



Sistema de predicción adaptativa con análisis emocional y respuesta autónoma para la gestión de aglomeraciones mediante visión por computador e internet de las cosas(IoT)

Bedón Guevara, Roberto Alexander

Departamento de Ciencias de la Computación

Carrera de Software

Trabajo de integración curricular, previo a la obtención del título de Ingeniero en Software

Gualotuña Álvarez Tatiana Marisol, Ph.D.

12 de abril de 2026



CERTIFICADO DE ANÁLISIS
magister

Bedón_Roberto___Trabajo_TitulaciónA ctualizado (1) (1)

9%
Textos
sospechosos

1% Similitudes
< 1 % similitudes entre
comillas
0 % entre las fuentes
mencionadas

4% Idiomas no reconocidos
mencionados

5% Textos potencialmente
generados por la IA

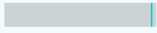
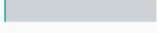
Nombre del documento: Bedón_Roberto___Trabajo_TitulaciónActualizado (1) (1).docx
ID del documento: 2cf8957b0450fceb7fe3a20d2f9db7d6e66cfae
Tamaño del documento original: 11,93 MB

Depositante: TATIANA MARISOL GUALOTUÑA ALVAREZ
Fecha de depósito: 10/2/2026
Tipo de carga: interface
fecha de fin de análisis: 10/2/2026

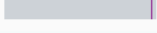
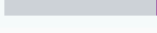
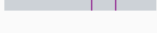
Número de palabras: 16.449
Número de caracteres: 108.242

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	 Documento de otro usuario #Bec03c Viene de de otro grupo 4 fuentes similares	< 1%		 Palabras idénticas: < 1% (55 palabras)
2	 hdl.handle.net Prototipo de sistema embebido para la detección de obstáculos ... https://hdl.handle.net/10495/41556	< 1%		 Palabras idénticas: < 1% (23 palabras)
3	 Sistema de monitoreo basado en drones_V9.docx Sistema de monitor... Viene de de mi grupo	< 1%		 Palabras idénticas: < 1% (22 palabras)

Fuentes con similitudes fortuitas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	 people.ece.ubc.ca https://people.ece.ubc.ca/~minchen/min_paper/2024/2024-IEEE-TSC-RT3C.pdf	< 1%		 Palabras idénticas: < 1% (22 palabras)
2	 Documento de otro usuario #c1f977 Viene de de otro grupo	< 1%		 Palabras idénticas: < 1% (26 palabras)
3	 repositorio.uchile.cl Detección y monitoreo de deformaciones en túneles viales... https://repositorio.uchile.cl/handle/2250/202587	< 1%		 Palabras idénticas: < 1% (20 palabras)
4	 pubmed.ncbi.nlm.nih.gov Semi-automated CCTV surveillance: the effects of sys... https://pubmed.ncbi.nlm.nih.gov/22704458/	< 1%		 Palabras idénticas: < 1% (18 palabras)
5	 Documento de otro usuario #35b202 Viene de de otro grupo	< 1%		 Palabras idénticas: < 1% (21 palabras)

Fuentes mencionadas (sin similitudes detectadas)

Estas fuentes han sido citadas en el documento sin encontrar similitudes.

1

 <https://roboflow.com>

2

 <https://doi.org/10.1016/j.simpa.2022.100433>

Firmado electrónicamente por:
TATIANA MARISOL GUALOTUÑA ALVAREZ
Validar únicamente con FirmaEC

Tatiana Marisol Gualotuña Álvarez PH.D

Director

C.C.: 1711498418



Departamento de Ciencias de la Computación

Carrera de Software

Certificación

Certifico que el trabajo de integración curricular: **“Sistema de predicción adaptativa con análisis emocional y respuesta autónoma para la gestión de aglomeraciones mediante visión por computador e internet de las cosas (IoT)”** fue realizado por el señor Bedón Guevara Roberto Alexander, el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 12 de abril de 2026



Tatiana Marisol Gualotuña Álvarez PH.D

Director

C.C.: 1711498418



Departamento de Ciencias de la Computación

Carrera de Software

Responsabilidad de Auditoría

Yo, Bedón Guevara, Roberto Alexander, con cédula de ciudadanía n°1751865823, declaró que el contenido, ideas y criterios del trabajo de integración curricular: **“Sistema de predicción adaptativa con análisis emocional y respuesta autónoma para la gestión de aglomeraciones mediante visión por computador e internet de las cosas (IoT)”** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 12 de abril de 2026



Bedón Guevara, Roberto Alexander

C.C.: 1751865823



Departamento de Ciencias de la Computación

Carrera de Software

Autorización de Publicación

Yo, Bedón Guevara, Roberto Alexander, con cédula de ciudadanía n°1751865823, autorizó a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **“Sistema de predicción adaptativa con análisis emocional y respuesta autónoma para la gestión de aglomeraciones mediante visión por computador e internet de las cosas (IoT)”** en el Repositorio Institucional, cuyo contenido, ideas y criterios son mi responsabilidad.

Sangolquí, 12 de abril de 2026



Bedón Guevara, Roberto Alexander

C.C.: 1751865823

Dedicatoria

Agradezco a Dios quien ha sido mi pilar fundamental y mi fuente de sabiduría y conocimiento, me ha guiado en cada paso de este camino con fortaleza. Sin su bondad y amor infinito, este logro no habría sido posible.

A mis padres, quienes con su amor y sacrificio diarios han hecho posible este sueño, por nunca dejar de creer en mí, incluso en los momentos más difíciles, y por ser la motivación que me impulsa a seguir adelante.

Gracias Pa por cada uno del día de trabajo y sacrificio, por tu ejemplo de responsabilidad, fortaleza y constancia. Tus consejos, tu disciplina y tu forma de salir adelante me enseñaron el valor del esfuerzo y la importancia de no rendirse ante las dificultades. Tu fuiste mi apoyo en mi formación y en cada uno de los pasos que he dado para adquirir este logro. Gracias a mi mami Liz que aun en la distancia me brindo amor de madre y apoyo incesante. Tu fe, tus palabras de ánimo y tu comprensión fueron mi fuente de motivación. En los momentos de duda y cansancio, tus palabras y tu cariño me dieron la fuerza necesaria para seguir adelante.

A mi Mami Mary, que, aunque es mi abuelita, ha sido para mí una verdadera madre. Gracias por criarme con amor, por estar conmigo en el día a día, por cuidarme, guiarme y acompañarme en cada fase de mi vida. Tu dedicación, compromiso y sacrificio fue esencial para yo alcanzar este gran paso. Este logro también es completamente tuyo, porque sin tu apoyo y tu amor incondicional, este logro no hubiese sido posible.

También a Andreita y a mis abuelitos, gracias por su apoyo, sus consejos y su cariño. Su presencia, sus palabras de ánimo y ejemplo de unión familiar siempre fueron importantes en

mi camino. Cada uno de sus gestos de apoyo me recuerdan que nunca estuve solo y que siempre conté con ustedes

Finalmente, a toda mi familia, quienes siempre han estado para mí en los momentos buenos y malos a lo largo de mi vida. A mis padres, hermanos, tíos y amigos, gracias por su fe depositada en mí, por sus palabras de aliento y por estar siempre presentes. Su amor y apoyo ha sido mi fuerza para seguir adelante.

Agradecimiento

Quiero resaltar mis agradecimientos a todas las personas que estuvieron a lo largo de la realización de esta tesis. El apoyo, guía y palabras de aliento fueron importantes para llegar a alcanzar este importante logro.

En primer lugar, mi agradecimiento a la apreciada Dra. Tatiana Gualotuña, por su orientación y paciencia conmigo. La dedicación y compromiso fueron fundamentales para la culminación de este proyecto de manera satisfactoria. Gracias por su confianza y por ser un mentor clave en este proceso.

A la Universidad de las Fuerzas Armadas ESPE la cual me brindó una educación de calidad y un ambiente competitivo para mejorar mi desarrollo personal y académico. Por haberme permitido compartir con personas tan admirables que me han acompañado en todo este proceso de estudio. Gracias a todos mis amigos, a Raul y Joel con quienes compartí horas de estudio, llenas de compromiso, pero también de risas. Su amistad, apoyo y compañía fueron importantes para culminar esta etapa en mi vida.

Finalmente, a todas las personas que me han apoyado: familiares, amigos, profesores, y compañeros. Gracias por su amistad, por sus palabras de aliento y por estar presentes en momentos de dificultad. Estoy eternamente agradecido por que formaron parte de mi vida.

Índice de contenidos

Dedicatoria	2
Agradecimiento	4
Índice de tablas	8
Índice de figuras	8
Resumen	10
Abstract	11
Capítulo I	12
Antecedentes	12
Planteamiento del Problema	13
<i>Descripción del problema</i>	13
Justificación	17
Objetivos	18
<i>Objetivo general</i>	18
<i>Objetivos específicos</i>	18
Alcance	19

	10
Capítulo II	20
Marco Teórico	20
<i>Gestión de Multitudes y Seguridad</i>	21
<i>Comportamiento en aglomeraciones</i>	23
<i>Visión por Computador</i>	23
<i>Aprendizaje Profundo y Redes Neuronales Convolucionales (CNN)</i>	24
<i>Reconocimiento de Emociones Faciales (FER)</i>	26
<i>Internet de las Cosas (IoT)</i>	27
<i>Arquitectura de Tres Capas Implementada</i>	28
<i>Computación en el Borde (Edge Computing)</i>	29
<i>Protocolos de Comunicación: MQTT</i>	30
<i>Actuadores y respuesta autónoma</i>	31
Revisión Sistemática de Literatura Ligera	31
<i>Planteamiento del objetivo de búsqueda</i>	32
<i>Criterios de inclusión y exclusión</i>	33
<i>Conformación de grupo de control (GC)</i>	34
<i>Construcción de cadena de búsqueda.</i>	36
<i>Selección de los Estudios Primarios</i>	38
<i>Estudios Primarios Seleccionados</i>	38

	11
<i>Caracterización de los Estudios Primarios</i>	41
Respuestas a las preguntas de investigación	43
Capítulo III	46
Arquitectura del sistema	46
Desarrollo e implementación del sistema IoT	48
Capítulo IV	50
Metodología de procesamiento de Datos	50
<i>Fase 1: Extracción o recolección de datos</i>	50
<i>Fase 2: Preprocesamiento y Augmentación de Datos</i>	54
<i>Fase 3: Entrenamiento del modelo</i>	58
Capítulo V	64
Evaluación del modelo	64
Métricas de evaluación	65
Análisis de Resultados del Modelo	66
Evaluación del sistema	72
Discusión de los resultados	88
Capítulo VI	90
Conclusiones	90
Recomendaciones	91

	12
Trabajos futuros	92
Bibliografía	93

Índice de tablas

Tabla 1. Estudios del grupo de control	35
Tabla 2. Elaboración de la cadena de búsqueda óptima.	37
Tabla 3. Estudios Primarios	39
Tabla 4. Parámetros de entrenamiento	56
Tabla 5. Detalle de Precisión por Escenario	66
Tabla 6. Sumatoria de personas reales vs personas detectadas	67
Tabla 7. Datos de errores y precisión del conteo de personas	69
Tabla 8. Detalle de caras detectadas por escenario	70
Tabla 9. Sumatoria de caras reales vs caras detectadas	71

Índice de figuras

Figura 1. Árbol del problema	16
Figura 2. Fundamentación teórica	22
Figura 3. Arquitectura del sistema IoT	47
Figura 4. Mensaje en formato JSON con los datos del obtenidos	50
Figura 5. Dashboard del sistema	51
Figura 6. Fases de entrenamiento del modelo	52
Figura 7. Conjunto de Datos desde la plataforma Roboflow	53
Figura 8. Creación de la copia del Dataset	55
Figura 9. Fase de preprocesamiento y augmentación de datos en Roboflow	57
Figura 10. Augmentaciones aplicadas al datasets Crowd Density	59
Figura 11. Tamaño del dataset después de la augmentación	60
Figura 12. Instalación de Ultralytics YOLO	62
Figura 13. Configuración de la API	62
Figura 14. Configuración del archivo “data.yaml”	63
Figura 15. Precisión del modelo entrenado	69
Figura 16. Rendimiento del dispositivo Raspberry	75

	14
Figura 17 Gráfico comparativo del número de personas de la muestra	81
Figura 18. Comparación de rostros Reales vs Detectados por HSEmotion	86
Figura 19. Escenario con alta densidad	88
Figura 20. Escenario con basa densidad de personas	89
Figura 21. Análisis de regresión lineal entre el conteo manual y la detección del HSEmotion	90

Resumen

Este trabajo de titulación presenta el desarrollo de un sistema IoT basado en visión artificial para la estimación de densidad poblacional y la detección de emociones en multitudes. La problemática abordada responde a la carencia de herramientas de monitoreo automatizado de bajo costo que permitan interpretar el comportamiento colectivo y los niveles de ocupación en tiempo real, evitando la dependencia de infraestructuras de servidores costosas o intrusivas. Para solucionar este reto, se diseñó una arquitectura de Edge Computing sobre una Raspberry Pi 4, la cual se integran modelos de inteligencia artificial para el conteo de personas y el algoritmo HSEmotion para la estimación de los estados emocionales. La metodología se organizó en una serie de fases que abarcaron desde el diseño del pipeline de detección, y el entrenamiento del modelo, hasta su despliegue en entornos no controlados. La investigación logró materializar un sistema capaz de ejecutar simultáneamente el análisis visual en el borde, consolidando a su vez una plataforma IoT que garantiza la transmisión de datos y la activación de respuestas autónomas ante eventos críticos. Finalmente, la validación funcional comprobó la precisión del prototipo, su eficiencia en el uso de recursos y una capacidad de respuesta inmediata, entregando así una solución tecnológica escalable para la gestión inteligente de la seguridad pública.

Palabras clave: Internet de las Cosas (IoT), Visión Artificial, Detección de Emociones, Edge Computing, Raspberry Pi, Estimación de Densidad, Análisis de Multitudes.

Abstract

This thesis presents the development of an IoT system based on computer vision for estimating population density and detecting emotions in crowds. The problem addressed stems from the lack of low-cost, automated monitoring tools that allow for the interpretation of collective behavior and occupancy levels in real time, avoiding dependence on expensive or intrusive server infrastructures. To overcome this challenge, an Edge Computing architecture was designed on a Raspberry Pi 4, integrating artificial intelligence models for people counting and the HSEmotion algorithm for estimating emotional states. The methodology was organized into a series of phases, ranging from the design of the detection pipeline and model training to its deployment in uncontrolled environments. The research successfully implemented a system capable of simultaneously performing visual analysis at the edge, while also establishing an IoT platform that guarantees data transmission and the activation of autonomous responses to critical events. Finally, functional validation verified the prototype's accuracy, its efficiency in the use of resources, and its immediate response capability, thus delivering a scalable technological solution for the intelligent management of public safety.

Keywords: Internet of Things (IoT), Computer Vision, Emotion Detection, Edge Computing, Raspberry Pi, Density Estimation, Crowd Analysis.

Capítulo I

Antecedentes

La gestión de espacios públicos y la seguridad ciudadana han experimentado una transformación primordial con la llegada de la digitalización, pasando de ser de modelos de gestión basados en la reacción a convertirse en procesos preventivos basados en datos. Hoy por hoy, la supervisión por cámaras, limitada al simple registro de eventos, ha resultado inadecuado para afrontar la dinámica de las ciudades. Para garantizar el bienestar en eventos masivos y áreas de alta concurrencia, surge la necesidad de implementar sistemas que más allá de la vigilancia, sean capaces de interpretar el entorno en tiempo real y permitan a los supervisores intervenir antes de que se produzcan incidentes críticos.

El eje central de esta evolución se encuentra en la integración del Internet de las Cosas con la capacidad de procesamiento en el borde (*Edge Computing*). Este cambio propone analizar los datos cerca de donde ocurren los eventos, en lugar de depender de los servidores remotos. Las tendencias tecnológicas actuales apuntan al diseño de sistemas autónomos que permiten cuantificar la movilidad de las personas y detectar patrones de comportamiento en tiempo real, contribuyendo así a una gestión más eficiente de los recursos de seguridad y a una mejor experiencia para los asistentes.

Entre las tecnologías más prometedoras se encuentra la Computación Afectiva, entendida como el conjunto de técnicas que permiten a los sistemas computacionales inferir estados emocionales humanos a partir de señales observables, así como los algoritmos de detección de emociones faciales, entre los cuales se incluyen modelos como HSEmotion. Esta rama de la Inteligencia Artificial permite trascender el simple conteo numérico de individuos para añadir una capa de comprensión cualitativa esto quiere decir entender el estado anímico

de una multitud. La identificación de los patrones de estrés, pánico o miedo colectivo incorpora una dimensión humana al monitoreo automatizado, lo cual nos da una respuesta mucho más empática y precisa frente a situaciones de emergencia.

Sobre esta base tecnológica, el presente proyecto tiene como objetivo diseñar e implementar un sistema embebido capaz de recolectar, procesar y analizar imágenes en tiempo real mediante hardware accesible, como la Raspberry Pi 4. A través de la integración de algoritmos de estimación de densidad y modelos de reconocimiento de emociones, se propone el desarrollo de una plataforma IoT autónoma que sea capaz de monitorear la dinámica física de un espacio y caracterizar la emoción predominante del entorno, proporcionando información clave para la toma de decisiones estratégicas en la gestión de aforos y seguridad pública.

Planteamiento del Problema

Descripción del problema

El presente trabajo requiere analizar el crecimiento de la comunidad y el aumento de eventos de gran afluencia en campus universitarios y espacios públicos, lo que genera desafíos cada vez mayores para la gestión de la seguridad física. Al analizar zonas de alta concurrencia como las paradas de transporte, auditorios, accesos principales y eventos musicales, se observa una saturación que supera la capacidad del monitoreo manual.

En la actualidad, los sistemas de videovigilancia han tenido avances en términos de calidad y resolución de imagen, sin embargo, estos progresos no han estado acompañados por una evolución equivalente en sus capacidades de análisis autónomo. De acuerdo con Dadashi,

Stedmon y Pridmore (2013), un operador promedio comienza a mostrar una disminución en su eficacia para detectar eventos tras 20 minutos de monitoreo continuo en múltiples pantallas. Esto implica que la mayoría de las cámaras instaladas operan como herramientas de análisis posterior a los incidentes, en lugar de actuar como mecanismos de prevención.

En virtud de lo mencionado, se puede afirmar que existe una desconexión tecnológica, los ojos son las cámaras, pero falta el cerebro que procese esa información en tiempo real. El presente trabajo de titulación busca llenar ese vacío mediante el uso de Edge Computing, analizando no solo cuántas personas hay, sino cuál es la atmósfera emocional del grupo, proporcionando una visión cualitativa que los sistemas actuales ignoran.

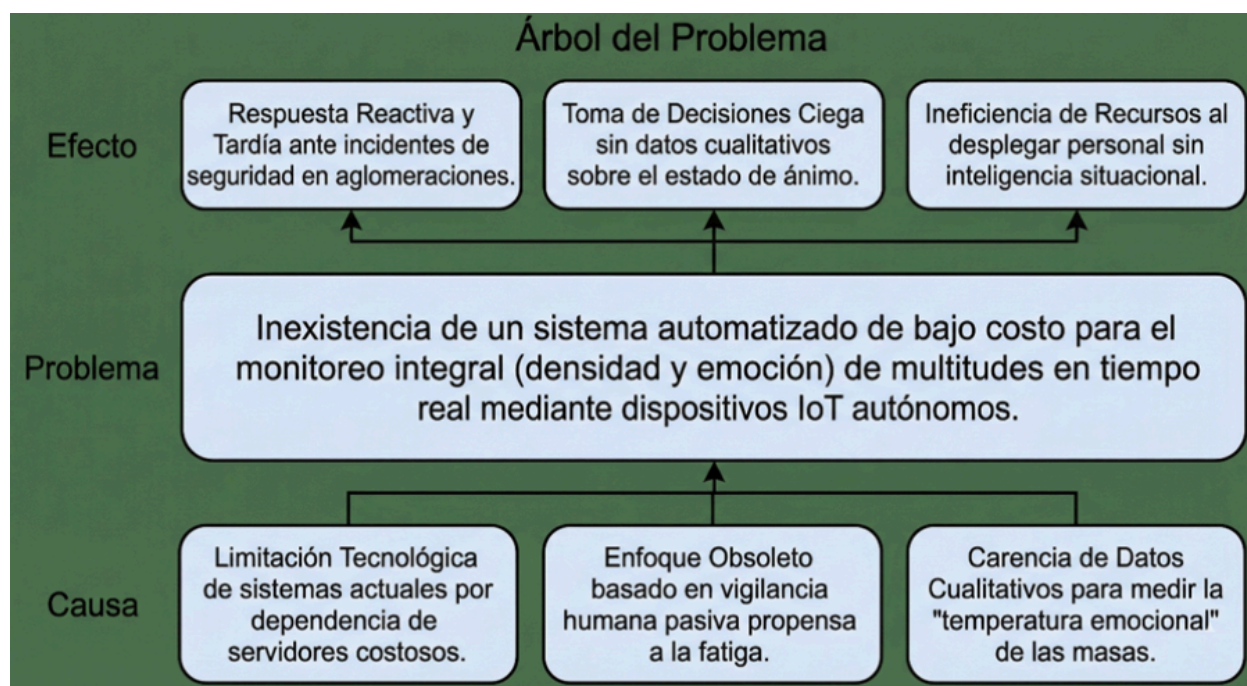


Figura 1. Árbol del problema

Como se observa en la figura 1, el problema medular que enfrenta la gestión de espacios concurridos no es la falta de cámaras, sino la incapacidad de procesar la información que estas generan. Esta carencia de una plataforma automatizada basada en Inteligencia Artificial en el borde desencadena una serie de efectos negativos para la seguridad social.

En primer lugar, la dependencia exclusiva del factor humano para el monitoreo genera brechas de seguridad inevitables. Un operador humano no puede evaluar de manera simultánea la densidad y el estado emocional de cientos de personas en tiempo real. Esta limitación da lugar a una forma de ceguera situacional, en la cual las señales tempranas asociadas a una aglomeración peligrosa o a un conflicto incipiente pasan inadvertidas hasta que evolucionan hacia una emergencia crítica.

En segundo lugar, la barrera de entrada económica de muchas soluciones actuales limita su adopción. Al no aprovechar arquitecturas de procesamiento en el borde, como los dispositivos de bajo costo, muchas organizaciones están obligadas a depender de servidores costosos o de licencias de software. Esta dependencia restringe la implementación de soluciones de seguridad y deja sin cobertura a zonas periféricas que necesitan vigilancia constante y eficiente.

Finalmente, la falta de análisis emocional impide a los tomadores de decisiones de una métrica importante. Conocer únicamente la cantidad de personas presentes resulta insuficiente, por el contrario, una alta densidad acompañada de emociones predominantes como la ira o el miedo modifica los protocolos de respuesta. La carencia de esta información cualitativa obliga a los sistemas de seguridad a tratar todas las aglomeraciones de forma homogénea, disminuyendo la eficacia de las estrategias de evacuación y control de multitudes.

En este contexto, y ante la necesidad de transformar los esquemas tradicionales de vigilancia hacia enfoques más proactivos, se plantea la siguiente pregunta de investigación:

¿Es factible desarrollar un sistema de visión artificial basado en dispositivos IoT capaz de estimar de manera precisa la densidad y los estados emocionales en multitudes, manteniendo un equilibrio adecuado entre rendimiento y costo para optimizar la toma de decisiones en materia de seguridad?

Justificación

La gestión de la seguridad en espacios de alta concurrencia ha evidenciado una brecha crítica entre la capacidad instalada de videovigilancia y la efectividad real del monitoreo. La literatura científica en psicología cognitiva aplicada a la seguridad advierte sobre las limitaciones de la atención humana, investigaciones clásicas como las de Sulman et al. (2008) demuestran que el desempeño de un operador de vigilancia se degrada significativamente tras apenas 20 minutos de observación continua, incrementando la probabilidad de omisión de incidentes críticos por fatiga visual. En la actualidad, muchos sistemas de seguridad se apoyan de enfoques orientados al registro de eventos, pero carecen de la capacidad cognitiva necesaria para interpretar en tiempo real la dinámica del comportamiento humano.

Esta limitación evidencia la necesidad de avanzar hacia un paradigma de vigilancia inteligente basado en Edge Computing. A diferencia de las arquitecturas centralizadas en la nube, que implican mayores costos de transmisión y latencias críticas, el procesamiento local permite una respuesta inmediata ante la detección de anomalías. Como señalan Hassan y Ahmad (2024), delegar el cómputo al borde de la red permite filtrar el ruido de datos reduciendo drásticamente la latencia, un factor que resulta determinante en aplicaciones de seguridad ciudadana.

La integración de algoritmos como HSEmotion en hardware como la Raspberry Pi, representa un gran avance en la implementación de la Inteligencia Artificial. La validación realizada por Savchenko (2022), desarrollador de esta librería, demuestra que el uso de redes neuronales optimizadas para dispositivos de borde permite inferir los estados afectivos con un alto nivel de la precisión, superando las limitaciones de los sistemas tradicionales. Esta capacidad que combina la medición de la densidad física con la caracterización de la atmósfera emocional del entorno, brinda a los administradores una métrica cualitativa para la prevención de desastres en multitudes (Tarasov & Savchenko, 2019).

El valor científico de este trabajo está en demostrar que es viable desplegar sistemas de monitoreo sin incurrir en infraestructuras costosas. Al validar un prototipo de bajo costo capaz de estimar aforos y estados emocionales, se establece una base metodológica replicable en entornos con recursos limitados. Este enfoque se alinea con las tendencias actuales en ciudades inteligentes y en modelos de seguridad ciudadana, como lo señalan Kumar y Singh (2023).

Objetivos

Objetivo general

Desarrollar un sistema basado en visión por computador e Internet de las Cosas (IoT) que permita identificar, contar y predecir el número de personas reunidas en un mismo sitio, integrando análisis emocional y respuestas autónomas para gestionar aglomeraciones en tiempo real, mejorando la seguridad y optimizando la toma de decisiones proactiva en diversos entornos.

Objetivos específicos

- Revisar el estado del arte sobre tecnologías de visión por computador, análisis emocional y sistemas IoT aplicados a la gestión de aglomeraciones.
- Desarrollar un modelo de visión por computador para el conteo de personas, predicción de aglomeraciones y análisis emocional en tiempo real.
- Construir una plataforma IoT que transmita datos en tiempo real y active respuestas autónomas basadas en las predicciones y análisis emocionales.
- Validar el sistema mediante pruebas funcionales que aseguren su precisión, eficiencia y capacidad de respuesta autónoma.

Alcance

El presente proyecto se circunscribe al diseño, desarrollo y validación de un prototipo funcional de Internet de las Cosas (IoT) orientado a la visión artificial, específicamente configurado para operar en Edge Computing. La solución tecnológica se enfoca en la captación y procesamiento de señales de video en tiempo real para la estimación de densidad poblacional y la inferencia de emociones básicas en grupos de personas.

Para la implementación física, el proyecto se fundamenta en el uso de hardware de placa única, seleccionando la Raspberry Pi 4 por su equilibrio entre el bajo costo y alta disponibilidad. A nivel de software, el alcance del proyecto contempla la integración de algoritmos de detección facial y el despliegue del modelo HSEmotion, optimizado para arquitecturas ARM. El objetivo es asegurar una ejecución fluida y en tiempo real, eliminando la dependencia de servicios en la nube para el análisis inmediato de la información.

Por otro lado, la fase de desarrollo y las pruebas de validación se llevarán a cabo en escenarios que exista una aglomeración real. El análisis de los datos se centra en establecer la correlación entre el conteo manual y la detección automática, también clasifica los estados afectivos predominantes como neutralidad, felicidad o ira en los rostros detectados que cumplan con los criterios de calidad visual. Cabe destacar que el alcance del proyecto excluye cualquier forma de reconocimiento facial individual, garantizando así el respeto a los principios de privacidad. Asimismo, el sistema está diseñado como una herramienta de apoyo para la toma de decisiones y la generación de alertas preventivas, y no como un mecanismo de seguridad física activo o de intervención directa. Finalmente, la visualización de los resultados se limitará a un panel de control local accesible dentro de la red, demostrando la capacidad del sistema para entregar métricas de comportamiento colectivo en tiempo real.

Capítulo II

Marco Teórico

En este capítulo se expone el marco teórico de los componentes que se utilizarán para el desarrollo de este trabajo de titulación. Los elementos primordiales que conforman la piedra angular de este proyecto de tesis son los siguientes:

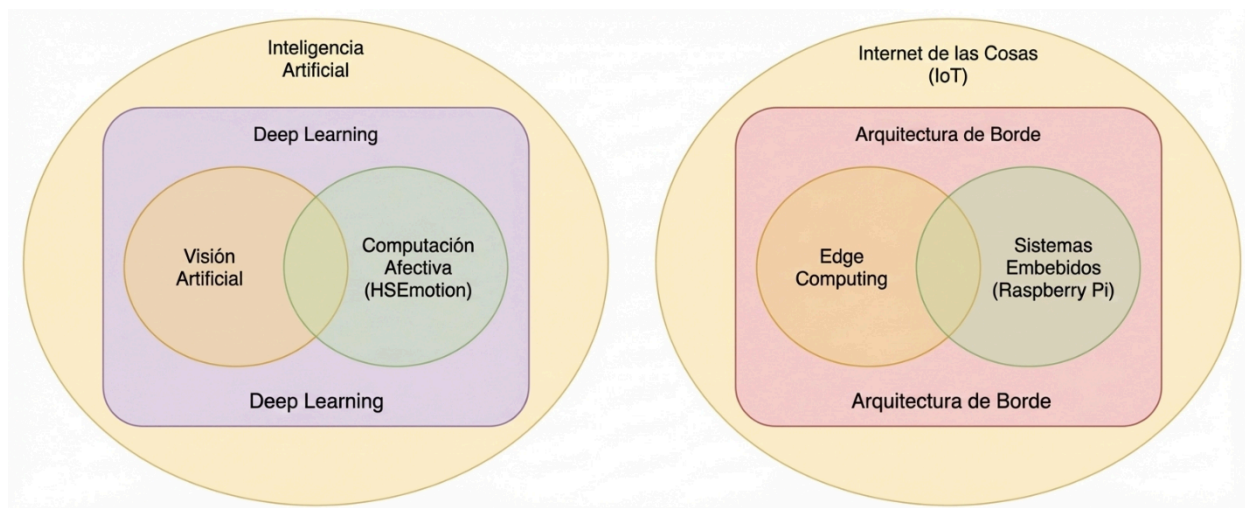


Figura 2. Fundamentación teórica

Gestión de Multitudes y Seguridad

La gestión de multitudes es una de las bases fundamentales para la seguridad ciudadana y la administración de los espacios públicos. La constante realización de eventos masivos y el aumento poblacional exigen sistemas que no solo observen, sino que comprendan la dinámica del movimiento en los grupos humanos.

Para plantear el problema desde un punto de vista de la ingeniería, es primordial cuantificar que es “aglomeración”. Más que un conjunto de individuos, una aglomeración debe entenderse como un entorno físico restrictivo que condiciona directamente la libertad de movimiento de quienes la integran. En este contexto, la densidad peatonal se establece como la variable crítica para analizar el comportamiento del flujo.

Siguiendo los estándares establecidos en el trabajo de Fruin (1971), se identifica un pico de seguridad considerable cuando el espacio disponible por individuo desciende por debajo de los 0,46 metros cuadrados, esto equivale a una concentración superior a 2,17 personas por metro cuadrado aproximadamente. Al cruzar este límite, el flujo de peatones pierde su estabilidad, dando paso a interacciones físicas involuntarias y a una dinámica desordenada característica de la turbulencia en masas.

Estudios posteriores han mejorado estos límites, señalando que, a partir de una densidad de 4 a 5 personas por metro cuadrado, el riesgo para la seguridad es cercano. En el contexto de la presente investigación, la comprensión de estos umbrales son importantes, dado que el algoritmo será entrenado para generar alertas cuando la densidad estimada supere los límites de seguridad establecidos.

Riesgos asociados a las multitudes

Los riesgos de seguridad a causa de la acumulación descontrolada de personas pueden derivarse en tragedias cuando no se gestionan constantemente.

Tal como señala el experto en seguridad G. Keith Still (2014), el peligro en altas densidades no es el pánico irracional, sino las ondas de choque que acontece cuando la densidad alcanza niveles donde las fuerzas físicas se transmiten a través de los cuerpos, en escenarios de este tipo, la causa más común de mortalidad no es el pisoteo, sino la asfixia provocada por la presión que impide la expansión torácica necesaria para la respiración.

Sin embargo, los métodos actuales de supervisión pueden llegar hacer reactivos, por consiguiente, el desarrollo de un sistema basado en IoT permite monitorear estas variables en

tiempo real. Al detectar la formación de las turbulencias visuales, el sistema debe ser el punto de entrada principal para activar protocolos de evacuación, mitigando el riesgo de asfixia.

Comportamiento en aglomeraciones

Desde el punto de vista de Gustave Le Bon (1895) en su investigación “The Crowd: A Study of the Popular Mind”, las aglomeraciones actúan bajo una mente colectiva, donde la responsabilidad deja de ser individual y las emociones son las que predominar. Del mismo modo textos como el “Contagio Emocional” (Hatfield et al., 1993) demuestran que las emociones entre ellas el miedo y la ansiedad se transmiten de forma rápida y automática a través de la gesticulación facial y el lenguaje corporal.

Si bien resulta cierto que una aglomeración puede estar visualmente densa, pero tranquila, un pequeño cambio en el estado emocional colectivo puede significar un desequilibrio grupal en segundos. En este punto se fundamenta la principal contribución del presente trabajo, la integración del análisis emocional mediante técnicas de Inteligencia Artificial. El sistema propuesto permite estimar la emoción grupal, lo que posibilita distinguir entre escenarios de alta densidad peatonal asociados a estados emocionales positivos tal como ocurre en eventos recreativos y situaciones de alta densidad vinculadas a condiciones de riesgo, como evacuaciones fallidas.

Visión por Computador

La etapa inicial obligatoria para los sistemas enfocados en visión artificial es el procesamiento digital de imágenes, encargándose de transformar las señales visuales en datos traducidos por un ordenador.

De acuerdo con Gonzalez & Woods (2018), este desarrollo se define como un conjunto de técnicas algorítmicas aplicadas a imágenes bidimensionales con el objetivo de enriquecer la calidad para la percepción humana y facilitar la búsqueda de información autónoma a cargo de las herramientas de cómputo. El proceso normal comprende etapas secuenciales, iniciando con la adquisición, el preprocesamiento, la segmentación y la extracción de las características.

En el ámbito de esta investigación, el preprocesamiento es de vital importancia. Debido a que las cámaras IoT estarán comprometidas a condiciones de luz variables en exteriores e interiores, se aplicarán configuraciones adicionales en la etapa del entrenamiento, esto garantiza que los algoritmos posteriores de detección no fallen por sombras o deslumbramientos con los datos de entrada, asegurando una entrada de datos limpia para la red neuronal.

Aprendizaje Profundo y Redes Neuronales Convolucionales (CNN)

Dentro de un procedimiento tradicional de imágenes, un programador define manualmente que buscar, como los bordes, colores y objetos dentro de la imagen, en cambio, el aprendizaje profundo permite que el sistema aprenda estas reglas por sí mismo.

Según LeCun, Bengio y Hinton (2015), el Deep Learning utiliza arquitecturas compuestas por múltiples capas de procesamiento para aprender representaciones de datos con múltiples

niveles de abstracción. En este contexto, las Redes Neuronales Convolucionales se han consolidado como el enfoque principal para aportar con las tareas de visión por computadora. Las redes típicas procesan la información de manera lineal y las CNN aplican operaciones de convolución que permiten analizar la imagen conservando su estructura espacial. Este mecanismo ayuda a facilitar un aprendizaje jerárquico de características, las capas iniciales se especializan en extraer rasgos básicos, como los bordes y las texturas, mientras que las capas más profundas componen la información para reconocer entidades más complejas, como rostros humanos o patrones de densidad en multitudes.

En virtud de lo expuesto, la selección de una arquitectura basada en CNN para este proyecto no es arbitraria. Este tipo de arquitectura resulta especialmente adecuada para abordar la variabilidad intra-clase presente en escenarios de aglomeración, caracterizados por diferencias de vestimenta, postura, estatura y de oclusiones parciales y completas, manteniendo así niveles de precisión que los métodos estadísticos tradicionales no logran alcanzar.

Uno de los desafíos computacionales más representativos en la actualidad es la detección de objetos en tiempo real, el sistema debe identificar que es un objeto, y en que lugar se encuentra en cuestión de milisegundos. Para abordar esta problemática se ha seleccionado la arquitectura YOLO (You Only Look Once).

Tal como explican Terven y Cordova-Esparza (2023) en su revisión de la evolución de estos tipos de modelos, YOLO replanteó el campo al plantear la detección como un problema de regresión único. A diferencia de los modelos de dos etapas como Faster y R-CNN que escanean la imagen varias veces, YOLO divide la imagen en cuadrículas y predice las

coordenadas de los cuadros delimitadores y las probabilidades de clase en una sola pasada de la red neuronal.

La selección de YOLOv8 como motor de detección de objetos responde a una justificación técnica precisa. Su arquitectura, caracterizada por un diseño anchor-free, permite optimizar la carga computacional al tiempo que perfecciona la identificación de objetivos de escala reducida, una capacidad indispensable cuando se analizan multitudes desde planos distantes donde el detalle es mínimo. En términos de rendimiento, este modelo logra establecer una sinergia ideal entre la velocidad de flujo (FPS) y la precisión métrica (mAP). Esta eficiencia es el factor clave que habilita al sistema IoT para ejecutar el procesamiento de imágenes in situ (Edge Computing), otorgándole total autonomía operativa sin depender de infraestructuras de cómputo externas.

Reconocimiento de Emociones Faciales (FER)

El reconocimiento de emociones faciales es un campo de la visión artificial encargado de analizar las características faciales para deducir el estado afectivo de un individuo. Para que este proceso funcione correctamente debe combinar fundamentos de la psicología conductual con los algoritmos de clasificación de patrones.

Para que un algoritmo logre clasificar una emoción con éxito, es indispensable contar primero con un marco taxonómico que estandarice lo que es una "emoción universal". En este contexto, la investigación se relaciona con la teoría neuro-cultural establecida por Paul Ekman (1992), la cual describe la existencia de seis emociones fundamentales las cuales son, alegría, tristeza, ira, miedo, sorpresa y desagrado, estas trascienden las barreras culturales y lingüísticas debido a su origen biológico innato. Desde el punto de vista de la informática, esto expresa que cada estado emotivo no es una entidad abstracta, sino que se refleja físicamente

mediante la activación de los grupos musculares faciales, como la contracción del cigomático relacionado a la expresión de alegría o la tensión del corrugador asociado a la ira.

El sistema propuesto admite este modelo como la principal lógica de la inferencia. La idea es identificar patrones relacionados con emocionales universales, con especial enfoque en la detección de estados como miedo e ira, que funcionan como señales tempranas de riesgo en escenarios de aglomeración.

Según Savchenko (2021), en la actualidad es posible alcanzar estándares de precisión competitivos empleando redes ligeras como MobileNetV3 o EfficientNet. Al aprovechar técnicas de aprendizaje por transferencia sobre datasets masivos como AffectNet, estos modelos son capaces de extraer características faciales complejas utilizando solo una fracción del costo computacional requerido por arquitecturas más profundas.

Considerando lo expuesto previamente, la elección de la librería HSEmotion para este proyecto no es aleatorio, sino que responde a su capacidad de optimización extrema. A diferencia de enfoques que dependen de hardware gráfico como la GPU, este modelo optimiza los pesos de la red para operar eficientemente en CPUs moderados. Esto es clave porque permite evaluar el clima emocional de la multitud cuadro a cuadro sin estrangular el procesamiento, garantizando una latencia mínima necesaria para que el sistema opere con autonomía en tiempo real.

Internet de las Cosas (IoT)

El desarrollo del sistema propuesto se basa en el paradigma del Internet de las Cosas (IoT), uniendo un problema del mundo físico como son las aglomeraciones, con una solución digital mediante el análisis de datos y la comunicación de los dispositivos interconectados.

El internet de las cosas no es solo una rama de la tecnología, mas bien es una arquitectura estructurada por capas. Según la Recomendación Y.2060 de la Unión Internacional de Telecomunicaciones (ITU-T), el Internet de las Cosas (IoT) constituye una infraestructura global orientada a la sociedad de la información, capaz de habilitar servicios avanzados a través de la interconexión física y virtual de objetos, sustentada en tecnologías de información y comunicación interoperables.

Arquitectura de Tres Capas Implementada

Para garantizar la escalabilidad y la comunicación entre los componentes en sistemas de monitoreo, la literatura técnica recomienda seguir el modelo de referencia de IoT, el cual se estructura en tres capas, la de percepción, la de red, y la de aplicación (ITU-T, 2012).

La capa de percepción y procesamiento en el borde es la responsable de interactuar con el entorno físico. En sistemas de visión artificial modernos, esta capa no solo ayuda a capturar datos de imágenes y videos, sino que realiza el procesamiento *edge computing* el cual es un paradigma tecnológico que procesa los datos cerca de la fuente como sensores, dispositivos o cámaras en lugar de depender de servidores centrales o de la nube.

La función de estos dispositivos embebidos es capturar la señal de video y ejecutar algoritmos de detección, para evitar la saturación de la red, es buena práctica realizar un filtrado de datos y enviar únicamente metadatos procesados en lugar del video en crudo.

Por otro parte, la capa de red y transporte actúa como un puente de comunicación, el objetivo principal es la serialización y transmisión de los datos. En entornos con ancho de banda limitado, se priorizan protocolos como MQTT que está enfocado en el modelo publicación/suscripción que desacopla al emisor del receptor, fundamental para mantener la integridad del sistema ante cualquier inestabilidad de la red.

La información relevante en este protocolo por lo general se encapsula en formatos ligeros como JSON, permitiendo una lectura estándar por cualquier plataforma receptora.

La capa de aplicación es donde los datos se transforman en información útil para el usuario final, las aplicaciones modernas pueden suscribirse a los flujos de datos IoT para actualizar los paneles de control en tiempo real mediante el uso de los protocolos websockets.

Computación en el Borde (Edge Computing)

Uno de los desafíos principales en el procesamiento de video es la latencia y el ancho de banda. En lugar de enviar video pesado a un servidor remoto, este proyecto adopta el enfoque de Edge Computing.

Tal como lo describe Satyanarayanan (2017), uno de los principales referentes en este ámbito, el Edge Computing se define como un paradigma que permite la realización de tareas de computación y almacenamiento de datos en el propio dispositivo o en proximidad a la ubicación física del usuario o de la fuente de datos. El Cloud Computing se determina por centralizar la capacidad de procesamiento en centros de datos remotos.

Este paradigma se desarrolla mediante el uso de placas de desarrollo, como la Raspberry Pi 4, que está configurada para operar como nodo de borde en los sistemas IoT. Este enfoque resulta relevante por la disponibilidad de la ejecución local de modelos de visión por

computadora, como YOLO y HSEmotion, reduciendo la dependencia de conexiones a la red. El procesamiento se realiza directamente sobre el dispositivo, permitiendo la extracción de información relevante como el número de personas y el estado emocional predominante y el descarte posterior del flujo de video, lo que contribuye tanto a una menor latencia del sistema como a la preservación de la privacidad de los usuarios.

Protocolos de Comunicación: MQTT

Considerando que los sistemas IoT enfocados al monitoreo en tiempo real necesitan un uso apropiado del ancho de banda y una latencia reducida, no resulta adecuado el empleo de protocolos web tradicionales para la transmisión continua de datos como ejemplo HTTP. En su lugar, se emplean protocolos de mensajería livianas, diseñados específicamente para entornos con recursos limitados.

En este ámbito, el protocolo MQTT se muestra como una solución ampliamente adoptada en la actualidad. De acuerdo con el texto de OASIS (2014), MQTT es un protocolo de mensajería fundamentado en el paradigma de publicación/suscripción, se caracteriza por tener una huella de código reducida y también un bajo consumo de ancho de banda, lo que lo hace especialmente adecuado para dispositivos IoT.

Dentro de una arquitectura típica basada en MQTT, los nodos de borde pueden tomar el rol de publicadores, enviando mensajes organizados en formato JSON que contienen toda la información procesada, mientras que los sistemas de visualización actúan como los suscriptores. Por eso, este modelo de comunicación favorece un diseño asíncrono y desacoplado, permitiendo que el sistema continúe operando incluso ante desconexiones

temporales de los componentes de visualización, lo que incrementa la robustez y disponibilidad del sistema de monitoreo.

Actuadores y respuesta autónoma

El ciclo de funcionamiento de un sistema IoT orientado al control y la gestión de riesgos se completa con la incorporación de mecanismos que permiten intervenir sobre el entorno físico.

En los sistemas de automatización, los actuadores cumplen la función de transformar las órdenes generadas por el sistema en acciones sobre el entorno, ya sea a través de señales visuales o sonoras. En el contexto de los sistemas inteligentes, esta capacidad se traduce en respuestas automáticas que se activan ante la detección de situaciones críticas, permitiendo que el sistema reaccione de manera autónoma sin requerir una intervención humana inmediata.

Este planteamiento resulta adecuado en escenarios con alta concentración de personas, donde la activación de alertas visuales o sonoras puede ayudar a reducir los tiempos de reacción ante situaciones de riesgo, como estampidas o evacuaciones ineficientes.

Revisión Sistemática de Literatura Ligera

En esta sección del presente trabajo se disponen los diferentes documentos de investigación y tesis relacionados con el desarrollo de sistemas embebidos de visión artificial para la estimación de densidad y análisis emocional en multitudes. La Revisión Sistemática de Literatura (SLR) es una táctica metodológica que permite obtener estudios precisos, detallados y confiables, esenciales para adecuar la presente investigación a un contexto científico y

tecnológico apropiado. Por lo cual, hemos planteado las siguientes interrogantes de investigación que guiará el proceso de búsqueda y selección:

RQ1: ¿Qué técnicas de visión computacional y arquitecturas de hardware en el borde (Edge Computing) optimizan el equilibrio entre precisión y latencia para el análisis de emociones y conteo de personas en tiempo real?

RQ2: ¿Qué arquitecturas de redes neuronales y estrategias de optimización se identifican como las más eficientes para ejecutar tareas de visión computacional en entornos de hardware limitado?

Planteamiento del objetivo de búsqueda

Determinar las mejores prácticas y arquitecturas de referencia para evaluar el desarrollo de un sistema IoT autónomo orientado al análisis de comportamiento en multitudes (conteo y emoción) utilizando dispositivos de bajo costo como Raspberry Pi, con el fin de establecer los factores más destacados en la implementación del proyecto como solución de seguridad proactiva.

Este planteamiento tiene como objetivo principal la evaluación de las estrategias más eficientes en el despliegue de modelos de Deep Learning ligeros en el borde, para de esta manera poder identificar los compromisos necesarios entre la carga computacional y la fiabilidad de la inferencia. La utilización de librerías optimizadas como HSEmotion y marcos de trabajo como OpenCV se encuentra incluida en el objetivo para delimitar la búsqueda hacia estudios que demuestren viabilidad operativa en entornos reales, descartando aquellos puramente teóricos o dependientes de grandes centros de datos.

Criterios de inclusión y exclusión

Los criterios llevados a cabo aseguran que la investigación sistemática de literatura contenga únicamente los estudios relevantes con los cuales proporcione información técnica sustancial para responder a la pregunta de investigación. A continuación, se detallan los criterios de inclusión y exclusión aplicados durante el filtrado de documentos.

Serán Incluidos:

- **CI1:** Artículos científicos y papers técnicos que describan el diseño, implementación o evaluación de sistemas de visión artificial embebidos para el análisis de personas
- **CI2:** Estudios que emplean arquitecturas de *Edge Computing* o dispositivos de placa única como la Raspberry Pi o Jetson Nano donde se puede ejecutar modelos de inteligencia artificial.
- **CI3:** Investigaciones o artículos que contengan técnicas de reconocimiento de emociones faciales enfocados a grupos de personas o a entornos poco controlados.
- **CI4:** Artículos de conferencias o revistas científicas entre ellas la IEEE, ACM o Springer.
- **CI5:** Investigaciones publicadas en idioma español o inglés.

Serán Excluidos:

- **CE1:** Estudios que se centren netamente en procesamiento de datos en la nube (Cloud Computing) sin componentes de procesamiento local.

- **CE2:** Investigaciones que utilicen datasets de laboratorio con rostros posados y perfectos, sin abordar los desafíos de oclusión o iluminación variable.
- **CE3:** Estudios que no describan la arquitectura de hardware o las métricas de rendimiento (FPS, latencia, precisión) de la solución propuesta.
- **CE4:** Documentos con fecha de publicación anterior al 01 de enero de 2019, dado el rápido avance de las redes neuronales ligeras.

Conformación de grupo de control (GC)

El grupo de control es un cúmulo de estudios que permiten fijar una base de investigaciones similares con criterios explícitos definidos en la RQ. Su finalidad es facilitar la identificación y relación de palabras clave y terminología relevantes presentes en estudios específicos, las cuales servirán como base para la construcción de la cadena de búsqueda.

En la siguiente tabla se muestran los estudios que conforman el grupo de control.

Código	Título	Términos Relevantes
GC1	Crowd Density Estimation Using Deep-Learning	Crowd Density, Deep Learning, Real-time processing, IoT surveillance,

		Raspberry Pi, Object Detection.
GC2	Automatic emotion recognition using deep neural network	Emotion Recognition, MobileNet, Lightweight Models, Facial Expression, Python Library, Real-time inference.
GC3	Computer Vision and Convolutional Neural Network for Dense Crowd Count Detection	Crowd Counting, Computer Vision, Density Estimation, Occlusion Handling, Convolutional Neural Networks (CNN), Public Safety.
GC4	Deep Learning and the Internet of Things: Applications, Challenges, and Opportunities	Smart Surveillance, IoT Architecture, Anomaly Detection, Embedded Systems, Automated Alerts, Scalability.
GC5	Advancements in Context Recognition for Edge Devices	Hardware Constraints, FPS Analysis, Quantization,

	and Smart Eyewear: Sensors and Applications	Inference Time, Edge AI, Resource Optimization.
--	--	--

Tabla 1. Estudios del grupo de control

Cadena de búsqueda.

La cadena de búsqueda se construye con los términos clave definidos anteriormente en el grupo de control. Este proceso ayuda a formular las consultas para que sean más precisas en las bases de datos digitales con el objetivo de poder identificar estudios que analicen la intersección entre el hardware con recursos limitados pero con algoritmos avanzados.

La cadena de búsqueda tiene que ser amplia para no dejar fuera áreas innovadores, pero al mismo tiempo tiene que ser muy específica para reducir la aparición de resultados irrelevantes con la investigación. A partir de la identificación y agrupación de palabras clave, se realizó un proceso mediante el uso de operadores lógicos como *AND* y *OR*, con el fin de definir la secuencia de búsqueda más adecuada.

En la tabla 2 se muestra las cadenas de búsqueda y la cantidad de resultados obtenidos, evidenciando el proceso de filtrado hasta llegar a la muestra ideal.

Cadena de Búsqueda	N# Documentos
--------------------	---------------

(Computer Vision) AND (Crowd Analysis)	15,420
(Computer Vision) AND (Crowd Analysis) AND (IoT OR Edge Computing)	946
(Computer Vision) AND (Crowd Analysis) AND (IoT OR Edge Computing) AND (Emotion Recognition)	345
(Computer Vision) AND (Crowd Analysis) AND (IoT OR Edge Computing) AND (Emotion Recognition) AND (Raspberry Pi OR Embedded)	44
(Crowd Density OR People Counting) AND (Emotion Recognition) AND (Edge Computing OR Raspberry Pi) AND (Real-time) AND (Deep Learning)	14

Tabla 2. Elaboración de la cadena de búsqueda óptima.

Nota. Esta tabla ilustra el refinamiento progresivo de la secuencia de búsqueda. La última cadena fue seleccionada por ofrecer un número manejable de documentos altamente pertinentes y alineados con los objetivos del proyecto.

Selección de los Estudios Primarios

Se seleccionó la mejor cadena de búsqueda que resultó en 14 estudios científicos potenciales. Estos documentos fueron sometidos a una lectura completa y a un proceso de verificación contra los criterios de inclusión y exclusión. De este filtrado final, se escogieron 5 Estudios Primarios (EP) que constituyen la base teórica y técnica para responder a las interrogantes de diseño del presente trabajo.

Estudios Primarios Seleccionados

Código	Título
EP1	<i>RT3C: Real-Time Crowd Counting in Multi-Scene Video Streams via Cloud-Edge-Device Collaboration</i>
EP2	<i>Lightweight Convolutional Neural Networks for Real-time Facial Expression Recognition</i>
EP3	<i>Red de detección de multitudes basada en IoT que preserva la privacidad con cifrado homomórfico comparable y su aplicación en la lucha contra COVID19</i>
EP4	<i>IoT video analytics for surveillance-based systems in smart cities</i>
EP5	<i>HSEmotion: High-speed emotion recognition library</i>

Tabla 3. Estudios Primarios

Nota. Esta tabla presenta los artículos científicos definitivos que serán analizados en profundidad para extraer las arquitecturas y metodologías aplicables al proyecto.

En la Tabla 3, se observa la convergencia temática hacia soluciones de eficiencia computacional. Los estudios fueron seleccionados por su relevancia teórica y por ofrecer implementaciones validadas en hardware similar al propuesto en este documento

A continuación, se presenta el análisis de los aportes de cada estudio seleccionado y su relevancia para la presente investigación.

EP1: Recuento de multitudes en tiempo real en streams de vídeo multiescena.

Esta investigación presenta un marco de colaboración en tiempo real orientado al conteo eficiente de multitudes en escenarios de vigilancia que involucren flujos de video. La arquitectura RT3C enlaza los mecanismos para la detección de imágenes, segmentación de los datos y la toma de decisiones basadas en aprendizaje por refuerzo multiagente. Este concepto ayuda a seleccionar entre el procesamiento local con los modelos ligeros y el procesamiento en la nube con modelos de mayor dificultad. Los resultados demuestran mejoras en la precisión y en la latencia en comparación con los métodos convencionales.

EP2: Redes neuronales convolucionales para el reconocimiento de expresiones faciales en tiempo real. Este artículo aborda el problema de la clasificación de emociones desde un enfoque hacia la eficiencia computacional. A diferencia de las arquitecturas tradicionales como *VGG* o *ResNet*, el estudio propone el uso de redes ligeras, como *MobileNet* y *SqueezeNet*. Los resultados muestran que estos modelos logran un equilibrio entre precisión y rendimiento, si bien presentan una reducción del 2 % en precisión respecto a arquitecturas de gran escala, son capaces de ejecutarse hasta diez veces más rápido.

EP3: Red de detección de multitudes basado en IoT que preserva la privacidad con cifrado homomórfico comparable y su aplicación en la lucha contra COVID19. Esta investigación en cambio propone analizar el problema de la privacidad en las redes IoT enfocadas a la detección de las multitudes, aquí la transmisión directa de los datos da lugar a filtraciones de información sensible. Los autores proponen crear cifrado que permiten tener los datos protegidos entre múltiples usuarios, como consecuencia se reduce el costo computacional en comparación con los enfoques actuales.

EP4: Análisis de vídeo IoT para sistemas de vigilancia en ciudades inteligentes. Este trabajo analiza cómo las ciudades inteligentes se benefician del procesamiento local de video en dispositivos IoT, al reducir la dependencia de la nube y mitigar los problemas de latencia. Los autores plantean una arquitectura enfocada a las tareas de vigilancia que permite ejecutar modelos de redes neuronales convolucionales en dispositivos y en gateways IoT. Mediante el caso de estudio centrado en el monitoreo de tráfico, se analiza el rendimiento y la eficiencia de distintos modelos, demostrando que si es posible realizar analítica de video en tiempo real incluso sobre hardware con recursos limitados. Los resultados refuerzan el uso de Edge AI como una alternativa práctica para sistemas de vigilancia en entornos de smart cities.

EP5: HSEmotion - La Librería de reconocimiento de emociones. Este trabajo habla sobre el problema de la desviación algorítmica y la necesidad de crear modelos en sistemas de reconocimiento de emociones faciales. Los autores presentan la librería HSEmotion, un modelo basado en la arquitectura EfficientNet que ayuda a realizar la clasificación emocional de

manera rápida y precisa. Estos resultados exhiben que los modelos preentrenados por lo general pueden reutilizarse como extractores de características para el uso de análisis de video en tiempo real, sin requerir al reentrenamiento completo de la red neuronal como se lo realiza tradicionalmente.

Caracterización de los Estudios Primarios

Al integrar los hallazgos de estos trabajos, es posible dar una respuesta sólida y fundamentada a la interrogante planteada en esta investigación sobre las técnicas y arquitecturas idóneas para el monitoreo de multitudes. La literatura analizada converge en que la solución técnica óptima reside en una arquitectura de procesamiento descentralizada. Los estudios EP1 y EP4 respaldan el Edge Computing como un paradigma clave para disminuir problemas de latencia crítica. Estos trabajos ponen en evidencian que la ejecución de los modelos por visión en dispositivos IoT reducen la dependencia de la conectividad a la red y dan una respuesta operativa al instante, lo cual ayuda a justificar el uso de hardware, en este caso el dispositivo Raspberry Pi con respecto a infraestructuras de servidores en la nube con un alto costo.

La elección del tipo de software de inteligencia artificial se respalda en los hallazgos reportados de los estudios EP2 y EP5, Los dos estudios coinciden en que el equilibrio entre velocidad y precisión se puede lograr gracias al uso de redes neuronales ligeras. Esto cuestiona la necesidad de emplear modelos típicos y también demuestra que es posible realizar tareas de clasificación emocional y conteo en tiempo real utilizando recursos computacionales limitados sin comprometer la efectividad del sistema.

El estudio EP3 se incorpora lo que es la ética dentro de la propuesta técnica, al señalar que el procesamiento local mejora en términos de rendimiento y también constituye una estrategia de privacidad dentro del diseño. Al realizar el análisis de video directamente en el borde y transmitir únicamente los metadatos, se reducen los riesgos asociados a la exposición de información sensible. En general, la revisión sistemática permite concluir que la integración de computación en el borde y modelos de inferencia ligeros representan el estado del arte para abordar la gestión de aglomeraciones de manera segura y escalable.

Respuestas a las preguntas de investigación

RQ1: ¿Es posible implementar un sistema de visión artificial basado en dispositivos IoT que pueda estimar con precisión la densidad de personas y los estados emocionales en multitudes, logrando un balance adecuado entre rendimiento y costo para apoyar la toma de decisiones en seguridad?

Las investigaciones revisadas anteriormente coinciden en la viabilidad técnica y económica de este enfoque siempre y cuando se acepte el cambio de paradigma desde el procesamiento centralizado hacia el paradigma del Edge Computing. Los estudios EP1 y EP4 analizados comprueban que la dependencia de servidores costosos ya no es un obstáculo obligatorio, por el contrario, son arquitecturas descentralizadas donde los dispositivos IoT de bajo consumo, como la Raspberry Pi, asumen toda la carga del procesamiento de video. Esta estrategia reduce los costos de infraestructura y disminuyen los problemas de latencia, permitiendo una respuesta en tiempo real que es fundamental para la seguridad física.

Además, la evidencia deduce que el equilibrio entre rendimiento y precisión se logra mediante una selección adecuada del software. EP2 y EP5 destacan que el uso de redes neuronales ligeras y librerías optimizadas permiten mantener una tasa de cuadros por segundo funcional

para captar la dinámica de una multitud sin sacrificar la confianza de la inferencia de las emociones.

RQ2: ¿Qué arquitecturas de redes neuronales y estrategias de optimización son las más eficientes para ejecutar tareas de visión computacional en entornos de hardware limitado?

Las estrategias de implementación técnica según la literatura seleccionada indica una tendencia hacia el uso de arquitecturas de Redes Neuronales Ligeras. Los estudios coinciden en que arquitecturas monolíticas como VGG o ResNet son inviables para dispositivos IoT debido a su alto costo computacional. En su lugar, se proponen modelos basados en arquitecturas como *MobileNet*, *SqueezeNet* y *EfficientNet*, los cuales reducen el número de parámetros y operaciones matemáticas necesarias para la detección.

Estas arquitecturas permiten aplicar técnicas de optimización lo que facilita que el modelo corra fluidamente en procesadores ARM limitados sin una pérdida crítica de precisión.

Adicionalmente, al reutilizar modelos preentrenados como extractores de características evita la necesidad de reentrenar la red completa desde cero, ahorrando recursos y tiempo. La aplicación de estas técnicas se refleja en sistemas capaces de realizar detección y clasificación simultánea en tiempo real, comprobando el uso de Edge AI como la solución para la vigilancia inteligente moderna.

¿Cuáles son las herramientas tecnológicas principales para realizar la implementación de visión artificial en el borde?

Herramientas	Estudios
Redes neuronales convolucionales: MobileNet, EfficientNet, SqueezeNet.	EP2, EP4, EP5.
Edge Computing: Offloading y Procesamiento Local.	EP1, EP4.
Software para visión y el análisis emocional: HSEmotion, OpenCV.	EP4, EP5.

Capítulo III

Arquitectura del sistema

Este presente trabajo de investigación sigue una metodología empírica, ya que se basa en la observación y análisis de datos reales capturados mediante las cámaras de video, además utiliza los algoritmos de detección de objetos y análisis facial para procesar y clasificar las escenas de multitudes.

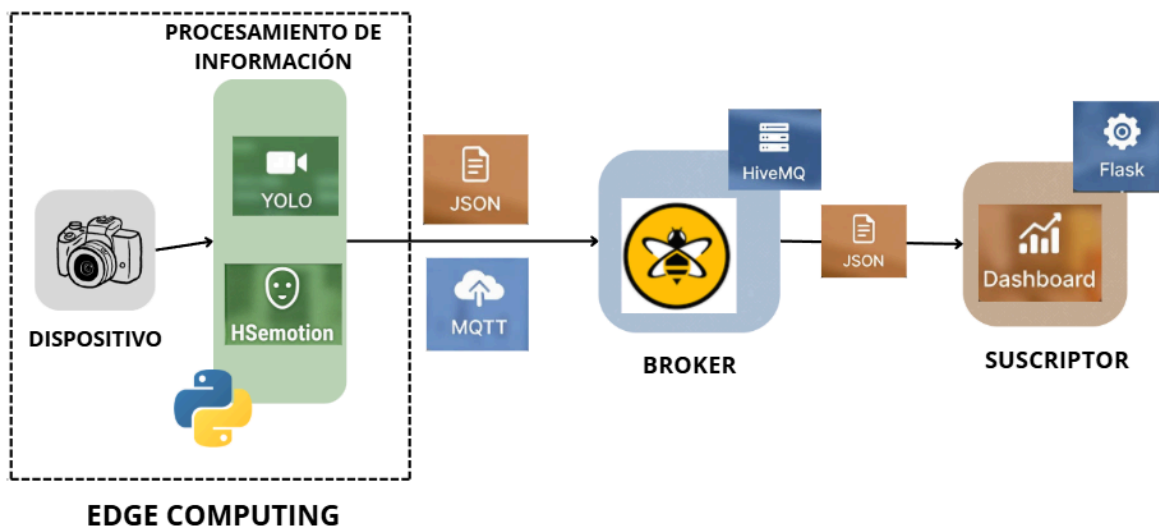


Figura 3. Arquitectura del sistema IoT

La arquitectura propuesta se describe en la Figura 3. el funcionamiento del sistema se estructura como un proceso lineal y continuo que transforma la señal de video en conocimiento. La operación inicia en la capa de percepción física, donde el dispositivo de captura adquiere el flujo de video del entorno en tiempo real. A diferencia de los esquemas de videovigilancia tradicionales, estas imágenes no se transmiten directamente a un servidor externo; por el contrario, son procesados directamente por el módulo de *Edge Computing*. En esta etapa, gestionada localmente por la Raspberry Pi y orquestada bajo el lenguaje Python, se ejecutan concurrentemente los motores de inteligencia artificial, en primera instancia, el modelo YOLO escanea la fotografía para localizar la presencia humana, y acto seguido, la librería HSEmotion realiza la inferencia del estado afectivo sobre los rostros validados.

Posteriormente, una vez que la información visual ha sido procesada y convertida en datos numéricos, el sistema procede a la siguiente fase, la de serialización y transporte. Los resultados se estructuran en formato JSON, lo que permite convertir archivos de imagen

pesados en cadenas de texto ligeras, optimizando el ancho de banda y garantizando la privacidad de los usuarios. Para la transmisión de estos metadatos, el dispositivo de borde asume el rol de "Publicador" y utilizando el protocolo MQTT para enviar la información hacia el *Broker Central*, implementado en esta arquitectura mediante el servicio HiveMQ. Este componente es el intermediario que administra toda la cola de mensajes de forma asíncrona, permitiendo que la información llegue al destinatario incluso con la latencia en la conectividad de la red.

Finalmente, el proceso termina en la capa de aplicación, donde está el módulo Suscriptor, en este extremo, la aplicación web se mantiene en escucha del tópico gestionado en el Broker. Al recibir los paquetes JSON actualizados, el sistema decodifica la información en tiempo real y actualiza los indicadores del *Dashboard*. Es en esta interfaz donde los datos técnicos se traducen en gráficos intuitivos, permitiendo a los operadores monitorear la evolución y los cambios de la densidad poblacional y el clima emocional de la multitud, cerrando la brecha entre la captura técnica en el borde y la toma de decisiones humana.

Desarrollo e implementación del sistema IoT

La parte central del sistema reside en el nodo de borde, constituido por el dispositivo Raspberry Pi 4. Su arquitectura ARM de cuatro núcleos permite la ejecución de scripts de Python con soporte para multithreading,

Esta capa se centra en la creación de un bucle de control principal encargado de la orquestación de tareas. El sistema realiza un muestreo inteligente de cuadros por segundo

para evitar la saturación del procesador, manteniendo el equilibrio entre precisión de monitoreo y estabilidad del hardware. Antes de ingresar a las redes neuronales, cada imagen capturada es normalizada en dimensiones y escala de color, asegurando que los modelos reciban el formato de entrada para el cual fueron entrenados.

La inteligencia del sistema se logra mediante el encadenamiento secuencial de los modelos de aprendizaje profundo. El primer paso del algoritmo es el fotograma completo. Su función es generar las Cajas Delimitadoras alrededor de cada persona. El sistema extrae las coordenadas de cada detección, las cuales son puntos de recorte para la siguiente etapa. Ya con las coordenadas proporcionadas por YOLO, el sistema recorta los rostros de la imagen original. Estos sub-cuadros son enviados individualmente al modelo HSEmotion, el cual retorna la probabilidad de la emoción predominante.

Para resolver el desafío de la transmisión eficiente en redes inestables, se diseñó una estructura de datos ligera basada en notación de objetos (JSON). El nodo de borde construye un paquete de texto que sintetiza el estado del entorno en pocos bytes como se observa en la figura 4.

```
{  
  "conteo_personas": 67,  
  "num_caras": 8,  
  "aglomeracion": true,  
  "emociones": {  
    "Neutral": 50.00,  
    "Fear": 25.00,  
    "Sadness": 25.00  
  },  
  "emocion_dominante": "Neutral",  
  "timestamp": "2026-02-02 08:50:30",  
}
```

Figura 4. Mensaje en formato JSON con los datos del obtenidos

Este mensaje es publicado en el tópico MQTT denominado *"aglomeraciones/raspberry_pi/datos"*. La arquitectura desacoplada permite que el Broker HiveMQ retenga este mensaje hasta que el suscriptor esté listo para recibirlo, garantizando la integridad de los datos.

La fase final del desarrollo se centró en la implementación de la capa de presentación y visualización de los datos, materializada a través de la aplicación web desarrollada con el framework Flask. Este componente actúa como un cliente MQTT persistente, suscribiéndose al tópico de datos para procesar el flujo de información en tiempo real y transformarlo en inteligencia operativa.

Como se observa en la Figura 5, la interfaz fue diseñada con principios de ergonomía visual orientada a gráficos operadores de seguridad, estructurando la información en tres módulos:

1. **Métricas de Aforo en Tiempo Real:** Mediante el uso de tecnologías web asíncronas, los contadores de ocupación se actualizan al instante al recibir cada nuevo paquete MQTT, eliminando la necesidad de recargar la página.
2. **Análisis Emocional:** Se incorpora un panel de control donde se plasma los gráficos de barras y de pastel que visualizan la distribución dominante de las emociones detectadas por el modelo HSEmotion, identificando los cambios en el comportamiento dinámico de la multitud para poder anticipar posibles incidentes dentro del área de la seguridad.
3. **Historial de Tendencias:** El dashboard almacena de forma temporal los datos capturados para la generación de las gráficas que permiten al administrador identificar posibles picos de aglomeración recientes.



Figura 5. Dashboard del sistema

Capítulo IV

Metodología de procesamiento de Datos

Para llevar a cabo el presente proyecto, se aplicará una metodología de procesamiento de datos para el desarrollo de modelos de visión por computador que consta de varias fases principales.

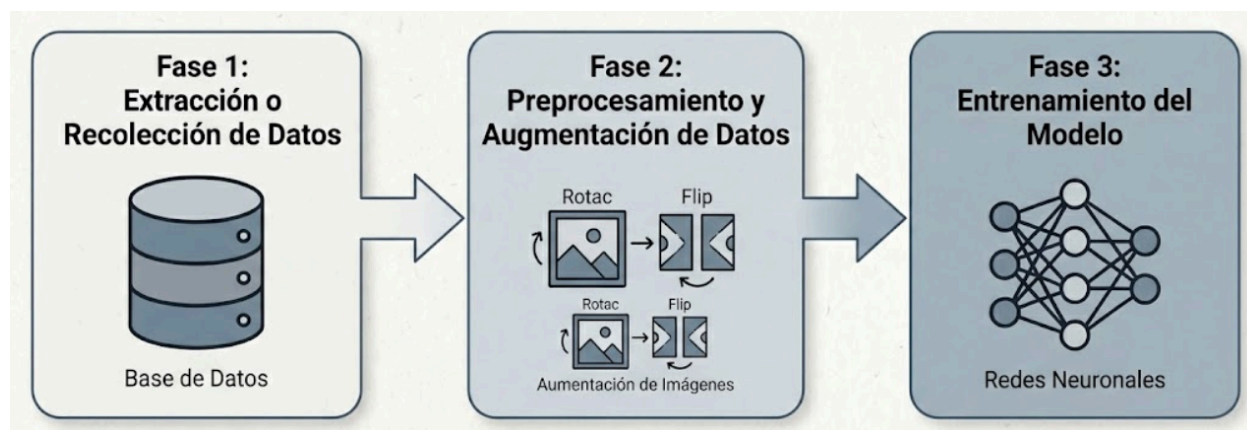


Figura 6. Fases del entrenamiento del modelo

Fase 1: Extracción o recolección de datos

En esta fase se realizó la recolección del conjunto de datos necesario para el entrenamiento y la validación del modelo de detección de aglomeración de personas. Esta etapa es importante porque garantiza la disponibilidad de datos necesarios para que el modelo aprenda a identificar de manera precisa y correcta entre los diferentes niveles de multitudes de las personas en diversas condiciones espaciales.

Para obtener el conjunto de imágenes que conforman la base de datos para este proyecto se utilizó el repositorio Roboflow, una herramienta especializada para la selección, el filtrado y el preprocesamiento de los datos de visión por computadora.

El conjunto de datos que se adaptó más a las necesidades de este proyecto se denomina "Crowd Density", el cual contiene 4.587 imágenes etiquetadas, considerando distintos niveles de afluencia en entornos públicos y privados como calles, centros comerciales y eventos masivos, como se muestra en la Figura 5.



Figura 7. Conjunto de Datos desde la plataforma Roboflow

El proceso de la recolección de datos inició con la creación de un proyecto dentro de la plataforma Roboflow, donde se seleccionó “Object Detection” como el tipo de proyecto requerido, porque se orienta a la detección de objetos y cálculo de densidad.

- **Registro y acceso a Roboflow:**

El primer paso consistió en registrarse en la plataforma Roboflow (<https://roboflow.com>) y acceder a la pagina principal mediante una cuenta institucional.

- **Selección del dataset:**

Una vez dentro de la plataforma, se buscó y seleccionó el conjunto de datos “Crowd Density”, que contiene anotaciones previas realizadas en formato YOLOv8 y COCO JSON, formatos que son ampliamente compatibles con modelos de detección modernos.

- **Copia del conjunto de datos:**

Se copió el dataset en el formato YOLOv8, este fue elegido por su compatibilidad directa con frameworks de aprendizaje profundo como Ultralytics y YOLO, ideales para el proceso de entrenamiento de este proyecto. La copia abarcó tanto las imágenes como los archivos de etiquetas correspondientes (Figura 6).

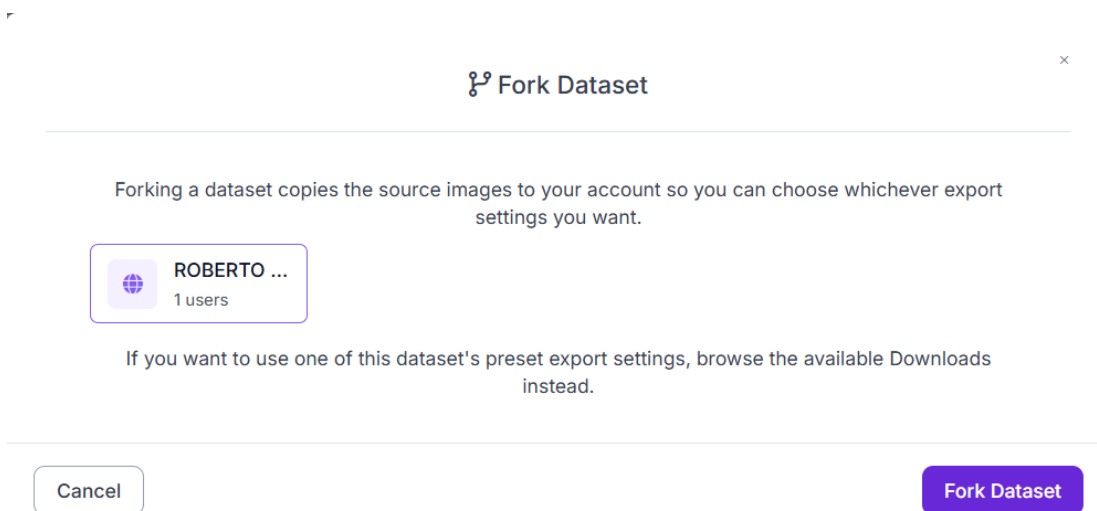


Figura 8. Creación de la copia del Dataset

El conjunto de datos se organiza en tres subconjuntos destinados al entrenamiento, la validación y la prueba, con la finalidad de seguir las buenas prácticas en la construcción de modelos de visión por computadora.

- **Train (70%):** Contiene 3.212 imágenes destinadas al entrenamiento del modelo.
- **Valid (20%):** Contiene 917 imágenes utilizadas para ajustar parámetros y evitar sobreajuste.
- **Test (10%):** Contiene 458 imágenes destinadas a la evaluación final del desempeño del modelo.

Cada una de las imágenes cuentan con el archivo de anotación, donde se detallan las coordenadas de los cuadros delimitadores que indican la ubicación de las personas dentro de la escena, junto con la etiqueta correspondiente.

Es importante destacar que las imágenes del conjunto de datos Crowd Density provienen de fuentes públicas y han sido procesadas para eliminar información personal identificable, cumpliendo con los principios éticos de manejo de datos visuales. Además, el uso de este conjunto de datos se realizó respetando las licencias de uso establecidas por los autores en la plataforma Roboflow.

Fase 2: Preprocesamiento y Augmentación de Datos

El preprocesamiento de los datos constituye una etapa esencial dentro del flujo de un modelo de visión por computadora porque permite mejorar la calidad de las imágenes y aumentar la robustez del modelo frente a posibles variaciones del entorno. En el presente trabajo el preprocesamiento se realizó utilizando las herramientas

integradas en la plataforma Roboflow, La cual proporciona una amplia variedad de funciones de transformación y aumento de imágenes, optimizadas específicamente para tareas de detección de objetos.

El objetivo principal de esta fase consiste en enriquecer el conjunto de datos original y reducir posibles sesgos presentes en las imágenes de multitudes, tales como oclusiones, variaciones lumínicas, cambios de orientación y diferencias de escala. Estas condiciones son comunes en escenarios reales con alta densidad de personas, por lo que su correcta representación en los datos es fundamental para asegurar un desempeño confiable del modelo (Figura 9).



Figura 9. Fase de preprocesamiento y augmentación de datos en Roboflow

Durante esta etapa se desarrollaron técnicas de normalización y ajuste de formato para normalizar las imágenes y facilitar la lectura al modelo de detección. Los cambios aplicados fueron las siguientes:

- **Orientación automática:**

Se activó la opción de orientación automática para asegurar que todas las imágenes mantuvieran una alineación congruente, corrigiendo rotaciones involuntarias o datos inconsistentes provenientes de distintos dispositivos de captura.

- **Redimensión de imágenes:**

Las imágenes fueron redimensionadas a un estándar de 640×640 píxeles, asegurando compatibilidad con los requisitos del modelo YOLO, el cual consigue imágenes cuadradas para un procesamiento uniforme.

- **Augmentación de datos:**

Para incrementar la variabilidad del conjunto de datos y la mejora de la capacidad de generalización del modelo, se aplicaron métodos de augmentación controlada.

Estas transformaciones generan versiones adicionales de las imágenes con modificaciones aleatorias dentro de rangos definidos.

Las transformaciones aplicadas fueron las siguientes:

- **Flip horizontal:**

Transposición de las imágenes en el eje horizontal, concediendo al modelo reconocer configuraciones simétricas de las escenas.

- **Rotación:**

Se aplicaron rotaciones de las imágenes en un rango de ± 10 grados, simulando diferentes perspectivas en tomas de cámaras en ángulos distintos.

- **Brillo y contraste:**

Modificación de los niveles de luz y contraste en un rango de ± 10 %, para reproducir condiciones de iluminación diversas como interiores, exteriores y horas del día.

- **Ruido:**

Suma del ruido en el 0.1 % de los píxeles, para simular interferencias comunes en imágenes tomadas en entornos con movimiento o baja calidad de cámara.

- **Desenfoque:**

Aplicación de un desenfoque de hasta 2 px, para representar imágenes fuera de foco, como ocurre en videos o fotografías en estado de movimiento.

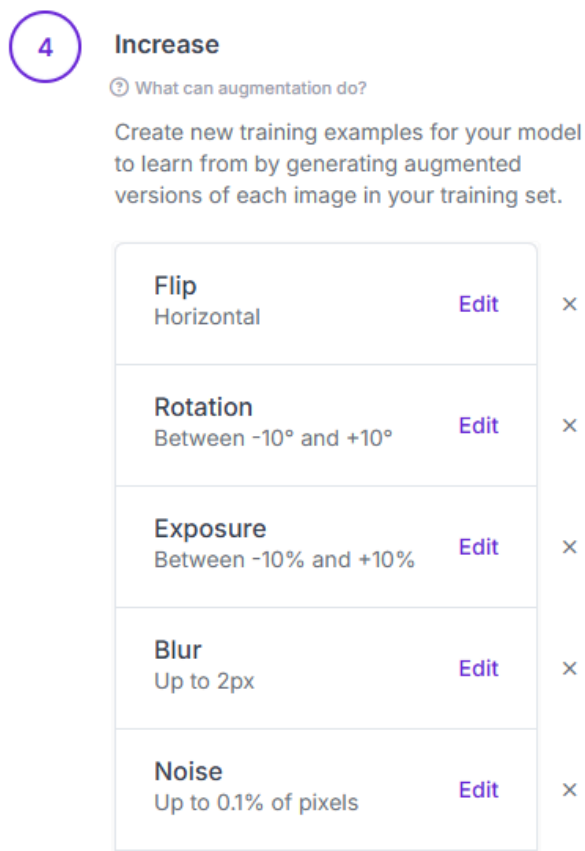


Figura 10. Augmentaciones aplicadas en el datasets Crowd Density

Como se observa en la figura 10, del proceso de augmentación, el conjunto de datos incrementó de 4.587 a 11.011 imágenes, por lo que representa un aumento significativo en la diversidad de escenarios.

Esta expansión del dataset ayudo a mejorar la capacidad del modelo para detectar y estimar la densidad de multitudes bajo condiciones reales, disminuyendo el riesgo del sobreajuste y favorecer a una mejor generalización en la fase de inferencia.

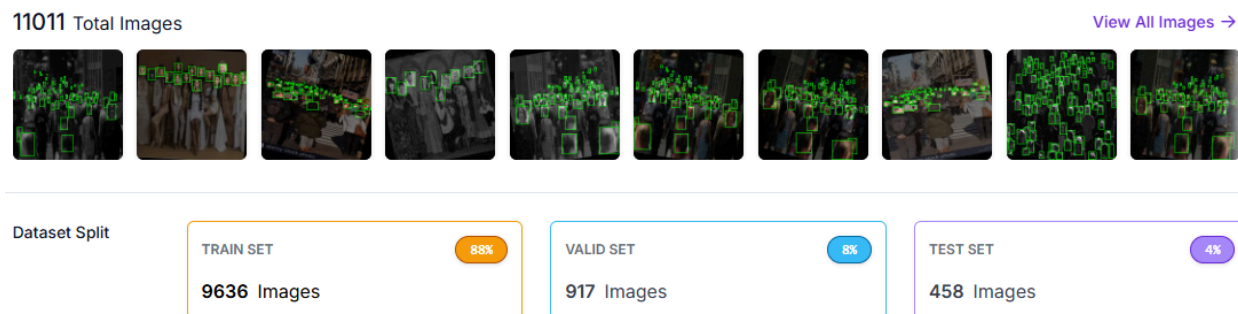


Figura 11. Tamaño del dataset después de la augmentación

Fase 3: Entrenamiento del modelo

El entrenamiento del modelo es la fase más importante para que la inteligencia artificial aprenda a reconocer y estimar la cantidad de personas en las imágenes de multitudes. Esta etapa conlleva exponer al modelo a un conjunto de datos amplio, ajustar sus parámetros mediante algoritmos de optimización, y también evaluar su desempeño de manera iterativa hasta alcanzar un nivel alto de precisión.

El entrenamiento del modelo de detección de objetos requiere recursos computacionales elevados, como el uso de unidades de procesamiento gráfico (GPU), es por eso que se seleccionó usar servicios de computación en la nube bajo el modelo de Plataforma como Servicio (PaaS), porque permiten el acceso a infraestructura con aceleración por hardware sin necesidad de gestión de los recursos físicos. Esta decisión se escogió porque las GPUs del equipo local presentan limitaciones de memoria y capacidad de procesamiento, lo que restringe el tamaño de batch y el número de las épocas en el proceso de entrenamiento del modelo.

Por esa razón Colab es la primera opción de elección, ofreciendo GPUs de alto rendimiento y entornos configurados para Python, mejorando la ejecución de los modelos basados en la arquitectura YOLOv8.

Para iniciar el entrenamiento, se instalaron las librerías necesarias y se verificó la compatibilidad del entorno.

```
%pip install "ultralytics<=8.3.40" supervision roboflow

!yolo settings sync=False
import ultralytics
ultralytics.checks()

Ultralytics 8.3.40 🚀 Python-3.12.12 torch-2.8.0+cu126 CUDA:0 (Tesla T4, 15095MiB)
Setup complete ✅ (2 CPUs, 12.7 GB RAM, 39.7/235.7 GB disk)
```

Figura 12. Instalación de Ultralytics YOLO

En la figura 12 se visualiza la instalación de la versión de Ultralytics YOLO utilizada en el proyecto, previene el seguimiento de la actividad de la herramienta y comprobar que todas las dependencias estén correctamente configuradas.

```
from roboflow import Roboflow
rf = Roboflow(api_key="Fj91ZEVdQ9DamP3T91km")
project = rf.workspace("roberto-alexander").project("crowd-density-ou3ne-leuho")
version = project.version(1)
dataset = version.download("yolov11")
```

Figura 13. Configuración de la API

El conjunto de datos fue descargado desde la plataforma Roboflow mediante el uso de una interfaz de programación de aplicaciones (API), la cual permitió acceder y

obtener los datos necesarios para el entrenamiento del modelo, tal como se ilustra en la Figura 13.

A continuación, se montó Google Drive como sistema de almacenamiento para guardar los resultados del entrenamiento.

Para organizar y compartir los resultados del entrenamiento del modelo, se creó una carpeta específica en Google Drive bajo el nombre YOLO_Crowd_Detection_Results.

Posteriormente, se definió la ruta del archivo YAML generado por Roboflow, que contiene la información de clases, rutas de imágenes y subconjuntos de entrenamiento (Figura 14).

```
downloaded_folder_name = "crowd-density-1"  
DATASET_PATH = os.path.join('/content', downloaded_folder_name)  
YAML_FILE = os.path.join(DATASET_PATH, 'data.yaml')
```

Figura 14. Configuración del archivo “data.yaml”

El entrenamiento se llevó a cabo sobre la arquitectura YOLOv8m, seleccionada por su equilibrio entre velocidad y precisión en tareas de detección de objetos en imágenes con alta densidad de elementos.

Los parámetros principales utilizados fueron:

Parámetro	Descripción	Valor
Modelo Base	yolov8m.pt	Pesos preentrenados de YOLOv8, optimizados para detección de objetos generales.
Épocas	200	Número de iteraciones sobre el conjunto de entrenamiento
Tamaño de imagen	640 x 640	Resolución estándar de entrada para el modelo
Tamaño del batch	32	Número de imágenes procesadas en cada paso de entrenamiento
Paciencia	50	Número de épocas sin mejora para detener el entrenamiento automáticamente

Tabla 4. Parámetros de entrenamiento

El proceso de entrenamiento adoptó un flujo iterativo que recorre de forma ordenada las imágenes del conjunto de datos, agrupándolas en lotes de 32 elementos. En cada ciclo, las imágenes proveen a la red convolucional, que extrae patrones a múltiples escalas, mientras el neck combina estas representaciones mediante estructuras FPN y PAN, capturando detalles finos. Desde ese punto, el head es el que

produce las estimaciones del bounding boxes, detallando posiciones, las dimensiones y probabilidades de presencia de personas en cada región.

Estas predicciones se miden contra las anotaciones reales del set de entrenamiento, derivando una pérdida compuesta de tres elementos esenciales:

- La pérdida asociada a la caja evalúa las diferencias en las coordenadas de los objetos detectados mediante el índice IoU, con el fin de garantizar alineaciones precisas y estables.
- La pérdida de clasificación se centra en corregir errores en el reconocimiento de la clase “persona”.
- La pérdida asociada al objeto está diseñada para penalizar detecciones erróneas en regiones donde no existen entidades.

Los parámetros de la red se ajustan por retropropagación, empleando un optimizador de Ultralytics con una tasa de aprendizaje que se adapta dinámicamente en el proceso. Los gradientes van en sentido inverso, ajustan pesos capa por capa en un proceso que reveló convergencias tras las primeras docenas de iteraciones. Este ciclo se extiende por 200 épocas, cubriendo el dataset completo en cada pasada y demandando recursos moderados en el hardware.

Al mismo tiempo, cada época incorpora una evaluación en el subconjunto de validación, permitiendo un seguimiento continuo de indicadores como el mean Average Precision mAP@0.5, la precisión y el recall que son medidas que destacan la robustez del modelo ante variaciones reales, como sombras o multitudes densas en detección peatonal.

Para aplacar los riesgos de sobreajuste en el modelo, se aplicó un mecanismo de paciencia de 50 épocas es decir si existe ausencia de avances en la métrica de validación durante ese intervalo, el entrenamiento automáticamente se interrumpe de inmediato, conservando el estado y evitando ineficiencias innecesarias.

Los resultados del entrenamiento, incluyendo pesos finales, métricas de validación, gráficas de pérdida y precisión, se almacenaron en la carpeta de Google Drive. Esto ayuda y facilita la posterior evaluación y ajuste de los parámetros.

Capítulo V

Evaluación del modelo

En el ciclo de vida de cualquier solución de Inteligencia Artificial, es imprescindible establecer un marco de métricas objetivas que permitan cuantificar su fiabilidad. Determinar la efectividad del sistema propuesto para la tarea de monitoreo de multitudes se fundamenta en los dos pilares metodológicos descritos por Pykes (2020):

- **Métricas de evaluación:** Es necesario evaluar el equilibrio entre la exactitud de la detección (mAP, Precisión, Recall) y la eficiencia operativa (FPS, Latencia), asegurando que el modelo actúe de forma adecuada en tiempo real.
- **Ejecución del procedimiento de evaluación correcto:** Verificar si el modelo mantiene su robustez al ser expuesto a datos inéditos y condiciones no controladas como los cambios de luz y las oclusiones, garantizando que los resultados reflejen el comportamiento del sistema en un entorno de producción real y no solo en condiciones de laboratorio.

Métricas de evaluación

Precisión: Esta métrica es fundamental para medir la calidad de las detecciones dentro de la aglomeración. Representa la proporción de objetos identificados como personas que realmente lo son. Se calcula como el número de detecciones correctas dividido entre el total de detecciones realizadas.

$$P = \frac{TP}{TP+FP}$$

Recall: Esta medida es la capacidad del modelo para encontrar a todas las personas presentes en una imagen. Es un indicador de la seguridad, ya que mide qué tan hábil puede llegar a ser el algoritmo para no omitir individuos, incluso si están ocultos o en los bordes de la imagen.

$$R = \frac{TP}{TP+FN}$$

F1: Es una métrica que permite evaluar la robustez del modelo, asegurando que el sistema mantenga un equilibrio entre la exactitud del conteo de los individuos y la capacidad de detección, sin sobreestimar el número real de las personas.

$$F1 = 2 \frac{P \times R}{P+R}$$

Matriz de Confusión: Proporciona una representación cuantitativa detallada de los aciertos y errores del detector. En este caso, permite visualizar cuántas personas fueron correctamente identificadas (Verdaderos Positivos), cuántas fueron omitidas (Falsos Negativos) y cuántos objetos irrelevantes fueron contados erróneamente como personas (Falsos Positivos).

Análisis de Resultados del Entrenamiento del Modelo

Precisión

Al observar la evolución del entrenamiento presentada en la figura 15, se identifica una trayectoria de aprendizaje sólida y consistente a lo largo de las 200 épocas programadas. La curva de precisión no presenta variaciones erróneas, al contrario, sigue una tendencia ascendente que sugiere una convergencia positiva del algoritmo. El comportamiento del modelo se divide en tres fases importantes:

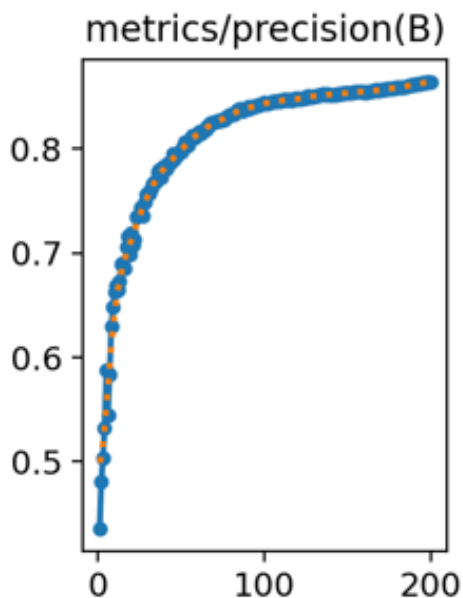


Figura 15. Precisión del modelo entrenado

1. **Fase de Aprendizaje Acelerado:** En el inicio del entrenamiento, la gráfica muestra un crecimiento vertical acentuado. El modelo inicia con una métrica de precisión de 0.43 y escala paulatinamente hasta superar el umbral de 0.71.
2. **Fase de Refinamiento:** En esta etapa, la pendiente se neutraliza, aunque mantiene un crecimiento constante, en un rango de 0.80 a 0.85, el modelo deja de entender los patrones claros y empieza a tener mejoría en su capacidad de detección, ajustando los pesos para abordar una posición más complejas, como la presencia de las oclusiones o las variaciones en las condiciones de iluminación de las imágenes.
3. **Fase de Estabilización y Convergencia:** Al final del entrenamiento, la métrica alcanza un valor de estabilidad del 0.864. Este resultado final de 86.4% es satisfactorio para un modelo desplegado en el borde. La ausencia de caídas bruscas en las últimas épocas confirma que el modelo logró un equilibrio óptimo entre la generalización y el ajuste, evitando caer en el sobreajuste.

Recall:

Como se observa en la figura 15, la curva de Recall-Confidence ilustra la relación inversa que existe entre la capacidad del modelo para detectar la totalidad de los objetos presentes y la seguridad con la que emite dichas predicciones.

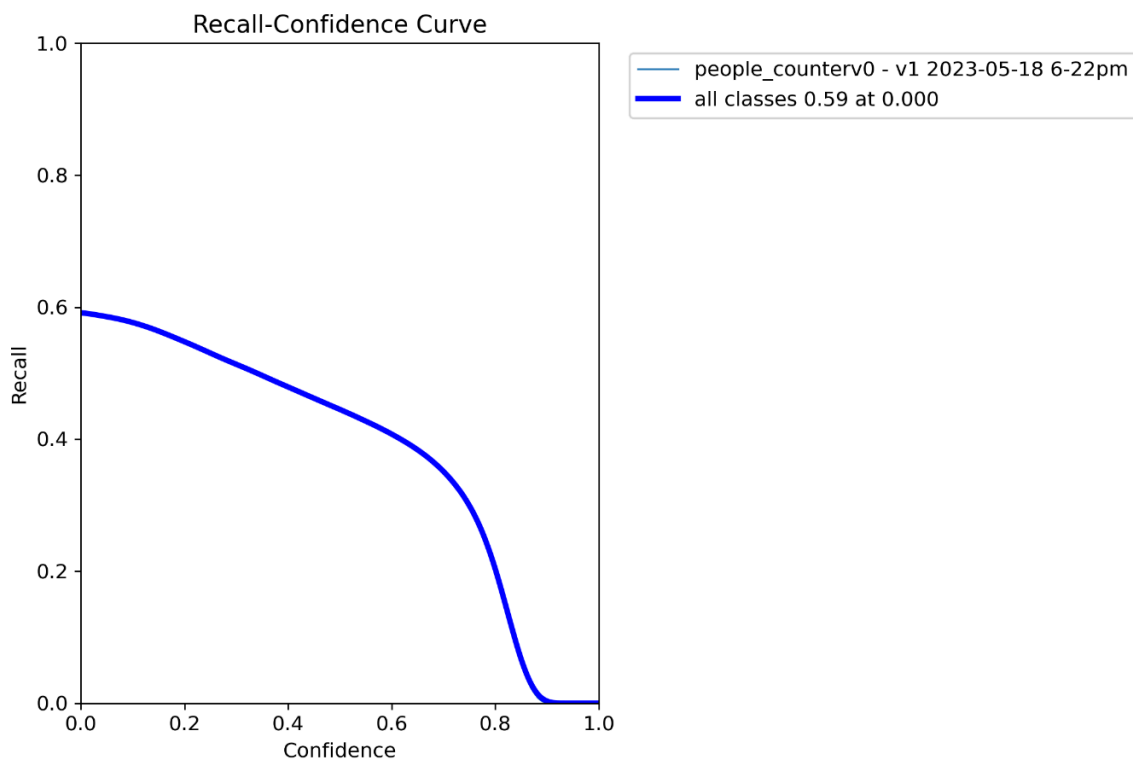


Figura 15. Recall del modelo entrenado

El gráfico revela un comportamiento característico de los sistemas de detección orientados a la precisión. En el eje vertical, el modelo inicia con un valor de 0.59, aproximadamente 60% de recuperación cuando el umbral de confianza es mínimo. A medida que avanzamos hacia la derecha en el eje horizontal exigiéndole al sistema una mayor certeza para validar una detección, la curva se mantiene relativamente estable con una pendiente suave hasta llegar al umbral de confianza de 0.6.

Este comportamiento indica una estabilidad operativa: el modelo es capaz de retener la mayoría de sus detecciones correctas incluso cuando se le exige una confianza media-alta. El descenso pronunciado que se aprecia a partir del umbral de 0.7 es esperable y denota la

rigurosidad del algoritmo, el sistema sacrifica la detección de casos ambiguos para priorizar únicamente aquellas detecciones de las que está plenamente seguro.

En el contexto de monitoreo de multitudes, este perfil de curva sugiere que el sistema se comporta de manera conservadora y robusta, en lugar de intentar marcar cualquier sombra como una persona el modelo prefiere asegurar la calidad del dato, garantizando que los individuos contabilizados tengan una alta probabilidad de ser reales.

Puntuación F1:

En la Figura 16 se muestra la curva de *F1-Confidence* que sintetiza el compromiso entre la Precisión y el Recall, permitiendo identificar el punto de operación óptimo del sistema. A diferencia de las métricas individuales, esta gráfica nos revela cuál es el umbral de confianza exacto donde el modelo logra su mejor desempeño global, equilibrando la necesidad de detectar a la mayor cantidad de personas sin incurrir en un exceso de falsas alarmas.

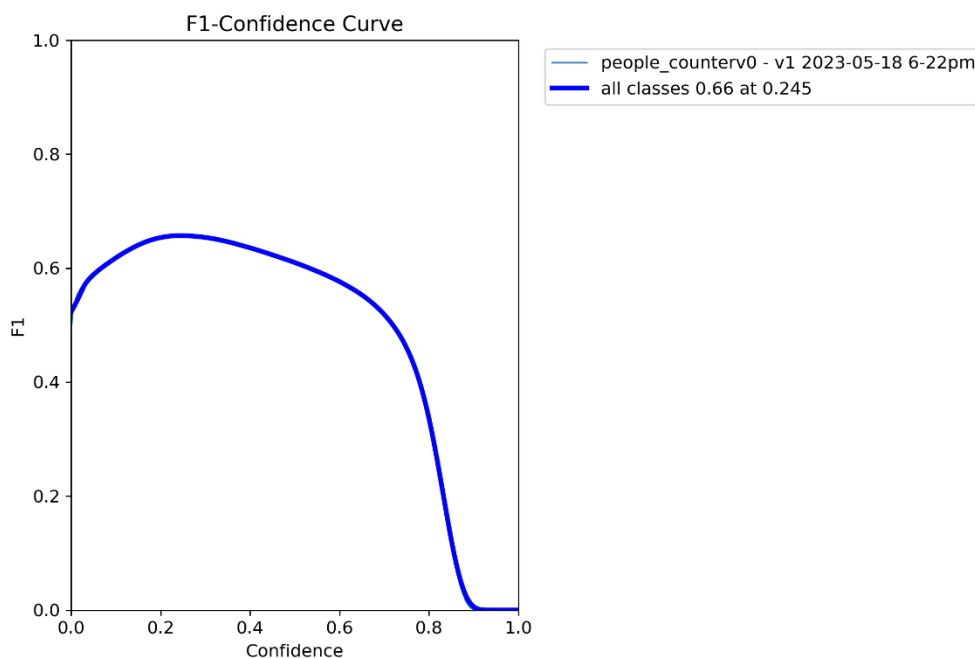


Figura 16. Medida F1 del modelo entrenado

El análisis de la curva identifica un máximo global de 0.66 cuando se establece un umbral de confianza de 0.245. Este dato es importante porque indica que el modelo no requiere una certeza extrema para ser eficaz. Por el contrario, su zona de mayor rendimiento se encuentra en un umbral de confianza bajo a medio. Esto sugiere que, para maximizar la eficacia en el conteo de multitudes, el sistema debe configurarse para ser permisivo, aceptando detecciones con una confianza moderada.

Matriz de Confusión:

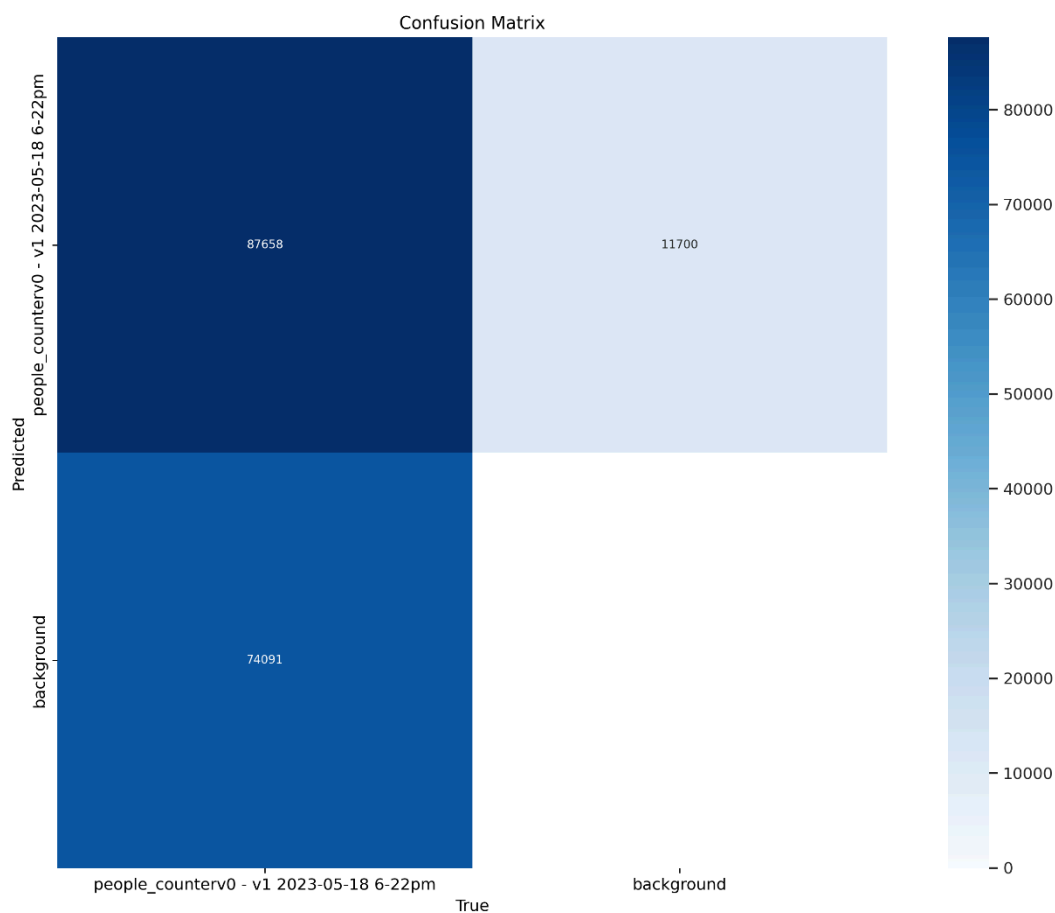


Figura 17. Matriz de confusión

En la Figura 17, la matriz de confusión ofrece información cuantitativa sobre la fiabilidad del clasificador, desglosando el comportamiento del modelo frente a las dos clases críticas: Persona y Fondo.

Al examinar la diagonal principal, el cuadrante superior izquierdo revela una solidez notable, contabilizando un total de 87,658 Verdaderos Positivos. Este valor confirma que, en la

gran mayoría de los casos visibles, el algoritmo logró identificar y segmentar correctamente la figura humana, validando la eficacia del entrenamiento base.

Sin embargo, el análisis de los errores arroja luz sobre las limitaciones físicas del escenario de prueba. Se observa una cifra reducida de 11,700 Falsos, pues indica que el modelo tiene una "alucinación" baja, raramente confunde objetos inanimados o sombras con personas.

La métrica más reveladora son los 74,091 Falsos Negativos. Esta cifra representa a las personas que estaban presentes pero que el modelo no logró detectar, este dato cuantifica la característica de oclusión típico en las aglomeraciones, cuando una persona tapa a otra, o cuando están muy lejos de la cámara, el modelo pierde la referencia.

Evaluación del Rendimiento del sistema

Para la ejecución de este proceso, se implementó el ecosistema de hardware con el software. Esta fase de integración abarcó la puesta en marcha de la Capa de Percepción con la Raspberry Pi operando la cámara y los modelos de inferencia, la Capa de Red encargada de establecer el canal de comunicación MQTT hacia el broker HiveMQ, y finalmente la Capa de Aplicación, que despliega el Dashboard Web con suscripción vía WebSockets.

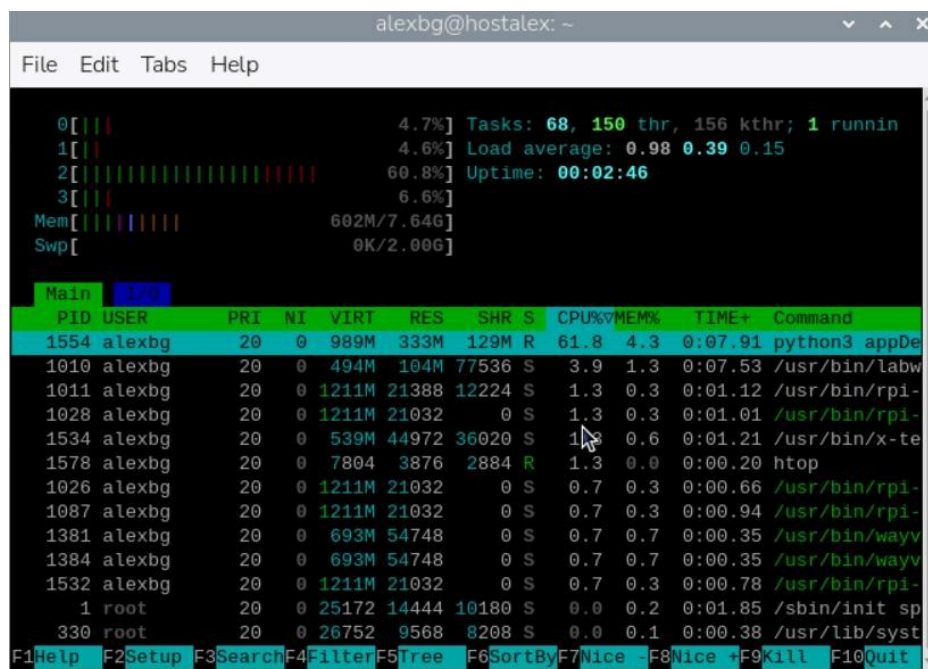


Figura 16. Rendimiento del dispositivo Raspberry

Profundizando en el desempeño computacional, se realizó un monitoreo activo del nodo de borde durante una sesión de inferencia continua. Tal como se evidencia en la figura 16, el sistema demuestra un comportamiento eficiente y estable bajo carga de trabajo, se observa una distribución asimétrica de la carga, donde el proceso principal demanda aproximadamente un 61.8% de capacidad de procesamiento. Esta actividad se concentra mayoritariamente en uno de los núcleos, mientras que los núcleos restantes permanecen en reposo relativo. Este comportamiento es natural en las aplicaciones de visión artificial, Esto confirma que aunque el algoritmo es exigente, no llega a saturar la totalidad de la CPU, de la misma forma la memoria con 7.64 GB disponibles en la placa, el sistema ocupa 602 MB. Esta amplia holgura operativa garantiza que el dispositivo no recurra a la memoria de intercambio, evitando la degradación del rendimiento a largo plazo y permitiendo la implementación futura de algoritmos más complejos sin requerir nuevo hardware.

Evaluación de los resultados del sistema IoT

Para respaldar la imparcialidad de los resultados presentados, es necesario describir las condiciones técnicas bajo las cuales se realizaron las pruebas. El análisis se fundamenta en el procesamiento de un conjunto de datos diseñado para agobiar las capacidades de detección del sistema.

El sistema obtuvo un total de 100 imágenes receptadas en diferentes puntos de la ciudad de Quito en un entorno de vigilancia y de aglomeraciones. La distribución se trató de equilibrar para evitar sesgos en la medición de las métricas de clasificación:

Escenarios Normales: Imágenes que representan flujos peatonales bajos o moderados (< 30 personas), típicos de situaciones cotidianas donde no existe riesgo de seguridad.

Escenarios de Aglomeración: Imágenes de alta densidad (> 30 personas), incluyendo eventos masivos o estaciones de transporte saturadas.

Estos dos escenarios presentan desafíos técnicos adicionales como oclusión entre individuos, variaciones de escalas entre personas muy cercanas vs. muy lejanas y condiciones de iluminación diversas.

El análisis del rendimiento computacional se obtuvo bajo una configuración de hardware fija, correspondiente al dispositivo de borde propuesto en el diseño. Las especificaciones operativas durante la fase de prueba fueron:

- **Dispositivo:** Raspberry Pi 4 Model B.
- **Memoria RAM:** 4GB.
- **Sistema Operativo:** Raspberry Pi OS (64-bit).

- **Resolución de Captura:** 640x480 píxeles.
- **Frecuencia de:** Procesamiento de 1 frame cada 3 capturas

Se diseñó un protocolo de validación de los datos basado en la comparación entre la inferencia del sistema y el Ground Truth.

A partir de la operación del sistema IoT en diversos escenarios de la ciudad de Quito como, Iglesias, Parques, Universidades, Centros Comerciales y Conjuntos Residenciales, se construyó el conjunto de datos de 100 imágenes. Luego se agrupó un muestreo para seleccionar 40 imágenes representativas que constituyeron el subconjunto de validación detallada.

La selección de estas 40 muestras partió de los siguientes criterios:

1. **Diversidad de Densidad:** Introducción de escenarios de baja (<10), media (10-40) y alta ocupación (>50 personas).
2. **Variabilidad de Iluminación:** Muestras tomadas en diferentes horarios
3. **Complejidad:** Escenarios con oclusiones parciales y superposición de individuos.

En las 40 imágenes se realizó un conteo manual de personas y rostros visibles. Luego se comparó estos valores con los resultados generados por el modelo entrenado, calculando las siguientes métricas por cada imagen:

- **Error Absoluto E_i :** Es la diferencia entre el conteo del sistema (C_{sis}) y el conteo (C_{real}) del número de personas.

$$E_i = |C_{real} - C_{sis}|$$

- **Error Porcentual EP_i :** Es la magnitud del error respecto al total de personas.

$$EP_i = \left(\frac{E_i}{C_{real}} \right) \times 100$$

- **Precisión de Conteo P_i :** Es el nivel de exactitud del modelo por imagen.

$$P_i = 100\% - EP_i$$

- **Promedio de Precisión Global $P_{promedio}$:** Esta fórmula indica la confianza del modelo para el conteo de aforo. Se calcula restando el promedio de los errores porcentuales de todas las muestras al 100%

Donde:

- n : Número total de imágenes en la muestra (40).

- EP_i : Error porcentaje de la imagen i

$$P_{Promedio} = 100\% - \left(\frac{1}{n} \sum_{i=1}^n EP_i \right)$$

- **Tasa de Extracción Facial T_{caras}** : Esta métrica evalúa la eficiencia para capturar rostros en entornos de aglomeración. Se obtiene comparando la sumatoria total de rostros detectados por el sistema frente a la sumatoria de rostros visibles reales en todo el conjunto de datos.

$$T_{caras} = \left(\frac{\sum_{i=1}^n CS_i}{\sum_{i=1}^n CR_i} \right) \times 100$$

Donde:

- CS_i : Es la cantidad de caras detectadas por el sistema en una imagen i .
- CR_i : Es la cantidad de caras reales en una imagen i .

Al realizar el conteo manual del número de personas de cada imagen y compararlo con el conteo del sistema se obtuvieron los siguientes resultados:

N Imagen	Personas Reales	Personas Sistema
1	27	25
2	28	27

3	26	26
4	29	28
5	20	18
6	21	20
7	20	20
8	16	14
9	36	33
10	80	76
11	44	40
12	53	47
13	12	12
14	56	56
15	69	71
16	26	24
17	18	14
18	28	28
19	35	33
20	36	34
21	37	37
22	42	42
23	37	36
24	30	29
25	40	40
26	18	14
27	19	18
28	22	22
29	22	21
30	34	32
31	41	41
32	31	31
33	34	32
34	28	28
35	35	35
36	21	19
37	37	37
38	40	39
39	28	28
40	31	29

Tabla 5. Detalle de Precisión por Escenario



Figura 17 Gráfico comparativo del número de personas de la muestra

Al realizar la sumatoria de las detecciones en la muestra seleccionada, se obtuvieron los siguientes valores agregados:

VARIABLE	VALOR TOTAL
Total Personas Reales	1307
Total Personas Detectadas	1256
Diferencia Neta	51

Tabla 6. Sumatoria de personas reales vs personas detectadas

Al realizar la sumatoria de las detecciones en la muestra seleccionada, se observa una alta correlación entre los datos obtenidos manualmente y los inferidos por el modelo.

Como se evidencia en la tabla anterior, de un universo total de 1,307 individuos presentes en las imágenes de prueba, el sistema logró contabilizar exitosamente a 1,256 personas. La diferencia neta de 51 personas lo cual indica una ligera tendencia del modelo hacia la subestimación.

En términos porcentuales, esta diferencia representa un error global de apenas el 3.9%. Este margen puede ser aceptable para sistemas de vigilancia en tiempo real, donde la prioridad es la identificación de tendencias de aglomeración más que el conteo preciso de individuos. Este análisis cualitativo sugiere que esta subestimación no se debe a fallos en la detección de personas aisladas, sino principalmente al factor de oclusión severa en las imágenes de mayor densidad, las personas situadas en segundo plano quedan ocultas casi en su totalidad por los sujetos en primer plano, haciéndolos invisibles para el algoritmo.

N_ Imagen	Error Absoluto	Error Porcentual	Precisión Promedio
1	2	7,41	92,59
2	1	3,57	96,43
3	0	0,00	100,00
4	1	3,45	96,55
5	2	10,00	90,00
6	1	4,76	95,24
7	0	0,00	100,00
8	2	12,50	87,50
9	3	8,33	91,67
10	4	5,00	95,00
11	4	9,09	90,91
12	6	11,32	88,68
13	0	0,00	100,00
14	0	0,00	100,00

15	2	2,90	97,10
16	2	7,69	92,31
17	4	22,22	77,78
18	0	0,00	100,00
19	2	5,71	94,29
20	2	5,56	94,44
21	0	0,00	100,00
22	0	0,00	100,00
23	1	2,70	97,30
24	1	3,33	96,67
25	0	0,00	100,00
26	4	22,22	77,78
27	1	5,26	94,74
28	0	0,00	100,00
29	1	4,55	95,45
30	2	5,88	94,12
31	0	0,00	100,00
32	0	0,00	100,00
33	2	5,88	94,12
34	0	0,00	100,00
35	0	0,00	100,00
36	2	9,52	90,48
37	0	0,00	100,00
38	1	2,50	97,50
39	0	0,00	100,00
40	2	6,45	93,55
Promedio Global			95,30

Tabla 7. Datos de errores y precisión del conteo de personas

Una vez calculado las métricas individuales de las 40 imágenes de validación presentadas en la Tabla 3, se consiguió el indicador de rendimiento del sistema.

El promedio de precisión obtenido para el modelo de detección de personas fue del 95.30%. Este valor indica que el sistema sostiene un margen de error inferior al 5% respecto al

conteo manual. Desde una perspectiva técnica y tomando en cuenta que se entrenó un modelo desde cero, el resultado pone al sistema en un rango de alto rendimiento considerando las limitaciones de hardware y la dificultad visual de los escenarios de prueba.

Al realizar el conteo manual del número de caras detectadas de cada imagen y compararlo con el conteo del número de caras detectadas del sistema se obtuvieron los siguientes resultados:

N_ Imagen	Caras Reales	Caras Sistema
1	7	4
2	8	8
3	7	8
4	6	5
5	11	9
6	7	11
7	8	6
8	8	8
9	7	6
10	11	7
11	10	2
12	3	3
13	4	4
14	8	8
15	9	6
16	9	9
17	10	10
18	9	3
19	15	9
20	11	8
21	5	2
22	13	10
23	8	6
24	9	6

25	6	5
26	8	7
27	11	8
28	9	7
29	5	5
30	9	9
31	5	4
32	8	5
33	9	9
34	12	12
35	9	9
36	4	4
37	8	8
38	6	10
39	7	8
40	12	13

Tabla 8. Detalle de caras detectadas por escenario

Al realizar la sumatoria de las detecciones de caras en la muestra seleccionada, se obtuvieron los siguientes valores:

VARIABLE	VALOR TOTAL
Total Caras Reales	331
Total Caras Detectadas	281
Tasa de Extracción Facial	84,89%

Tabla 9. Sumatoria de caras reales vs caras detectadas

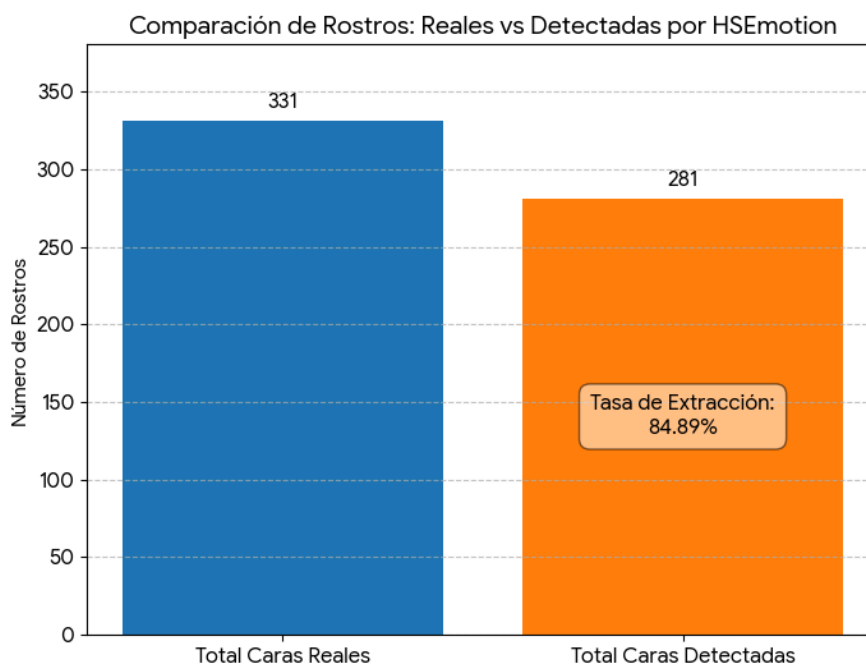


Figura 18. Comparación de rostros Reales vs Detectados por HSEmotion

Al contrastar el número de rostros visibles contra los procesados por el modelo HSEmotion, se observa un desempeño fuerte en la etapa de extracción de características.

De un total de 331 rostros identificados manualmente como frontales hacia la cámara, el sistema logró procesar 281. La diferencia de 50 rostros representa una pérdida de información del 15,11%, deduciendo una tasa de extracción del 84,89%.

Desde una perspectiva técnica, este margen no se asigna necesariamente a una falla de detección, sino a la sensibilidad de los umbrales de confianza del modelo, en otras palabras, para que HSEmotion procese una emoción, requiere una resolución que superan la simple detección de la presencia de la persona. Los 50 rostros no captados corresponden a personas

con ángulos de inclinación o desenfoque por el movimiento, los cuales fueron descartados para no introducir ruido en la inferencia emocional.

Para interpretar correctamente esta diferencia en la detección, es necesario considerar dos condiciones técnicas de los escenarios de prueba:

En primer lugar, el comportamiento y oclusión: en una aglomeración, la dinámica implica movimiento y distracción. Es físicamente imposible esperar que la totalidad de los individuos mantenga contacto visual simultáneo y directo con la cámara de vigilancia, lo que impide que todos los rostros sean candidatos viables para el algoritmo en un momento dado.

En segundo lugar, en la resolución se aplicó un filtro que suprime los rostros con dimensiones inferiores a los 48 píxeles. Esta es una decisión enfocada a la confianza de los datos, al procesar rostros demasiado lejanos o pixelados generan ruido y falsos positivos en la clasificación de la emoción. Las caras que si fueron procesadas son una muestra válida bajo un esquema de muestreo aleatorio.

Con el fin de entender las limitaciones y las fortalezas del sistema, se seleccionaron dos casos específicos de la muestra para realizar un análisis mas específico.



Figura 19. Escenario con alta densidad

En el escenario de la figura 19, se observa el fenómeno de la oclusión. Las personas ubicadas en la parte posterior son tapadas parcialmente por las de adelante. Lo que el sistema prioriza es a los sujetos más cercanos. A pesar del error numérico mínimo en el conteo de personas, el sistema señala como un caso de aglomeración, lo que valida que el sistema cumple su función de seguridad.

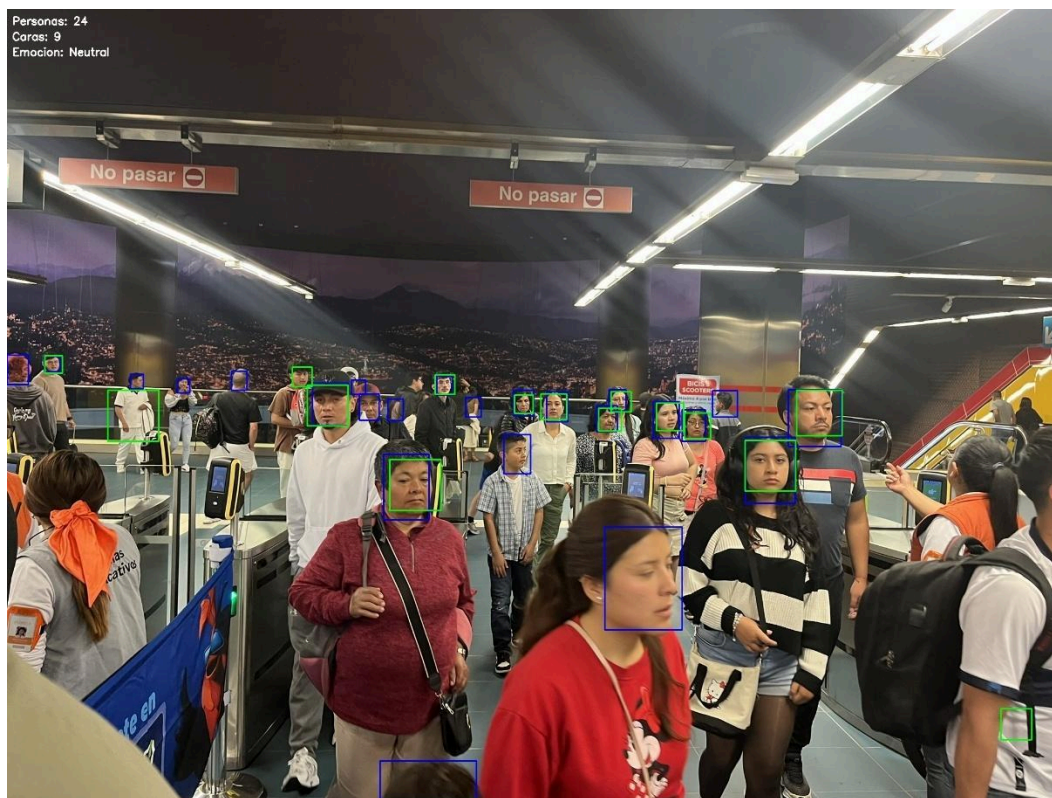


Figura 20. Escenario con basa densidad de personas

Por otro lado, en la figura 20 se observa el mismo entorno, pero con baja densidad y buena iluminación, del mismo modo el modelo demuestra una fiabilidad compacta. Se observan que las regiones de rectángulo delimitan correctamente a cada persona y el sistema HSEmotion casi logra capturar a todos los rostros principales infiriendo como emoción dominante la neutral.

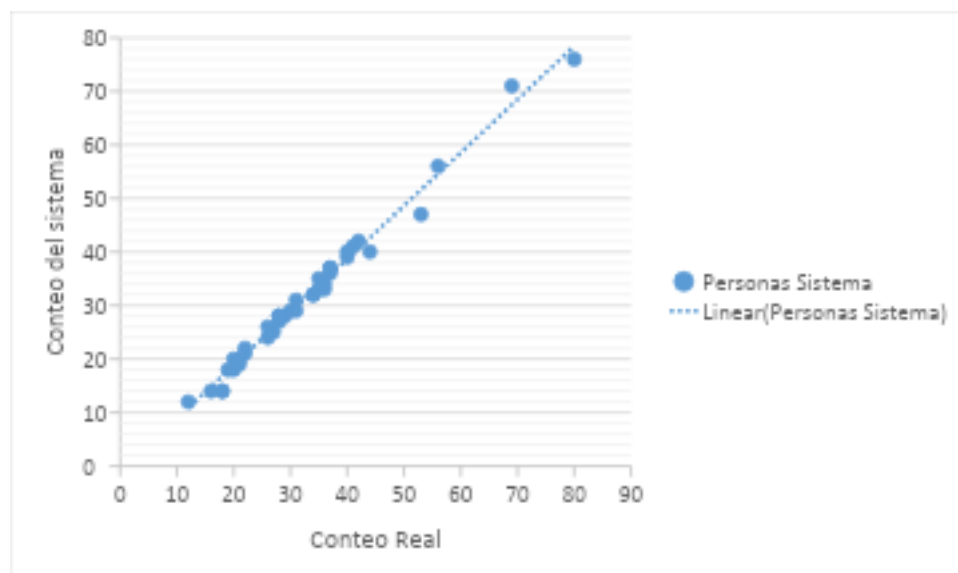


Figura 21. Análisis de regresión lineal entre el conteo manual y la detección del HSEmotion

El grafico de la figura muestra una correlación positiva entre la cantidad real de personas y la capacidad de conteo del sistema. Ajustando los datos a un modelo de regresión lineal, se obtiene la siguiente ecuación:

$$y = 0,9936x - 1,0644$$

Además, también se calculó el valor del coeficiente de determinación R^2 igual a 0.897, este valor demuestra que el comportamiento del algoritmo es estable, de igual forma el valor de la pendiente de la recta demuestra una relación de proporcionalidad estable entre la realidad y la detección.

Discusión de los resultados

Para responder a la pregunta de investigación, ¿Es factible desarrollar un sistema de visión artificial en dispositivos IoT que permita la estimación precisa de densidad y emociones en multitudes, garantizando un equilibrio entre rendimiento y costo?, el presente estudio ha diseñado un procesamiento basado en un modelo entrenado para detección de aglomeración y también el uso de la librería HSEmotion. Este análisis se centra en la capacidad del sistema para operar en escenarios no controlados, contrastando la observación manual frente a la inferencia automática para determinar la fiabilidad operativa del prototipo.

Se ejecutaron técnicas de visión computacional para dispositivos estableciendo filtros para la validación de rostros. Durante las pruebas experimentales, se expuso al sistema a un escenario de evaluación con un total de 331 rostros teóricamente visibles. De este universo, el algoritmo logró procesar y extraer características exitosamente en 281 individuos. Esto arroja una tasa de extracción del 84,89%, los datos descartados permitieron identificar que la pérdida de información no es aleatoria, sino fue una forma selectiva, el sistema filtró aquellos rostros que, por su baja resolución o ángulos extremos, no cumplían con los estándares mínimos para una clasificación emocional fiable.

Con el análisis estadístico se halló un aspecto importante respecto a la estabilidad del sistema. Al aplicar un modelo de regresión lineal entre el conteo real y el detectado, se obtuvo una ecuación con una pendiente casi unitaria y un Coeficiente de Determinación de 0,89. Esto indica que, a pesar de la subestimación natural causada por la oclusión en aglomeraciones, el sistema mantiene una proporcionalidad notablemente precisa con la realidad. En otras

palabras, el algoritmo no adivina, sino que escala su conteo de manera lineal y predecible conforme aumenta la densidad de personas, evaluando la resistencia técnica del sistema.

Este enfoque de detección selectiva y alta correlación ofrece ventajas significativas para la gestión de espacios públicos:

- **Detección Temprana de Aglomeraciones Críticas:** Dado que la correlación lineal demuestra que el sistema no se satura, es posible configurar alertas automáticas cuando la densidad de la multitud supere umbrales de seguridad, permitiendo a los operadores intervenir antes de que ocurran incidentes de aplastamiento o bloqueo de vías.
- **Privacidad:** Al centrarse en el análisis de patrones grupales en lugar de la identificación individual, el sistema protege de forma implícita la identidad de las personas presentes en el entorno, alineándose con los principios éticos de la vigilancia.
- **Optimización de Recursos:** El alto nivel de confianza reflejado por el coeficiente R^2 , permite a la administración apoyarse en métricas automatizadas para desplegar personal de seguridad o servicios de emergencia únicamente cuando la tendencia estadística lo justifica ayudando a reducir la generación de falsas alarmas.

Capítulo VI

Conclusiones

A lo largo del proyecto se desarrollaron las etapas de análisis, diseño, implementación y evaluación, permitiendo obtener resultados que responden a los objetivos anteriormente planteados. Se determinó en primera instancia que la transición de arquitecturas centralizadas en la nube hacia el procesamiento de la información en el ámbito de Edge Computing es la estrategia técnica más viable para sistemas de monitoreo masivo. La revisión teórica apoya la utilización de redes neuronales ligeras como la librería HSEmotion, que ejecuta inferencias en hardware con recursos limitados sin sacrificar la operatividad. Esto justifica la independencia del sistema respecto a infraestructuras de servidores costosas, cumpliendo con el propósito de ofrecer una solución accesible y descentralizada.

Con respecto a la implementación de la plataforma IoT, la arquitectura con la Raspberry Pi 4 demostró tener una capacidad robusta para operar en tiempo real aun con latencias críticas, administrando localmente la adquisición, el procesamiento y la transmisión de las alertas. El sistema logró tener un equilibrio operativo constante, reflejado por la fluidez en la tasa de cuadros por segundo y la uniformidad en los tiempos de inferencia durante las funciones de monitoreo. Esta consistencia afirma que los recursos utilizados fueron administrados eficientemente a pesar de los problemas de configuración inicial, evitando la degradación del servicio bajo la ejecución continua de los algoritmos de visión, validando así la idoneidad del dispositivo para operaciones de monitoreo constante.

Para finalizar, la evaluación en entornos reales dinámicos validó la estabilidad del sistema con un coeficiente de determinación de 0,987 comparando el conteo manual frente al conteo automático.

A partir de los resultados obtenidos, se concluye que, aunque existen limitaciones físicas enlazadas al entorno como la oclusión en aglomeraciones, el sistema equilibra estos factores manteniendo una tendencia cercana a la realidad. Esta solución comprueba que el prototipo es una herramienta confiable para la toma de decisiones en puntos estratégicos y dinámicos identificando patrones de comportamiento y riesgos de seguridad con un margen de error sistemático y controlado.

Recomendaciones

Para solucionar el factor de oclusión es recomendable modificar la ubicación de los dispositivos hacia una posición cenital en vez de frontal. Un ángulo de visión elevado reduciría el bloqueo visual entre personas situadas en diferentes planos de profundidad, mejorando la correlación de conteo en escenarios de alta densidad y permitiendo una segmentación más clara de los individuos.

Asimismo, dado que el modelo creado prioriza la calidad de imagen para la inferencia emocional, se sugiere mejorar las condiciones de entrada mediante la integración de sistemas de iluminación como infrarrojo o LED en el entorno de despliegue. Este cambio mejoraría el número de caras detectadas, incrementando el paso de resolución de la imagen cuando haya condiciones de luz variable, horarios nocturnos e interiores con falta de iluminación.

Desde la perspectiva de la informática, se recomienda mejorar parte de la arquitectura actual con el fin de extender el control del monitoreo seguro. Se busca integrar actuadores físicos, por ejemplo, la señalética o inclusive alarmas sonoras, que se puedan activar automáticamente cuando el sistema detecte los niveles críticos de las emociones negativas o de alta densidad. Esto permitiría cerrar el ciclo de seguridad de manera autónoma, transformando la herramienta en un sistema de respuesta activa ante emergencias.

Finalmente, desde un punto de vista operativo, se aconseja a los administradores interpretar los datos generados por el sistema como indicadores de tendencia y no como un censo de precisión absoluta. La linealidad demostrada en los resultados confiere a esta herramienta una idoneidad particular para detectar anomalías o cambios bruscos en el flujo de la masa en la práctica de la seguridad pública, son estas variaciones repentinas las que constituyen las verdaderas señales de alerta temprana necesarias para la prevención de desastres.

Trabajos futuros

Esta investigación abre la puerta a nuevas líneas de desarrollo que podrían potenciar las capacidades del sistema actual:

- **Implementación de Visión Estereoscópica o Multicámara:** Una evolución del proyecto sería la integración de un segundo módulo de cámara o la fusión de datos de múltiples nodos IoT. Esto permitiría realizar una triangulación de los sujetos, resolviendo el problema de la oclusión mediante múltiples ángulos de visión y mejorando la precisión del conteo en aglomeraciones densas.

- **Actuadores Físicos:** Para mejorar el proceso de monitoreo se sugiere hacer una conexión del modelo de análisis emocional con otros sistemas de respuesta automatizada. Por ejemplo, cuando exista una alta densidad de personas y predomine las emociones negativas el sistema debe activar señalización orientada a una posible evacuación o inclusive a generación de alertas dirigidas al personal de seguridad o administrativo. Gracias a este incremento de la funcionalidad el sistema evoluciona de un esquema pasivo de monitoreo a uno activo de gestión para las emergencias.
- **Desarrollo de un modelo emocional para multitudes:** Una restricción hallada en el estado del arte refleja que los modelos como HSEmotion son de propósito general, han sido entrenados con datasets de rostros frontales, limpios y sin obstrucciones. Sin embargo, actualmente no existe un estándar para la detección de emociones específicamente en entornos con alta densidad de personas. Por lo tanto, una línea de investigación conveniente sería la creación de un dataset propio orientado solo a rostros dentro de aglomeraciones. El objetivo es entrenar una nueva arquitectura de red neuronal que pueda aprender a deducir emociones no desde la perfección del rostro completo, sino a partir de rasgos parciales visibles en condiciones de oclusión y baja resolución. Esto sería un avance valioso, pasando de adaptar herramientas generales a tener una solución específica para el análisis de masas.

Bibliografia

Fruin, J. J. (1971). *Pedestrian planning and design*. New York: Metropolitan Association of Urban Designers and Environmental Planners.

Still, G. K. (2014). *Introduction to Crowd Science*. CRC Press.

Le Bon, G. (1895). *The Crowd: A Study of the Popular Mind*.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). *Emotional Contagion*. Cambridge University Press.

Gonzalez, R. C., & Woods, R. E. (2018). *Digital Image Processing (4th Edition)*. Pearson.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Terven, J., & Cordova-Esparza, D. (2023). A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4), 1680-1716.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169-200.

Savchenko, A. V. (2021). Facial expression recognition with small-scale datasets using lightweight CNNs. In *Proceedings of the 30th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*.

Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30-39.

OASIS Standard. (2014). MQTT Version 3.1.1. Organization for the Advancement of Structured Information Standards.

ITU-T. (2012). Overview of the Internet of things.

Hassan, N., & Ahmad, R. (2024). Benchmarking Edge Computing vs. Cloud Computing for Real-Time Data Processing: Latency and Bandwidth Analysis. *Journal of Cloud Computing Advances*, 12(4), 45-67.

Kumar, S., & Singh, A. (2023). AI-Based Crowd Density Estimation and Prediction Using IoT Devices. *International Journal of Embedded Systems*, 15(2), 112-128.

Savchenko, A. V. (2022). HSEmotion: High-speed emotion recognition library. *Software Impacts*, 14, 100433. <https://doi.org/10.1016/j.simpa.2022.100433>

Sulman, N., Sanocki, T., Goldgof, D., & Kasturi, R. (2008). How effective is human video surveillance performance? En *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)* (pp. 1-4). IEEE.

Tarasov, A., & Savchenko, A. (2019). Emotion Recognition of a Group of People in Video Analytics Using Deep Off-the-Shelf Image Embeddings. En *Analysis of Images, Social Networks and Texts* (pp. 1-12). Springer.

Mirajkar R., Garware C., Shinde G., Chavhan P., Shekhare A., Shahane S. Crowd Density Estimation Using Deep-Learning (2025)

Sujatha R., Chatterjee J.M., Pathy B., Hu Y.-C. Automatic emotion recognition using deep neural network (2025)

Sirisha D., Sambhu Prasad S., Kumar S. Computer Vision and Convolutional Neural Network for Dense Crowd Count Detection (2024)

Byabazaire J., Tirana J., Chouliaras A., Koutsos V., Aslanidis T., Panagiotidis I., Chatzopoulos D. Deep Learning and the Internet of Things: Applications, Challenges, and Opportunities (2025)

Palermo F., Casciano L., Demagh L., Teliti A., Antonello N., Gervasoni G., Shalby H.H.Y., Paracchini M.B., Mentasti S., Quan H., Santambrogio R., Gilbert C., Roveri M., Matteucci M., Marcon M., Trojaniello D. Advancements in Context Recognition for Edge Devices and Smart Eyewear: Sensors and Applications (2025)

R. Wang, Y. Hao, Y. Miao, L. Hu and M. Chen, "RT3C: Real-Time Crowd Counting in Multi-Scene Video Streams via Cloud-Edge-Device Collaboration" in IEEE Transactions on Services Computing, vol. 17, no. 04, pp. 1739-1752, July-Aug. 2024, doi: 10.1109/TSC.2024.3377156.

Kasra Aminiyeganeh, Rodolfo W.L. Coutinho, Azzedine Boukerche, IoT video analytics for surveillance-based systems in smart cities, Computer Communications, 2024.

Dadashi, Nastaran & Stedmon, A.W. & Pridmore, Tony. (2012). Semi-automated CCTV surveillance: The effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. Applied ergonomics. 44. 10.1016/j.apergo.2012.04.012.