores.

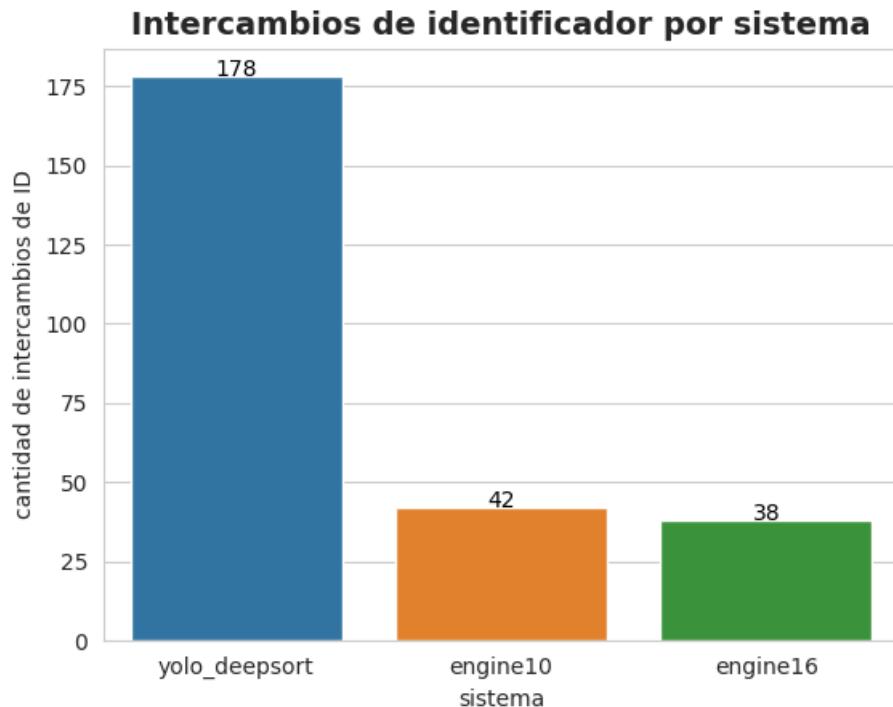


FIGURA 4.9. Intercambios de identificador por sistemas.

## 4.5. Simulaciones

Durante la realización de este trabajo no se tomaron muestras de videos de tiendas o recintos para realizar ensayos y validación, sino que se contó con el material de video disponible en internet. El problema es que la mayoría del material disponible es de muy corta duración (menor al minuto) o circulan muy pocas personas por el recinto (dos a tres personas). Debido a estos problemas fue muy difícil al comienzo validar las primeras etapas del trabajo.

El único video que se encontró en internet (youtube) que cumpliera con todas las características fue un video de una tienda [1] en donde aparecen más de 10 personas por más de un minuto. Este video se utilizó para las primeras demostraciones [2] para validar el sistema de re-identificación y manejo de occlusiones. El inconveniente con este video es que no resultó posible probar todas las funcionalidades de las zonas de interés, debido a que el espacio es bastante reducido y las personas durante la duración del video no llegan a transitar por más de una zona; como se observa en la figura 4.10.

Dado que se necesitaba un video de un recinto con más zonas de interés y personas que se involucraran más con el espacio se buscó simular el entorno. Para ello, se optó por utilizar el video juego "Los Sims 3"[3], en el cual es posible crear un ambiente totalmente a medida y personajes que interactúan en el. El juego permite crear personajes con diferentes personalidades por lo que cada uno interactúa de forma singular con el espacio pudiéndose obtener distintas métricas por el tiempo que se desee que dure el ensayo. En la figura 4.11 se observa el entorno creado para ensayar las últimas funcionalidades creadas en el sistema, como por ejemplo las zonas de interés, las métricas y la aplicación web.



FIGURA 4.10. Imagen del video de tienda utilizado al comienzo el trabajo.



FIGURA 4.11. Imagen del escenario creado en el video juego "Los Sims 3".

El video generado a partir del simulador fue un éxito y permitió terminar de ensayar y validar todas las funcionalidades del sistema sin inconvenientes, pero no fue fácil de elaborar. En los primeros videos de simulación el detector a penas conseguía capturar a los personajes en el video, dado que la resolución de grabación y la calidad de las luces y sombras era baja. Se necesitó elaborar diez configuraciones diferentes de captura de video para que los colores y los personajes se vieran lo más realista posible y el detector pudiera identificar a los personajes del juego como personas reales. El resultado final [4] se acerca a un escenario real, ya que el detector presenta más inconvenientes detectando a los personajes que a las personas reales en la tienda y es por ello que la precisión de seguimiento utilizando solo el detector y seguidor para este video es muy baja. Finalmente, el video generado como una simulación pudo utilizarse luego de encontrar la configuración de video adecuada y gracias a que el sistema de re-identificación ayuda a salvar todos los defectos del detector en el seguimiento de estos personajes ficticios.

## 4.6. Resultados

En esta sección se detallan los resultados y métricas obtenidas en la aplicación web luego de procesar el video de los Sims con el Engine v1.6. En la aplicación web las detecciones y las personas se ejemplifican con figuras geométricas, ya que no se realiza la transmisión del video original en la misma; como se observa en la figura 4.12.

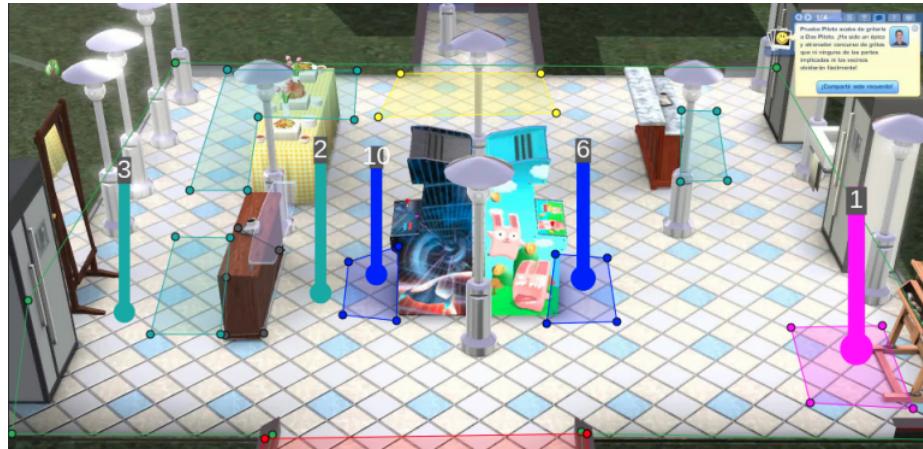


FIGURA 4.12. Imagen del monitoreo de personas en la aplicación web.

Además de contar con el seguimiento y monitoreo de cada persona en la aplicación web, a la izquierda se cuenta con un panel que se actualiza en tiempo real en el cual se detallan las zonas con las cuales interactúo cada persona del recinto, como se observa en la figura 4.13.

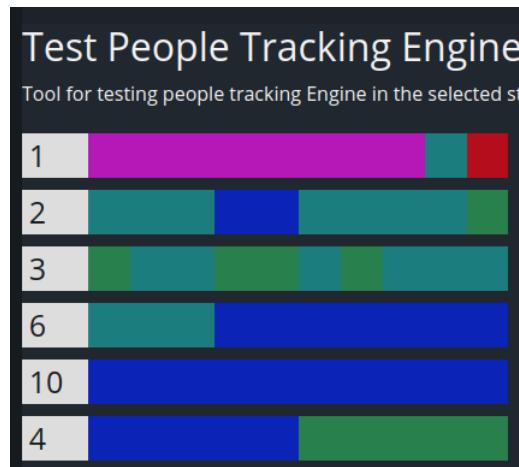


FIGURA 4.13. Interacción de las personas con cada zona dibujada de interés.

La aplicación web también arroja métricas sobre el recinto, ofrece en una tabla cuales fueron las zonas más transitadas o más populares como también el mapa de calor del recinto (por donde caminaron las personas), como se puede observar en la figura 4.14 y 4.15 respectivamente.

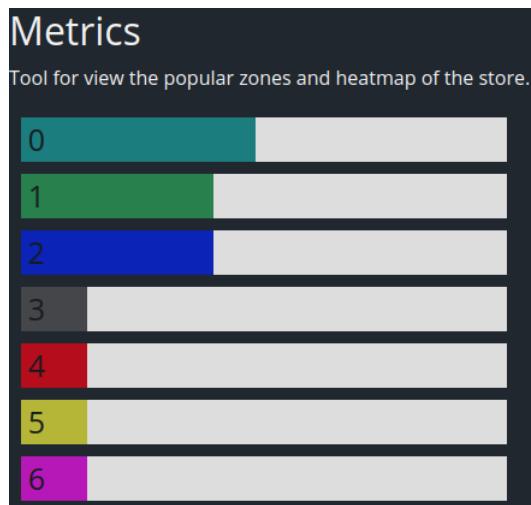


FIGURA 4.14. Métricas de las zonas de interés del recinto.

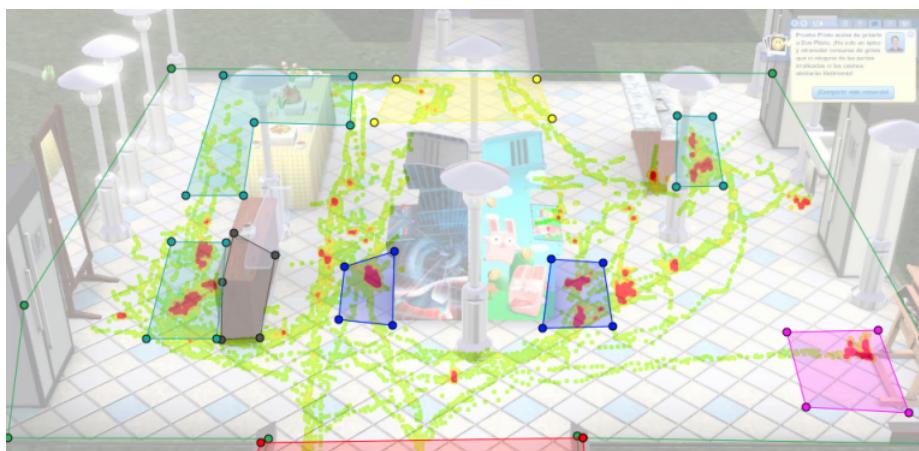


FIGURA 4.15. Mapa de calor del recinto.

De los resultados obtenidos se puede llegar a la conclusión que las zonas más visitadas del recinto son las zonas de comida y bebida (color cyan) y las zonas de juegos (color azul).



## Capítulo 5

# Conclusiones

### 5.1. Resultados obtenidos

Se desarrolló e implementó con éxito un sistema de monitoreo de personas aplicando conocimientos adquiridos a lo largo de todo el año de la Especialización de Inteligencia Artificial. Se alcanzó a cumplir el objetivo propuesto, el cual consistía en estudiar los movimientos que realiza una persona al ingresar y transitar un espacio durante al menos el 80 % del tiempo que la persona permanece en el espacio.

A continuación se listan los logros destacados del trabajo final:

- Cumplir con la planificación original.
- Alcanzar las métricas pautadas para el sistema de seguimiento.
- Entrar un modelo propio basado en OsNet.
- Combinar técnicas de aprendizaje profundo con técnicas de aprendizaje automático para re-identificación de personas.
- Re-identificar a las personas que salen de cámara o son ocluidas durante un tiempo considerable.
- Generar material en video simulado para ensayar el sistema.

A continuación se resaltan aquellas materias de mayor relevancia para este trabajo:

- Gestión de Proyectos: la elaboración de un plan de proyecto para organizar el trabajo final, facilitó la realización del mismo y evitó demoras innecesarias.
- Análisis de datos: el análisis de los datos de entrenamiento y su preprocesamiento permitieron mejorar los resultados de entrenamiento.
- Aprendizaje automático: el uso de técnicas de segmentación y clasificación para la re-identificación de personas.
- Visión por computadora I y II: la experiencia adquirida en el uso de algoritmos y modelos de detección de objetos fueron vitales para la realización de este trabajo.

## 5.2. Trabajo futuro

Utilizando la experiencia adquirida en la realización de este trabajo se encontraron diferentes aspectos de mejora del prototipo, necesarios para que se convierta en un producto comercial:

- Mejorar el modelo entrenado de OsNet utilizando *triple-loss* y datos de diferentes orígenes.
- Evaluar la utilización de *Tensorflow Real Time* para mejorar los tiempos de cómputo de los modelos de IA.
- Evaluar que tipo de dispositivo podría ejecutar el sistema completo en tiempo real.
- Evaluar el uso de cámaras de mayor ángulo visual para capturar más espacio del recinto.
- Evaluar incorporar modelos de detección de pose a fin de obtener más información de la actividad que desarrollada la persona en el recinto.

Durante la utilización de la interfaz del sistema también se encontraron aspectos a mejorar en cuanto a la usabilidad:

- Crear una base de datos para almacenar las métricas obtenidas en cada ensayo o ejecución.
- Disponer la posibilidad de visualizar varias cámaras o recintos desde la misma interfaz.
- Desarrollar una interfaz reducida para su visualización en dispositivos móviles.
- Elaborar alertas programables relativas a que está sucediendo en el local.

# Bibliografía

- [1] HDCCTV Cameras. *HD CCTV Camera video retail store.* <https://www.youtube.com/watch?v=KMJS66jBtVQc>. Dic. de 2017. (Visitado 25-12-2017).
- [2] Hernán Contigiani. *PBT: Demo de tienda.* [https://youtu.be/LzRwI\\_e-gh0](https://youtu.be/LzRwI_e-gh0). Jul. de 2021. (Visitado 03-07-2021).
- [3] Electronic Arts. *Los Sims 3.* <https://www.ea.com/es-es/games/the-sims/the-sims-3>. Jun. de 2009. (Visitado 02-06-2009).
- [4] Hernán Contigiani. *PBT: Demo de Los Sims.* <https://youtu.be/6CvBsHmHaSI>. Jul. de 2021. (Visitado 03-07-2021).