PI1830-1-ACT-CCTV

Identificación de actividades inusuales a partir del uso de CCTV

Eder Mauricio Abello Rodríguez

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2018

PI1830-1-ACT-CCTV

Identificación de actividades inusuales a partir del uso de CCTV

Autor:

Eder Mauricio Abello Rodríguez

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director

Ingeniero Enrique González Guerrero

Comité de Evaluación del Trabajo de Grado

Ingeniero César Julio Bustacara Medina

Ingeniero Carlos Alberto Parra Rodríguez

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~PI1830-1-ACT-CCTV

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
NOVIEMBRE, 2018

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Lope Hugo Barrero Solano

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Angela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortiz Pabón

Artículo 23 de la Resolución No. 1 de junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Deseo reconocer de manera especial mi agradecimiento a mi tutor, el profesor Enrique González Guerrero, que estuvo aportando constantemente con su conocimiento y experiencia al desarrollo del trabajo de investigación.

Quiero agradecer a la empresa Controles Inteligentes en la ciudad de Bogotá, por brindar los medios tecnológicos y los recursos necesarios para la ejecución de cada una de las etapas del proyecto.

A la coordinación de parqueaderos del centro comercial Oviedo en la ciudad de Medellín, por permitir la captura de los videos, la ejecución de entrevistas y el desarrollo del protocolo experimental en sus instalaciones.

A José López, Mauricio Ciro y Diego Hoyos, por su ayuda en la captura de videos y la ejecución de las pruebas en el centro comercial Oviedo.

A los profesores del departamento de Ingeniería de Sistemas, que aportaron ideas valiosas en los seminarios de investigación que fortalecieron el resultado de la investigación.

Finalmente, a mi esposa Adriana y a mi familia, por darme todo el apoyo, la comprensión y la motivación necesaria para superar las adversidades y seguir adelante en el desarrollo del proyecto de grado.

CONTENIDO

ABSTRACT

This work presents AC-CCTV, a system of recognition of human activities oriented to the detection of suspicious events. AC-CCTV has a design oriented to rational agents, which exploits the distributed characteristic of a CCTV by developing a scalable architecture and performing the recognition by means of the information provided by multiple cameras. AC-CCTV makes the identification of activities based on the development of classifiers in 3 levels of abstraction (poses, actions and activities), where their responses are combined in a final response based on assembly methods.

RESUMEN

En este trabajo se presenta AC-CCTV, un sistema de reconocimiento de actividades humanas orientado a la detección de eventos sospechosos. AC-CCTV cuenta con un diseño orientado a agentes racionales, que explota la característica distribuida de un CCTV al desarrollar una arquitectura escalable y realizar el reconocimiento por medio de la información proporcionada por múltiples cámaras. AC-CCTV realiza la identificación de actividades a partir del uso de clasificadores en 3 niveles de abstracción (poses, acciones y actividades), en donde sus respuestas son combinadas en una respuesta final a partir de métodos de ensamble.

RESUMEN EJECUTIVO

Una de las principales herramientas tecnológicas utilizadas por las compañías de vigilancia y seguridad son los CCTV (circuitos cerrados de televisión). Por lo general, los esquemas de seguridad de las compañías suelen contar con personal especializado en medios tecnológicos, que se encuentran encargados de realizar el monitoreo constante de las cámaras y de reportar cualquier evento delictivo u sospechoso [1]. Aunque este esquema de seguridad basado en CCTV reduce costos de operación y permite obtener un registro constante de video, autores como Langois et al muestran que las labores repetitivas asociadas al monitoreo constante de las cámaras vuelven perezoso al personal de seguridad y vigilancia [2].

Con el objetivo de proporcionar ayudas tecnológicas que permitan optimizar las labores del personal de vigilancia, la identificación de actividades humanas ha sido un tema que ha cobrado gran relevancia en la comunidad científica en los últimos años [3]. Si bien el problema de detección de actividades a partir del procesamiento de imágenes ha sido ampliamente abordado en el estado del arte, la mayoría de los modelos recaen en trabajos centralizados o semicentralizados que no explotan la característica distribuida que contiene un CCTV [4].

En este trabajo de investigación se presenta AC-CCTV, un sistema diseñado con el objetivo de proporcionar un modelo de detección de actividades que use la información contextual de múltiples cámaras, y cuente con un diseño distribuido que favorezca la escalabilidad del sistema. Uno de los valores agregados de AC-CCTV con respecto a otros trabajos desarrollados en el estado del arte es su diseño basado en agentes racionales. Gracias a la naturaleza distribuida de los sistemas CCTV, los sistemas orientados a agentes surgen como una alternativa de diseño descentralizado, gracias a características como el control de recursos compartidos [5], el uso de operaciones concurrentes [6] y el cumplimiento de metas a partir del uso de estrategias cooperativas [7]. Además, la mayoría de los agentes presentes en el sistema cuentan con un patrón productor-consumidor, que permite la implementación de técnicas en procesamiento en paralelo en modo pipeline [8].

El desarrollo del modelo de AC-CCTV se encuentra basado en el trabajo de Cristiani et al [9], en donde descompone el modelo de detección de actividades a partir de 3 niveles. El nivel bajo se encarga de identificar las personas que se encuentran en la escena y de realizar la extracción de las características asociadas a la pose. En el desarrollo del módulo de bajo nivel se utilizó la librería OpenPose, un framework de código abierto que obtiene el esqueleto de la persona a partir de la posición de sus articulaciones [10].

Luego de ejecutar la extracción de la pose y el cálculo de los descriptores, la etapa de nivel medio se encarga de realizar el rastreo (tracking) de la persona en escena. Las principales herramientas que el módulo de reidentificación utiliza para el rastreo de las personas son la transformación proyectiva, que realiza la conversión de coordenadas locales en la imagen en coordenadas globales, y la extracción de una medida de similitud a partir de la comparación de color. Los descriptores que se encuentran asociados a la persona son almacenados en una lista, y posteriormente enviados al módulo de alto nivel a medida que el sistema detecta la finalización de una actividad.

La última etapa del clasificador de acciones está compuesta por los módulos de alto nivel, en donde la clasificación de actividades se divide en 4 etapas: identificación de poses, identificación de acciones, identificación de actividades y ensamble. La división de las 3 primeras etapas se encuentra sustentado en el trabajo de Saad et al, el cual define una actividad en función de sus acciones, y una acción en función de sus poses [11]. Por otra parte, la implementación de la etapa de ensamble se elaboró con base en el argumento proporcionado por Dietterich, el cual establece que el uso de diferentes modelos en la solución de problemas de clasificación, combinados a partir de técnicas de ensamble, produce mejores resultados que el uso de clasificadores de forma individual [12].

La validación del modelo de detección de actividades se realiza con ayuda del CCTV instalado en el parqueadero del centro comercial Oviedo en la ciudad de Medellín. La red CCTV del centro comercial cuenta con 1093 cámaras, las cuales se encuentran instaladas a lo largo de un parqueadero que cuenta con 4 sótanos y 3 niveles. El protocolo experimental ejecutado en el centro comercial se desarrolló con base en las entrevistas desarrolladas al personal y vigilancia, en donde se clasifican un total de 3 actividades usuales y 4 actividades inusuales.

Los resultados obtenidos en los experimentos muestran un porcentaje de precisión de 93.1% para el mejor grupo de clasificadores, en la identificación de las 7 clases de actividades definidas en el protocolo experimental. A pesar de los argumentos proporcionados por Dietterich [12], el protocolo experimental comprobó que la mejor alternativa no corresponde precisamente al uso de técnicas de ensamble. Por otra parte, aunque los algoritmos seleccionados para la identificación de actividades mostraron una precisión sobresaliente, la medición de desempeño comprobó el alto nivel de recursos computacionales que requiere algunos módulos del sistema. Los altos recursos computacionales requeridos por el sistema generan retardos que limitan el procesamiento y la generación de alarmas en tiempo real.

Con el objetivo de lograr un sistema escalable que pueda generar alertas en tiempo real, se plantea como trabajo futuro la integración de OpenPose a dispositivos de procesamiento especializado, como es el caso del stick Movidius desarrollado por Intel [13]. Además, se propone la integración de AC-CCTV con la interfaz gráfica del sistema de guiado que opera en el centro comercial.

INTRODUCCIÓN

Uno de los problemas más frecuentes con los cuales deben enfrentar las ciudades a lo largo del mundo es la inseguridad. Tan solo en el contexto colombiano, el Departamento Administrativo Nacional de Estadística (DANE) estima que, en el 2015 el 11.3% de colombianos de 15 años o más fueron víctimas de hurto [14]. Algunas ciudades como Bogotá han adoptado políticas para la reducción de estos indicadores como la creación planes de intervención, aumento del patrullaje, y el desarrollo de planes de seguridad [15]. Sin embargo, muchas instituciones privadas optan por la contratación de compañías privadas para la gestión de la seguridad y vigilancia dentro de sus organizaciones. Esta cifra se ve reflejada en un estudio generado por la Superintendencia de Vigilancia y Seguridad Privada muestra una tasa de crecimiento del sector del 42.3% entre los años 2006 y 2014 [1].

El CCTV es una de las herramientas más utilizadas dentro de las compañías de seguridad, gracias a que proporcionan registros de video y optimizan las labores del personal de seguridad [16]. Para el control y la gestión de los sistemas CCTV, las compañías suelen tener centros de monitoreo que recopilan la información de múltiples cámaras y las despliegan por medio de un arreglo de pantallas. Los centros de monitoreo son controlados por operadores de medios tecnológicos, los cuales están encargados de la vigilancia constante de las cámaras, el control de alarmas y la coordinación de actividades [17].

Una de las aplicaciones que han aprovechado el CCTV para el desarrollo de equipos de parqueadero es el sistema de guiado. El sistema de guiado consiste en un conjunto de cámaras interconectadas que identifican, por medio de algoritmos de inteligencia artificial, la ocupación de las celdas de parqueo. Algunas arquitecturas de sistemas de guiado realizan montajes de una cámara por celda de parqueo, lo cual puede llegar a sumar un total de hasta 2000 cámaras instaladas. Aunque cada una de estas cámaras puede visualizarse desde el centro de monitoreo, es imposible que los operadores tecnológicos logren realizar la inspección de todas las cámaras de forma simultánea.

En este documento se presenta ACT-CCTV, un sistema para la detección automática de actividades inusuales a partir del uso de CCTV, el cual busca brindar ayudas tecnológicas que permitan al vigilante realizar un trabajo de mayor calidad. El desarrollo del sistema se encuentra basado en un modelo orientado a agentes racionales, brindando un grado de novedad al manejar e integrar múltiples cámaras con un enfoque distribuido y el uso de estrategias colaborativas.

En la primera sección del documento se desarrolla la identificación de la problemática y la formulación del problema. A continuación, se describe la formulación de objetivos y la metodología utilizada para el desarrollo del proyecto de investigación. Posteriormente, se realiza una recopilación de los trabajos representativos asociados al problema de investigación. En las siguientes secciones se desarrolla el diseño, la implementación y la evaluación del modelo de detección, y finalmente se generan las conclusiones.

1. DESCRIPCIÓN GENERAL

1.1. Oportunidad y problemática

Muchas empresas de seguridad han adaptado los medios tecnológicos como una herramienta fundamental para la prestación del servicio de vigilancia. Uno de los principales medios tecnológicos es el Circuito Cerrado de Televisión (CCTV). Para muchas empresas, los sistemas CCTV son una herramienta efectiva de seguridad, gracias a que reducen los costos de operación y permiten la obtención de registros de video [16]. Sin embargo, autores Leman-Langois et al sugieren que este tipo de sistemas pueden volver más perezoso y dependiente al personal de vigilancia, y critican la falta de reacción de este tipo de organismos al detectar un evento criminal [2].

Con el objetivo de proporcionar ayudas tecnológicas que faciliten las labores de vigilancia, el CCTV se ha convertido en uno tema importante de investigación para el desarrollo de aplicaciones basadas en Inteligencia Artificial. Smart CCTV es un concepto que ha evolucionado en la industria en los últimos años [3]. A través del uso de algoritmos de visión por computadora, los sistemas Smart CCTV brindan información adicional que apoyan al vigilante para realizar un trabajo de mayor calidad. Dentro de los escenarios de aplicación del Smart CCTV, la detección de actividades humanas ha cobrado gran relevancia dentro de la comunidad científica [18].

El modelo clásico propuesto para la detección de actividades a partir de secuencias de video cuenta con un proceso en 2 etapas [9]. La etapa de bajo nivel permite la detección de personas y la generación de descriptores a partir de métodos de preprocesamiento y clasificación. Algunas técnicas que permiten la generación de descriptores de bajo nivel son el flujo óptico (Optical Flow) [19] y SIFT (Scale Invariant Feature Transform) [20]. A partir de la información proporcionada por el módulo de bajo nivel, la etapa de alto nivel realiza la identificación de actividades a partir de técnicas de reconocimiento de patrones o métodos de análisis espacial y temporal de las imágenes [9].

Aunque por naturaleza los sistemas CCTV permiten el manejo de múltiples cámaras, no se encuentran muchos trabajos que exploten esta característica. Algunos trabajos como el de Weinland et al [21] y Kooij et al [19] utilizan la información obtenida de múltiples cámaras, al realizar una reconstrucción tridimensional de la escena y generar descriptores de mayor nivel. Sin embargo, estos trabajos se limitan al monitoreo de una sola cámara y la mayoría de ellos propone un modelo centralizado en el procesamiento de los datos.

Dada la naturaleza distribuida de un sistema CCTV, los modelos basados en agentes son una alternativa que ofrece un valor agregado al sistema. Una de las principales ventajas de un modelo basado en agentes es que favorece la descentralización y aumenta la escalabilidad del sistema [22]. Además, los modelos basados en agentes permiten reunir la información contextual proveniente de múltiples cámaras a través del uso de estrategias cooperativas. Poten-

cialmente, el uso de una información contextual generará descriptores más robustos, logrando un mayor porcentaje de precisión en la detección.

Aunque los modelos basados en agentes ofrecen una ventaja en el campo de la identificación de actividades inusuales, la mayoría de los trabajos reportados en el estado del arte se orientan a modelos centralizados que no garantizan la escalabilidad del sistema. Trabajos como el de Ejaz et al [4] implementan arquitecturas semi-distribuidas basadas en agentes, las cuales distribuyen la carga de procesamiento a través del uso múltiples agentes. Sin embargo, las arquitecturas semi-distribuidas representan un cuello de botella que se manifiesta cuando el número de cámaras alcanzan un valor significativo, limitando su escalabilidad y favoreciendo la reducción de costos.

Una aplicación sobre la cual el número de cámaras alcanza un valor considerable es el sistema de guiado. Los sistemas de guiado están compuestos por un conjunto de sensores que determinan la ocupación en tiempo real de las plazas de un parqueadero [23]. Entre los diferentes tipos de sistemas de guiado que se pueden encontrar se destacan los compuestos por sensores y cámaras. Aunque los sistemas de guiado con sensores son más baratos y fáciles de instalar, los sistemas basados en cámaras ofrecen un valor agregado al generar un registro continuo de video por cada celda de parqueo.

1.2. Formulación del Problema

El problema informático que atacará este proyecto de investigación es la identificación de actividades inusuales a partir de técnicas de inteligencia artificial. La identificación automática de actividades inusuales será el punto de partida para la creación de ayudas tecnológicas basadas en los sistemas Smart CCTV, que generen la prevención de actos delictivos dentro del contexto de seguridad colombiano. El objetivo de estas ayudas tecnológicas es mejorar los tiempos de respuesta y el nivel de desempeño de los vigilantes, permitiendo reducir su carga laboral y el riesgo de padecer trastornos como el síndrome de desgaste ocupacional.

El enfoque del proyecto de investigación se centra en el desarrollo y la implementación de técnicas de alto nivel para la identificación de actividades inusuales. La selección del enfoque se realiza de acuerdo con el área de énfasis y al conocimiento previo del estudiante y del profesor asesor. El alcance del proyecto se limita a sistemas CCTV instalados en recintos cerrados, debido a sus cámaras permiten obtener una mayor definición de la escena y un ambiente controlado de iluminación. El paradigma que se utilizará en el diseño del modelo corresponderá al desarrollo de sistemas basados en agentes racionales, generando un grado de novedad al manejar e integrar múltiples cámaras con un enfoque distribuido y el uso de estrategias cooperativas entre los agentes.

El caso de estudio definido para el desarrollo del proyecto es el CCTV instalado en el parqueadero del centro comercial Oviedo en la ciudad de Medellín. Este caso de referencia se selecciona debido la facilidad del investigador para acceder a los videos del sistema de seguridad. Como empresa aliada del proyecto se encuentra Controles Inteligentes SAS, organización que cuenta con 7 años de experiencia en el desarrollo de equipos de parqueaderos.

2. DESCRIPCIÓN DEL PROYECTO

2.1. Objetivo General

- Diseñar un sistema para la identificación de actividades inusuales, a partir de imágenes pre-procesadas, mediante del uso de técnicas de inteligencia artificial, que pueda ser aplicado a sistemas CCTV instalados en recintos cerrados.

2.2. Objetivos específicos

1. Analizar, a partir del estado del arte, las técnicas actuales de identificación de actividades inusuales, evaluando su aplicabilidad para el contexto de sistemas CCTV instalados en recintos cerrados.
2. Diseñar un modelo basado en agentes racionales, a partir del empleo de imágenes pre-procesadas provenientes de múltiples cámaras, para la identificación de actividades inusuales a través del uso de estrategias cooperativas y técnicas de reconocimiento de patrones.
3. Evaluar el desempeño, la precisión y usabilidad del modelo propuesto, a través de su implementación parcial en un sistema CCTV enmarcado dentro del contexto colombiano de seguridad y vigilancia.

2.3. Fases de desarrollo

Las fases de desarrollo del proyecto de investigación se encontrarán definidas por cada uno de los objetivos específicos. Aunque el objetivo del proyecto está orientado al desarrollo de un modelo de detección de actividades independiente del contexto, la caracterización del caso de estudio permite evaluar en un contexto real las técnicas identificadas en el estado del arte. El caso de estudio seleccionado se encuentra de forma transversal a cada una de las etapas, cuya caracterización se desarrolla a través de un estudio cualitativo que identifican elementos como herramientas tecnológicas, así como protocolos de trabajo y de seguridad.

Investigación y Análisis

La fase de investigación se basa en los 2 niveles del modelo de Cristiani et al [9]:

Análisis de bajo nivel: La etapa de análisis del módulo de bajo nivel identifica el conjunto de técnicas y herramientas que permiten realizar el procesamiento de bajo nivel. A partir de este análisis, se seleccionará de una forma rigurosa la herramienta que se utilizará en la fase de implementación y validación. Dado que el foco del proyecto no se encuentra en el procesamiento de imágenes, la evaluación del modelo no resulta afectada en caso de que el preprocesamiento de imágenes no pueda ser ejecutado en tiempo real.

Análisis de alto nivel: A través de un análisis sistemático y riguroso, la etapa de análisis de alto nivel busca identificar los conceptos, métodos y técnicas necesarias para el desarrollo del modelo de detección de alto nivel. Entre las actividades propuestas para esta etapa se incluyen la recopilación de artículos, la elaboración de mapas conceptuales y la elaboración del documento del estado del arte.

Diseño

La fase de diseño desarrolla el modelo del sistema multiagente utilizado para la identificación de actividades inusuales. Se divide en dos etapas:

Desarrollo del sistema multiagente: La primera etapa del diseño consiste en el desarrollo de la arquitectura del sistema basado en agentes racionales. Antes de realizar el desarrollo del modelo, el investigador debe identificar 3 actividades inusuales que tengan un alto grado de relevancia dentro del caso de estudio seleccionado. El modelo del sistema multiagente será desarrollado a partir de la metodología de diseño AOPOA [24], generando énfasis en el desarrollo de estrategias cooperativas enfocadas a la solución de problemas como la transmisión y el procesamiento eficiente de la información.

Desarrollo de la inteligencia del agente: El diseño del modelo de inteligencia se elaborará a partir de la metodología de diseño iterativo e incremental basada en prototipos [25], en donde cada incremento abarca el diseño de uno o más agentes. De acuerdo con la caracterización de roles y el mapeo de requerimientos, se define para cada agente si requiere el uso de inteligencia artificial. En caso de requerir inteligencia, se debe seleccionar cuál es la técnica adecuada de acuerdo con las capacidades ofrecidas por el módulo de bajo nivel y a los resultados reportados en el estado del arte. Los resultados de selección se apoyarán con la ejecución de una prueba de concepto, la cual contendrá sets de entrenamiento y validación capturados directamente del caso de estudio.

Implementación y Evaluación Experimental

La implementación del modelo del sistema multiagente se realizará de forma parcial, debido a la limitación de tiempo de TG-MISyC impide que no sea suficiente el desarrollo de una implementación completa. En esta etapa se seleccionan cuáles agentes se implementarán en el sistema y qué aspectos de inteligencia se desarrollan en los agentes. Los tipos de agentes y las funcionalidades se seleccionan con el objetivo de validar las técnicas de alto nivel y estrategias cooperativas desarrolladas en la fase de diseño.

La evaluación del modelo se desarrolla mediante el desarrollo de un protocolo experimental, el cual busca evaluar y caracterizar el desarrollo del modelo a partir del porcentaje de precisión obtenido con respecto a los trabajos identificados en el estado del arte. El protocolo experimenta definirá la recopilación de videos a capturar el sistema CCTV. Cabe resaltar que los videos se obtienen a través de escenas actuadas en un ambiente controlado, debido a que el centro comercial se negó a que su registro de videos fuera utilizado dentro del proyecto de investigación. Como aspecto cualitativo, se evaluará el grado de utilidad del sistema percibido por las empresas aliadas y el personal de seguridad y vigilancia.

3. MARCO TEÓRICO / ESTADO DEL ARTE

Según un informe generado por la empresa IHS Market en el año 2016, se estima que en cada año el mercado de los equipos de videovigilancia crece un 7% de forma anual [26]. La reducción de costos generada por la masificación de la producción de cámaras en el mercado chino, además los avances continuos en las tecnologías de comunicación han provocado que los sistemas CCTV integren cada día un mayor número de cámaras. En ciudades como Londres, se estima que por cada 11 personas existe 1 cámara instalada dentro de un circuito CCTV [27]. Aunque en Colombia no existen estudios que estimen de forma precisa el número de sistemas CCTVs instalados, ciudades como Bogotá han fomentado políticas públicas que promueven la instalación de CCTVs en puntos estratégicos urbanos [28].

Teniendo en cuenta el uso potencial de las aplicaciones orientadas a sistemas CCTV y su uso masivo dentro de los esquemas de seguridad y vigilancia, la identificación automática de actividades humanas ha sido un tema de gran interés entre la comunidad científica y la industria [18]. El término acción humana se encuentra asociado a una gran cantidad de definiciones, que se encuentra desde un simple movimiento de articulación, hasta una secuencia compleja de movimientos efectuadas por una o más personas [29]. Por otra parte, algunos trabajos orientados al desarrollo de aplicaciones de seguridad se encuentran orientadas a detectar el comportamiento de la persona más que sus actividades, con el objetivo de determinar si la persona se encuentra desempeñando alguna acción delictiva [30].

Con el objetivo de obtener una definición específica para el concepto de actividad, la primera sección del capítulo establece una definición según los trabajos presentes en el estado del arte. La definición brindara una base sólida para el diseño del modelo de clasificación de actividades. A partir de la definición de un modelo base de detección de actividades, se realizará un análisis de los trabajos presentes en el estado del arte que identificará los algoritmos y las técnicas necesarias para el desarrollo de cada uno de los componentes del modelo.

3.1. Modelo de detección de actividades

Un modelo de abstracción que muestra la definición de actividad se encuentra en el trabajo de Saad et al, los cuales proporcionan un modelo a partir de niveles de abstracción, que define el comportamiento en términos de una secuencia de actores, actividades, secuencias y poses [11]. En el modelo de Saad et al, la pose es el elemento base en la detección del comportamiento. Así como una secuencia de poses define una acción, una secuencia de acciones define una actividad. Saad et al definen una actividad como una secuencia de actividades que se encuentran orientados al cumplimiento de un único objetivo. Esta definición sugiere que la actividad de una persona se debe medir hasta el punto en que exista un cambio en las metas de sus acciones.

El modelo de propuesto por Saad de identificar las actividades a partir de una secuencia de acciones y poses es utilizado, en mayor o menor medida, en numerosos trabajos del estado

del arte [30] [31] [32]. Sin embargo, la mayoría de los modelos pueden asociarse en un proceso que es ejecutado en dos etapas: una etapa de bajo nivel, que se encarga de realizar la segmentación de personas, identificación de objetos y generación de descriptores. Luego de que las personas son identificadas dentro de las imágenes y los descriptores son calculados, el módulo de alto nivel se encarga de realizar la identificación de la actividad a partir de técnicas de inteligencia artificial. Este modelo general de reconocimiento de actividades es propuesto en el trabajo de Cristiani et al [9], el cual propone el uso de poses, gestos, e incluso el uso de señales auditivas en el desarrollo de modelos de detección de actividades.

Esta sección analizará el modelo de detección de actividades en 3 etapas: etapa de nivel bajo, etapa de nivel medio y etapa de nivel alto. Este trabajo extiende las 2 etapas del modelo original propuesto por Cristiani et al, debido a que el contexto sobre el cual se desarrolla este trabajo de investigación obliga a ejecutar procesos adicionales orientados a realizar el rastreo de personas sobre múltiples cámaras. La Figura 1 muestra la composición del modelo extendido de actividades inusuales.

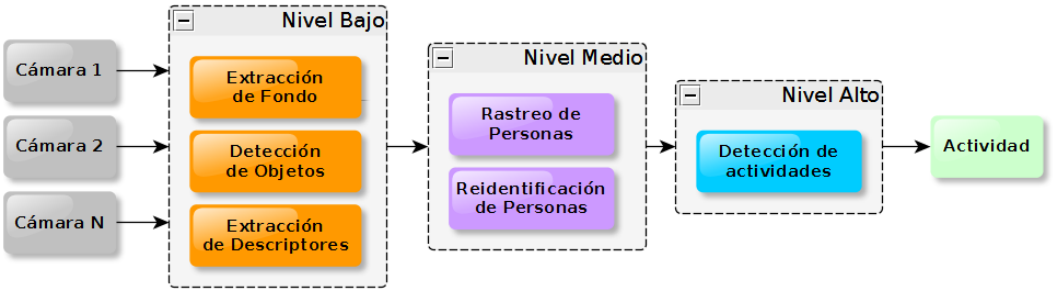

Figura 1: Modelo extendido de detección de actividades inusuales, basado en el modelo propuesto por Cristiani et al [9].

3.2. Técnicas de bajo Nivel

De acuerdo con el modelo propuesto por Cristiani et al, uno de los procesos fundamentales para realizar la detección de actividades es la segmentación de las personas que se encuentran en la escena [9]. La segmentación de personas se puede desarrollar a partir de dos grupos de técnicas. El primer grupo utiliza algoritmos de extracción de fondo para identificar los objetos móviles a lo largo de una secuencia de video. Una vez los objetos se encuentren identificados, se adiciona un módulo de clasificación que separa a las personas de los demás objetos que se encuentran en la imagen. El segundo grupo de técnicas ignora la información temporal que pueda proporcionar la secuencia de video, realizando la segmentación de las personas por cada una de las imágenes de forma independiente. Las siguientes secciones exponen las ventajas y desventajas del uso de cada uno de los grupos de técnicas.

Segmentación de personas basada en extracción de fondo

En el estado del arte se puede encontrar diversas técnicas para realizar la segmentación de las personas a partir de técnicas de extracción de fondo. Fauziah et al implementan un modelo de reconocimiento de poses a través de un algoritmo de extracción de fondo, basado en la sustracción de la imagen actual con una imagen de fondo que permanece estática a lo largo de la escena [33]. Existen modelos adaptativos que son robustos frente a cambios de iluminación y la presencia de sombras como el desarrollado por Kaewtrakulpong et al, el cual se basa en métodos probabilísticos desarrollados a partir de modelos de mezclas gaussianas (GMM por sus siglas en inglés) [34]. Otro trabajo que ejecuta la extracción de fondo basado en GMM es el modelo desarrollado por Chateau et al, que fue diseñado para realizar seguimiento de la cabeza y las manos de una persona [35]. Los resultados generados en el trabajo de Chateau et al son satisfactorios al realizar pruebas donde la iluminación ambiente no es contante.

En internet existen herramientas de uso libre y comercial que contienen el desarrollo de las librerías para implementar los algoritmos de extracción de fondo. Una de estas herramientas es OpenCV, que cuenta con más de 2500 algoritmos optimizados que abarcan desde técnicas clásicas hasta técnicas que se pueden encontrar en el estado del arte [36]. Otra de las herramientas que se encuentran disponibles para realizar la extracción de fondo es MATLAB. A diferencia de OpenCV, el uso de la herramienta MATLAB es restringido a la adquisición de una licencia.

Segmentación de personas sobre una imagen única

A diferencia de las técnicas de extracción de fondo que utilizan la información de las imágenes previas, las técnicas de imagen única utilizan únicamente la información de la imagen actual para realizar la segmentación de los objetos. Por lo general, las técnicas de imagen única requieren mayor capacidad de cómputo a comparación de las técnicas de extracción de fondo, ya que se basan la elaboración de modelos complejos que controlan aspectos como sombras, ruido en la imagen, oclusiones, entre otros [37]. Por otra parte, una de las principales ventajas del uso de técnicas de imagen única es que mitiga algunos problemas relacionados con las técnicas de extracción de fondo, entre los cuales se encuentran la identificación de objetos sin movimiento, los cambios súbitos de iluminación, fondos dinámicos en las imágenes, entre otros [38].

Uno de los métodos usados en el estado del arte para la identificación de objetos sobre imágenes estáticas son las cascadas de Haar. Las cascadas de Haar son un método de reconocimiento de objetos desarrollado por Viola et al, los cuales se basan en el concepto de características de Haar para la extracción de descriptores de una imagen [39]. La construcción del clasificador de Viola et al se realiza a partir del entrenamiento de clasificadores simples puestos en cascada, lo cual permite reducir en gran medida el tiempo de procesamiento sin afectar el desempeño del sistema [39]. El uso de las cascadas de Haar se ha implementado exitosamente en aplicaciones como el reconocimiento de placas vehicular (LPR por sus siglas en inglés), en el cual Peng et al reportan porcentajes de detección de 93% [40]. Por otra parte, el uso de las cascadas de Haar en la detección de personas está documentado en trabajos como

el de Siala et al, los cuales utilizan el modelo para contabilizar el número de personas que se encuentran presentes dentro de una imagen [41, 42]. Los experimentos desarrollados por Siala et al reportan una precisión del 85%, al utilizar diferentes clasificadores para realizar la detección de una persona en diferentes poses.

Otros autores han evaluado el uso de algoritmos basados en aprendizaje profundo (deep learning) para realizar la segmentación de las personas que se encuentran sobre una imagen. Angelova et al evalúan el desempeño de un clasificador basado en aprendizaje profundo, diseñado para la detección de peatones que circulan sobre espacios abiertos [43]. El diseño del clasificador de Angelova et al está basado en la adaptación del modelo de aprendizaje profundo propuesto por Benenson et al, el cual mejora en gran medida los tiempos de respuesta con respecto a los modelos convencionales. En su trabajo, Benenson et al logran realizar el procesamiento de imágenes a una tasa de 50 cuadros por segundo, utilizando una máquina de escritorio convencional que cuenta con una tarjeta gráfica instalada. Por otra parte, el desarrollo de Kinect y otros modelos que detectan el esqueleto de la persona a partir de cámaras estereoscópicas y sensores de profundidad, se han efectuado estudios que demuestran la posibilidad de detectar actividades en las personas a partir del análisis de la secuencia de sus poses. El SDK proporcionado por Microsoft para el dispositivo Kinect permite obtener la posición en el plano (x, y, z) de cada una de las articulaciones, además de obtener acceso a las imágenes emitidas por las cámaras RGB e infrarrojas [44]. Le et al utilizan la información proporcionada por el SDK para el desarrollo de un modelo de detección de acciones, cuyo porcentaje de precisión se encuentra alrededor del 80% [45].

Aunque la mayoría de los modelos de extracción del esqueleto se basan en el uso de cámaras estereoscópicas y sensores de profundidad, en los últimos años se han desarrollado modelos que han mostrado excelentes resultados al realizar la detección de un esqueleto por medio de una única imagen. OpenPose es un framework de código abierto desarrollado a partir de algoritmos de aprendizaje profundo y mapas de calor, el cual permite obtener el esqueleto de la persona, puntos faciales del rostro y la posición de los dedos en las manos [10]. A pesar de su reciente lanzamiento, autores como Qiao et al han efectuado estudios de cómo generar sistemas de comparación de poses orientados al reconocimiento de gestos [46]. OpenPose cuenta con interfaces de programación desarrolladas en lenguajes C++ y Python. La Tabla 1 muestra las características principales de los trabajos analizados en el estado del arte, correspondientes al procesamiento de bajo nivel.

3.3. Técnicas de Nivel Medio

Cristiani et al establecen en su modelo que uno de los aspectos más importantes en el proceso de identificación de actividades es el proceso de rastreo (tracking) de la persona en escena. Cuando la escena se encuentra monitoreada por una sola cámara, el rastreo de una persona se realiza mediante algoritmos que generen algún grado de robustez frente a cambios de iluminación y oclusiones. Uno de estos algoritmos es el filtro de Kalman, el cual es un método de rastreo que obtiene un probabilístico usado para predecir la ubicación del objeto en los siguientes instantes de tiempo [47]. El filtro de Kalman ha sido utilizado en trabajos como el de Mirabi et al, los cuales desarrollaron un modelo de rastreo de personas para espacios abiertos que es robusto frente a oclusiones [48].

Tabla 1: Recopilación de trabajos orientados al procesamiento de bajo nivel

Autor	Categoría	Técnica Utilizada	Velocidad de Respuesta del Modelo	Framework de Desarrollo
Fauziah et al	Extracción de Fondo	Extracción de Fondo estática	Alta	No Reporta
Kaewtrakulpong et al	Extracción de Fondo	Extracción de Fondo GMM	Alta	No Reporta
Chateau et al	Extracción de Fondo	Detección de tonos de piel en la imagen	Alta	No Reporta
Siala et al	Imagen única	Cascadas de Haar	Media	No Reporta
Angelova et al	Imagen única	Deep Learning	Alta	No Reporta
Quiao et al	Imagen única	Deep Learning, Mapas de Calor	Baja	OpenPose
Lee et al	Imagen única	Visión estereoscópica	Media	Kinect

Aunque el filtro de Kalman proporciona un método para realizar el rastreo de una persona para un sistema mono-cámara, la mayoría de CCTVs realizan el monitoreo de un espacio por medio de un arreglo de múltiples cámaras. La reidentificación de personas es un proceso en el cual un sistema detecta el paso de una persona a través de múltiples cámaras. La reidentificación es un problema que debe lidiar con varios retos, entre los cuales se encuentran los cambios de iluminación, la variación en los ángulos de visión y las oclusiones [49]. Los algoritmos de reidentificación de personas pueden clasificarse en 2 grupos, dependiendo si las imágenes emitidas por las cámaras se encuentran o no sobrelapadas. Cabe resaltar que los CCTV que cuentan con cámaras sobrelapadas son más robustos al momento de realizar la reidentificación, debido a que la persona nunca desaparece de la escena.

Algoritmos de rastreo utilizando cámaras no sobrelapadas

La mayoría de los métodos que soportan el manejo de cámaras no sobrelapadas se pueden clasificar en 2 grupos, que se encuentran asociados al tipo de descriptor que procesan: descriptores de color y textura, y descriptores de forma.

Los algoritmos basados en descriptores de color y textura se centran en identificar características de la ropa para realizar el rastreo. Un ejemplo del uso de descriptores de color es el desarrollado por Jang et al, los cuales calculan un histograma de color de la parte superior e inferior de las personas que se encuentran en la escena [50]. A través de una medida de similitud, el sistema puede generar una asociación entre 2 histogramas y determinar si estos pertenecen a una misma persona. Aunque el uso de los descriptores de color proporciona un método directo para realizar la reidentificación, una de las desventajas de este método es cuando dos personas se encuentran vestidas de forma idéntica.

Por otro lado, los algoritmos basados en descriptor de forma buscan obtener características que sean invariantes al color, lo cual genera robustez frente a cambios de iluminación y permiten discriminar personas que se encuentran vestidas de forma idéntica. Las técnicas por forma utilizan descriptores asociados a la silueta de la persona para realizar la reidentificación. En su artículo, Aziz et al utilizan descriptores SURF y SIFT para realizar la reidentificación, combinado con la extracción de posición de la persona a partir de la detección de su cabeza [51]. Aunque los algoritmos de forma generan robustez frente a los cambios de color entre cámaras, su desempeño puede verse comprometido si las cámaras no proporcionan el mismo ángulo de visión de la persona. Otros ejemplos de descriptores de forma son la relación alto-ancho de la persona, ubicación de las extremidades, señales de distancia del contorno, entre otros [52].

Algoritmos de rastreo utilizando cámaras sobrelapadas

Los algoritmos basados en descriptores de color y forma son independientes de la posición del objeto y no tienen en cuenta su relación espacial dentro de la escena. Estos algoritmos son útiles en el caso de que no se conozca la información topológica del CCTV, y no se cuente sobrelapamiento entre las cámaras instaladas dentro del sistema. Por otra parte, los sistemas que cuentan con cámaras sobrelapadas permiten rastrear la transición de una persona entre dos cámaras, ya que ésta nunca desaparece de escena. En el momento en que dos o más cámaras detectan una persona, se pueden utilizar algoritmos de reproyección para medir la proximidad espacial de los objetos y establecer una probabilidad de que todas las cámaras estén visualizando el mismo individuo.

Uno de los métodos para obtener la proximidad espacial entre 2 objetos es relacionar su posición dentro de la imagen con respecto al plano global, a través del uso de una transformación proyectiva u homografía. La homografía es una transformación que determina una correspondencia entre dos figuras geométricas planas, de forma que a cada uno de los puntos y las rectas de una de ellas le corresponden, respectivamente, un punto y una recta de la otra [53]. A través de una calibración de 4 puntos sobre cada cámara, se puede definir una matriz que permite transformar un punto en (x, y) ubicado en la imagen, sobre un punto (x, y) que determina la posición global de cada persona y viceversa.

La homografía es una técnica utilizada en trabajos como el de Eshel et al, en donde determina si dos individuos capturados por diferentes cámaras corresponden a la misma persona a través del cálculo de su posición global [54]. En otro trabajo relacionado orientado al estudio de multitudes, Khan et al utiliza la transformación proyectiva para realizar el rastreo de varios individuos dentro de una escena, los cuales se encuentran monitoreados a través de diferentes cámaras [55]. Una de las ventajas que Khan menciona en su trabajo acerca del uso de la homografía es la eliminación del uso de descriptores basados en color que pueden variar dependiendo de las características intrínsecas de cada cámara, y el uso de descriptores de forma que dependen de la ubicación de la cámara y del objeto dentro de la escena.

Aunque la homografía es un modelo que puede ser suficiente para determinar la posición global del objeto dentro de la imagen, existen otras técnicas que incluyen dentro del modelo aspectos como la distorsión de la cámara y el sesgo (skew). El modelo de cámara estenopeica (Pinhole Camera Model) es una descripción matemática que determina la relación matemáti-

ca entre una coordenada ubicada en el espacio tridimensional y su proyección dentro de la imagen. El modelo de la cámara estenopeica es utilizado ampliamente en trabajos relacionados con computación gráfica y realidad aumentada. Un ejemplo es el trabajo desarrollado por Zhu et al, los cuales diseñaron un sistema de entrenamiento para la milicia basado en el modelo de la cámara estenopeica, en donde el sistema ubica objetivos dentro de una escena real que la persona debe disparar [56]. Aunque con algunos ajustes el modelo de cámara estenopeica se comporta de forma similar a la transformación proyectiva, el modelo de cámara estenopeica se usa principalmente para proyectar un objeto tridimensional sobre una imagen de dos dimensiones. La Tabla 2 muestra las características principales de los trabajos analizados, orientados a realizar el procesamiento de nivel medio.

Tabla 2: Recopilación de trabajos orientados al procesamiento de nivel medio

Autor	Categoría	Técnica Utilizada	Robustez frente al ángulo de visión	Uso de información topológica
Jang et al	Cámaras no sobrelapadas	Comparación de Histogramas de color	Si	No
Aziz et al	Cámaras no sobrelapadas	Comparación de descriptores SURF y SIFT	No	Si
Eshel et al	Cámaras sobrelapadas	Homografía	Si	Si
Khan et al	Cámara sobrelapadas	Homografía	Si	Si

3.4. Técnicas de Nivel Alto

Una vez la persona se encuentra identificada en la escena y se haya realizado el rastreo de su paso por múltiples cámaras, el sistema realiza la medición de descriptores sobre la secuencia de imágenes, y a partir de estos descriptores se ejecuta la identificación de la actividad. El estado del arte propone dos grupos de técnicas para realizar el reconocimiento de actividades sobre una persona. El primer grupo corresponde al modelo propuesto por Cristiani et al, el cual establece la implementación de un método de análisis de trayectoria como base para realizar el reconocimiento de actividades. El segundo grupo de técnicas se basan en el modelo propuesto por Saad et al, el cual establece que una actividad puede identificarse a partir del análisis de poses ejecutadas por una persona.

Reconocimiento de actividades a partir del análisis de trayectoria

El estado del arte cuenta con diversos trabajos que realizan la clasificación de actividades a partir del análisis de trayectoria. Un aspecto común de los sistemas de reconocimiento basados en el análisis de trayectoria es que se limitan a clasificar la actividad entre 2 grupos, los cuales por lo general corresponden a actividades usuales o inusuales. Esta limitación tiene como sustentación el bajo nivel de información que puede entregar la trayectoria acerca de una actividad, en contraste con la información que proporciona una secuencia de poses. Al igual que muchas de las técnicas de inteligencia artificial, las técnicas de análisis de trayectoria pueden clasificarse en algoritmos de aprendizaje supervisado y no supervisado.

Las técnicas supervisadas requieren de una base de datos de actividades analizadas por personas expertas y etiquetadas a partir del comportamiento de la persona. Un ejemplo del análisis de trayectoria a partir de técnicas supervisadas es el desarrollado por Ng et al, el cual realiza la detección de la actividad por medio de un árbol de decisión [57]. En su modelo, Ng et al desarrolla el concepto de análisis de actividades basado en eventos, el cual identifica aspectos de la trayectoria asociados al contexto como la región de entrada, la región de salida y el tamaño del objeto. En el modelo, Ng et al utilizan estos aspectos adicionales como entrada al sistema de clasificación. Por otra parte, otros autores como Hao-Zhe et al desarrollan un modelo de análisis de la trayectoria utilizando modelos de clasificación de series de tiempo, en el cual se implementan técnicas clásicas como lo son las cadenas ocultas de Markov [58].

Por otro lado, el uso de métodos no supervisados tiene la característica de que no necesitan el conocimiento de personas expertas, ya que se limita a medir el grado de diferencia entre una nueva actividad y las actividades anteriores. Si suponemos que las actividades comunes son actividades que son ejecutadas con algún grado de repetición por las personas que transitan por la escena, se puede generar una medida de divergencia en la trayectoria que establezca, con algún grado de certeza, si la actividad puede ser catalogada como usual o inusual. Esta técnica es utilizada en el trabajo de Calderara et al, la cual establece una medición de divergencia a partir del análisis de trayectoria por medio de celdas de Voronoi [30]. Según el análisis de los autores, el uso de celdas de Voronoi permite reducir la incertidumbre al realizar el análisis de trayectoria, a comparación de una grilla contenida dentro de un plano cartesiano.

Reconocimiento de actividades a partir del análisis de poses

A diferencia del análisis de trayectoria, el análisis de las poses corporales permite obtener descriptores adicionales de la actividad que permiten realizar una detección más precisa. Los métodos de extracción de descriptores orientados al análisis de poses se pueden clasificar en 2 grupos principales: análisis a partir de siluetas o análisis a partir del esqueleto de la persona.

El primer grupo de análisis es la extracción de características de la silueta o el contorno de la persona. La silueta de la persona puede extraerse de una forma directa a través de algoritmos de extracción de fondo. Dependiendo del método de análisis del módulo de alto nivel, los algoritmos de procesamiento pueden generar información acerca del centroide, el contorno de la imagen, los contornos de la silueta, entre otros. En su artículo, Chaaraoui et al desarrollaron un método de análisis de actividades por medio de la identificación de poses, las cuales son organizadas en el tiempo y comparadas con una base de datos de secuencia de poses a través del clasificador K-Nearest Neighbor [52]. Uno de los aspectos claves que aporta el trabajo de Chaaraoui es el concepto de Poses Clave (Key Poses). En vez de realizar un análisis sobre toda la secuencia de poses ejecutadas por una persona, se identifican las poses más representativas a partir de métodos de aprendizaje no supervisado (K-Means). El uso de Poses Clave en el desarrollo del modelo limita el número de poses analizadas en el sistema a un tamaño fijo, y reduce el efecto que el ruido pueda generar dentro del procesamiento del video.

El otro grupo de análisis para la detección de actividades es la extracción de características a partir del esqueleto de la persona. En la sección de descripción de las técnicas de nivel bajo, se describieron métodos para realizar la extracción del esqueleto a partir de cámaras estereoscópicas (Kinect) o modelos basados en aprendizaje profundo (deep learning). Por lo general,

el esqueleto de una persona se encuentra descrito por un arreglo de puntos en 2 o 3 dimensiones, en el cual cada punto describe la posición de una de las articulaciones del esqueleto. A partir del arreglo de puntos, autores como Gedat et al obtienen los descriptores a partir de los ángulos que forman las articulaciones, argumentando que estos descriptores son invariantes de la posición y escala [32]. Por otra parte, autores como Du et at utiliza como característica la posición relativa de los puntos asociados al esqueleto, ejecutando una normalización cuya base dependerá de la posición mínima y máxima de las articulaciones durante el desarrollo de la actividad [31].

El estado del arte muestra diversas técnicas de clasificación mediante inteligencia artificial, que pueden ser aplicados para la identificación de actividades a partir de secuencias de poses. El modelo desarrollado por Du et al utiliza el concepto relacionado de mapas de distancia de articulaciones (JDM por sus siglas en inglés), el cual codifica la información espaciotemporal de una secuencia de poses en una única imagen. A partir de la generación de las imágenes, se puede aplicar técnicas de clasificación que han demostrado excelentes resultados como las redes neuronales convolucionales (CNN por sus siglas en inglés) [59]. A través del modelo propuesto, Du et al lograron obtener porcentajes de precisión superiores al 90% al clasificar las secuencias en 11 categorías de acciones [31] [60]. Por otra parte, Xia et al realiza la clasificación de las actividades a partir del análisis de secuencias de poses por medio de cadenas ocultas de Markov. Para una base de datos que contiene 9 acciones, Xia et al reportan porcentajes de precisión del modelo superiores al 90% [61]. Así mismo, Liu et al implementan un modelo de reconocimiento de actividades a partir de acciones, realizando la extracción de características a partir del método BoW (Bolsa de palabras por sus siglas en inglés) [62]. El modelo desarrollado por Liu et al reporta porcentajes de precisión hasta del 98%. La Tabla 3 muestra la recopilación de los trabajos analizados en el estado del arte, que corresponden al procesamiento de la etapa de nivel alto.

Tabla 3: Recopilación de trabajos orientados al procesamiento de nivel alto

Autor	Categoría	Tipo de clasificación	Técnica de IA Utilizada	Descriptores Utilizados	Uso de poses clave
Ng et al	Trayectoria	Aprendizaje supervisado	Árbol de decisión	Detección de eventos en la escena	No aplica
Hao-Zhe et al	Trayectoria	Aprendizaje supervisado	HMM	Secuencia de puntos de trayectoria	No aplica
Calderara et al	Trayectoria	Aprendizaje no supervisado	Medición de divergencia	Trayectoria - Celdas de Voronoi	No aplica
Chaaraoui et al	Siluetas	Aprendizaje supervisado	K-NN	Parámetros de la silueta de la persona	Si
Gedat et al	Poses	Aprendizaje no supervisado	HMM	Ángulo de articulaciones	Si
Du et al	Poses	Aprendizaje no supervisado	CNN	Posición de las articulaciones	No
Xia et al	Poses	Aprendizaje no supervisado	HMM	Histograma de posición de articulaciones	Si
Liu et al	Acciones	Aprendizaje supervisado	BoW	Uso de n-gramas. N=1, 2 ,3	No

3.5. Procesamiento multicámara y uso de agentes en el reconocimiento de actividades

Debido a la característica multicámara que poseen los sistemas CCTV, se pueden encontrar múltiples trabajos en el estado del arte que explotan la información proporcionada por múltiples cámaras para mejorar el modelo de detección de actividades. Los modelos multicámara para la detección de actividades se pueden clasificar en 2 grupos. El primer grupo corresponde a los modelos que utilizan la información de múltiples cámaras para la reconstrucción de un modelo tridimensional que permita obtener descriptores de mayor nivel. La reconstrucción tridimensional es utilizada en trabajos como el de Kooij et al [19] y Weinland et al [21]. Sin embargo, estos y otros trabajos que utilizan la reconstrucción tridimensional se limitan al monitoreo de una única escena, y por lo general proponen un modelo centralizado para el manejo de los datos. El otro grupo de modelos multicámara corresponde a las cámaras que se encuentran distribuidas en múltiples escenas, y que requieren técnicas de reidentificación que permitan realizar el rastreo de una persona a lo largo de múltiples cámaras. En su trabajo, Wang et al menciona una lista de aspectos que se deben atacar al momento de desarrollar este tipo de modelos, entre los cuales se encuentran el cálculo de parámetros intrínsecos y extrínsecos de las cámaras, el uso de la topología del sistema, el rastreo de personas, entre otros aspectos [63].

A pesar de que la característica multicámara es una constante dentro de los sistemas CCTV, son pocos los trabajos que propone modelos basados en metodologías distribuidas que permitan obtener algún grado de escalabilidad en el sistema. Algunos autores han propuesto el desarrollo de modelos de detección de actividades, gracias a que favorece el desarrollo de un diseño descentralizado y aumenta la escalabilidad del sistema [22]. En su trabajo, Ejaz et al implementan un modelo basado en sistemas multiagentes, los cuales realizan de forma distribuida la detección de actividades anormales a través de algoritmos basados en flujo óptico [4]. Por medio del establecimiento de un sistema jerárquico, los agentes reportan la detección de una actividad inusual a un agente controlador, el cual se encuentra encargado de reportar al operador del CCTV.

Aunque el modelo de Ejaz et al desarrolla modelo basado en un sistema multiagente con algún nivel de escalabilidad, el modelo no establece protocolos de cooperación que permitan aprovechar la información contextual de múltiples cámaras. En su trabajo, Remagnino et al desarrollaron un modelo que busca que exista continuidad en la información procesada, por medio de mecanismos de cooperación entre los agentes [64]. Las técnicas de cooperación desarrolladas por Remagnino et al permitan elaborar el rastreo de una persona al pasar por medio de múltiples cámaras. Sin embargo, uno de los problemas identificados en el modelo es el uso de algoritmos de trayectoria en la identificación de actividades, lo que limita el número de clases que puede clasificar de acuerdo con lo expuesto en la sección de técnicas de nivel alto.

3.6. Técnicas de Ensamble

Los trabajos encontrados en el estado del arte realizan la identificación de la actividad a partir de la construcción de un único sistema inteligente, que es entrenado y caracterizado a través

de sets de entrenamiento y validación. Sin embargo, autores como Dietterich han argumentado que el uso de diferentes modelos en la solución de problemas de clasificación, combinados a partir de técnicas de ensamble, produce mejores resultados que los modelos individuales [12]. El argumento que sustenta esta afirmación se encuentra en el campo estadístico, dado que el uso de métodos de ensamble mitiga el efecto generado por problemas como el sobre entrenamiento y los mínimos locales. Si existe un grado de diversidad en los clasificadores, el uso de técnicas de ensamble une las respuestas de cada una de las hipótesis, reduciendo el riesgo de seleccionar el clasificador equivocado.

Una de las técnicas más usadas para la implementación de las técnicas de ensamble en un sistema inteligente es el voto ponderado [65]. En esta técnica de ensamble, a cada respuesta del clasificador se le asigna un factor de ponderación, que puede estar relacionado con el nivel de precisión validado en el modelo. Luego de culminar con el proceso de votación, la clase ganadora corresponderá a la que presente un mayor número de votos. Por otra parte, las técnicas de ensamble proporcionan mecanismos para generar múltiples hipótesis a partir de un único modelo de clasificación, manipulando únicamente los datos utilizados en el set de entrenamiento. Este método se conoce como Bagging, en donde cada clasificador es entenado a partir de la selección aleatoria de un subconjunto del set de entrenamiento [12]. Una variante de este método es el AdaBoost desarrollado por Freud et al, en donde cada nuevo entrenamiento se realiza con mayor énfasis en los ejemplos en los que el clasificador anterior presentó errores. La técnica AdaBoost ha sido implementado exitosamente en algoritmos de detección de objetos, en combinación con la extracción de conjuntos de descriptores conocidos como las características de Haar [39].

Con el objetivo de desarrollar un modelo que solucione las limitaciones de los modelos propuestos por Ejaz et al y Remagnino et al, el trabajo de investigación propone la elaboración de AC-CCTV, de un sistema orientado a agentes racionales basado en el modelo de reconocimiento de actividades propuesto por Cristiani et al [9]. Con el objetivo de probar la hipótesis generada por Dietterich, el modelo AC-CCTV evaluará la precisión del sistema a partir de la integración de técnicas de ensamble. El diseño del sistema multiagente se desarrollará en el capítulo 5 de este documento. El diseño del modelo de inteligencia basado en el análisis del estado del arte se desarrollará en el capítulo 6.

4. CARACTERIZACIÓN DEL CASO DE ESTUDIO

El caso de estudio seleccionado para el proyecto de investigación es el sistema CCTV instalado en el parqueadero del centro comercial Oviedo de la ciudad de Medellín. Como empresa tecnológica que proporciona la instalación y el mantenimiento del sistema es Controles Inteligentes, empresa colombiana especializada en la fabricación e instalación de equipos especializados en control y monitoreo de parqueaderos. El CCTV instalado en el centro comercial hace parte del sistema de guiado, el cual consiste en un conjunto de cámaras y pantallas informativas instalados a lo largo del parqueadero. Las cámaras del sistema de guiado están orientados a detectar la presencia de vehículo en cada una de las celdas de parqueo, con el objetivo brindar información oportuna al usuario que permita reducir los tiempos de estacionamiento [23]. A diferencia de otros sistemas que cuentan con sensores, el uso de cámaras como sensores de ocupación permite obtener una secuencia de video desde el momento de ingreso de un vehículo hasta su retiro.

En particular, el centro comercial Oviedo en la ciudad de Medellín cuenta con un sistema de guiado operado con sensores de imagen, que se encuentran instalados a través de una red CCTV. La red CCTV del centro comercial cuenta con 1093 cámaras, instaladas a lo largo de un parqueadero que cuenta con 4 sótanos y 3 niveles. En las siguientes secciones se realizará una descripción de la arquitectura del sistema, las características de los equipos instalados y las entrevistas efectuadas al personal de seguridad.

4.1. Arquitectura del sistema

El sistema de guiado instalado en el centro comercial Oviedo cuenta con un conjunto de cámaras e indicadores de ocupación ubicados a lo largo del parqueadero. Cada una de las cámaras que integran el CCTV realiza el monitoreo de una única celda de parqueo, por lo que el número de cámaras que contiene el sistema será equivalente al número de celdas presentes en el parqueadero. Un riel ubicado en el centro del carril brinda los puntos de apoyo para la instalación de las cámaras y los sensores de ocupación. Los sensores de ocupación cuentan con un código de colores que indican verde en caso de que exista una celda disponible en la zona, o rojo en el caso contrario. Adicionalmente, el sistema de guiado cuenta con un conjunto de pantallas informativas ubicadas en las intersecciones, que permiten guiar al conductor hacia las zonas que se encuentran con mayor disponibilidad.

Las cámaras que componen el sistema CCTV cuentan con una interfaz de comunicación ejecutada bajo protocolo TCP/IP, lo cual permite el uso de interfaces de red para realizar la captura de video. Las cámaras instaladas en el centro comercial Oviedo son compatibles con la capa de comunicación CGI de Foscam [66]. El total de las cámaras que componen el sistema CCTV es de 1087, las cuales se encuentran comunicadas a través de una red de conmutadores instalados a lo largo del parqueadero. La Tabla 4 describe las características técnicas de cada una de las cámaras que componen el sistema.

Tabla 4: Características técnicas de las cámaras instaladas en el CCTV del centro comercial

Resolución	640 x 480 píxeles
Cuadros por segundo	2 FPS
Capa de comunicación	Protocolo CGI
Formato Multimedia	MJPEG
Tipo de conexión	POE
Voltaje de alimentación	12 voltios
Compensación de Luz de Fondo	Si
LEDs Infrarrojos	No

Para realizar el procesamiento de imágenes que detecta la presencia de vehículo en la celda, cada cámara cuenta con un sistema de cómputo que implementa algoritmos de inteligencia artificial. La información entregada por los sistemas de cómputo determinará el color con el cual se encenderán los indicadores de iluminación, así como el conteo que se mostrará en cada una de las pantallas informativas. Existen 2 arquitecturas que Controles Inteligentes utiliza para la implementación de los sistemas de cómputo. Una de estas arquitecturas se basa en un diseño distribuido basado en tarjetas Raspberry Pi, donde cada una de las tarjetas realiza el procesamiento de hasta 8 celdas de parqueo. La segunda arquitectura se basa en un diseño semi-distribuido, en el cual el procesamiento de imágenes se realiza bajo un arreglo de servidores ubicados dentro de la red. Para el caso específico del centro comercial Oviedo, el sistema está compuesto por una arquitectura semi-distribuida que cuenta con dos servidores de procesamiento.

Cada uno de los sistemas de cómputo instalados en el sistema se encuentran conectados a un servidor principal, que se encarga de guardar los parámetros de configuración del sistema y de proporcionar la interfaz de comunicación con el sistema de monitoreo. En el caso de la arquitectura semi-distribuida, el servidor principal cuenta con menores capacidades que los servidores de procesamiento, dado que no requiere la implementación de algoritmos de inteligencia artificial. Las características técnicas del servidor principal y de los servidores de procesamiento se puede visualizar en la Tabla 5. Por otra parte, el sistema de monitoreo proporciona la interfaz de usuario que permite a los vigilantes conocer el estado de ocupación de las celdas en tiempo real, visualizar los videos de cada una de las cámaras, realizar búsquedas de registros y obtener informes de estadísticas y permanencia. La Figura 2 muestra el diagrama físico de la arquitectura.

4.1. Análisis de las imágenes

El sistema CCTV instalado en el sistema de guiado se diferencia de los CCTVs convencionales por el gran número de cámaras instaladas y su poca área de cobertura. Con el objetivo de obtener una imagen que permita la digitalización de la placa a través algoritmos de reconocimiento de caracteres (LPR por sus siglas en inglés), las cámaras usualmente son instaladas de tal forma que el mayor porcentaje de la imagen corresponde al vehículo, dejando poco espacio para la captura de las acciones que ocurren a su alrededor.

Tabla 5: Características técnicas del servidor principal y los servidores auxiliares del sistema de guiado

Característica	Servidor Principal	Servidor Auxiliar
Procesador	Intel Core i7-3930K @3.2GHz 12 núcleos	Intel Core i7-3930K @3.2GHz 12 núcleos
Memoria RAM	24 GB	24 GB
Disco	100 GB	6TB
Tarjeta de Video	No	NVIDIA GeForce GT 610
Tarjetas de red	2	1

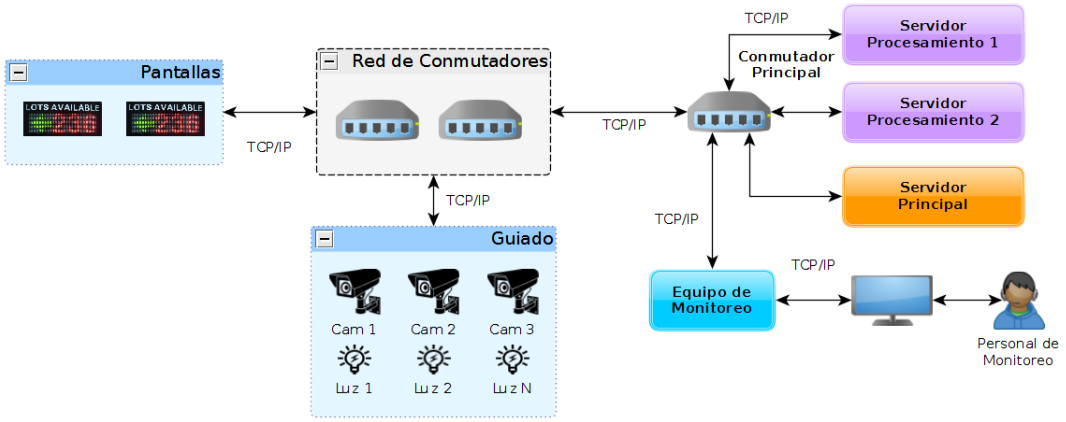

Figura 2: Diagrama físico de la arquitectura del sistema instalado en el centro comercial Oviedo.

La Figura 3 muestra ejemplos de imágenes obtenidas por el sistema de guiado del centro comercial Oviedo. De la Figura se puede observar que la poca área de cobertura genera algunos inconvenientes, ya que en muchas ocasiones se genera una captura parcial de la persona (Figura 3b). Esta característica del sistema propone que el modelo propuesto debe permitir la identificación de actividades, en personas cuya captura del cuerpo se haya elaborado de forma parcial.

Otra característica que se puede observar en las imágenes que se muestran en la Figura 3 es el sobrelapamiento que existe entre las cámaras que se encuentran contiguas. En la sección correspondiente al estudio del estado del arte, se observó que la reidentificación de personas en cámaras con sobrelapamiento es más robusta que las cámaras sin sobrelapamiento, debido a que la persona no desaparece dentro de la escena. Esta condición genera un escenario favorable para el sistema CCTV, ya que además se conoce de antemano la topología de la instalación. Adicionalmente, la presencia de marcas comunes entre las cámaras como las líneas del piso y los tope llantas, permite validar el funcionamiento de los algoritmos de transformación proyectiva (homografía) que pueden ser usados en la reidentificación de personas.

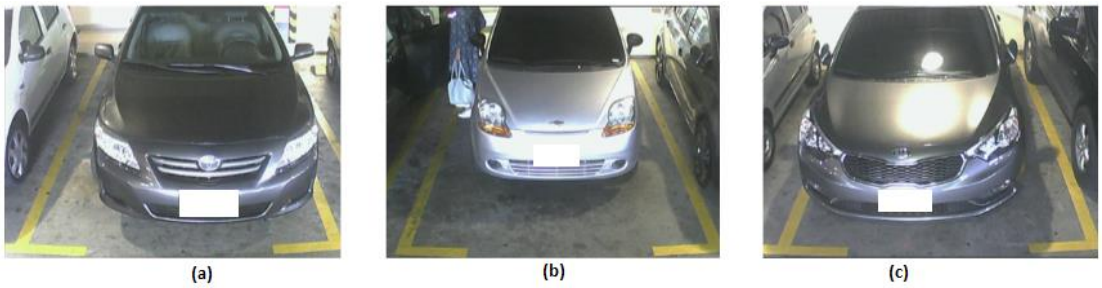

Figura 3: Ejemplos de imágenes capturadas en sistema CCTV del centro comercial Oviedo

4.2. Análisis de operación del CCTV y entrevistas al personal de seguridad

La operación del CCTV que compone el sistema de guiado de Oviedo es ejecutada por 2 personas desde un cuarto de monitoreo, la cual se ejecuta las 24 horas al día en los 7 días de la semana. Adicional a la operación del sistema de guiado, el personal de monitoreo se encuentra encargado de dirigir las acciones del personal de vigilancia que se encuentra en campo, contestar las llamadas de emergencia que se generan de los ascensores, realizar la inspección de otros sistemas CCTV instalados en el centro comercial, entre otras. En el transcurso del día, el personal de vigilancia utiliza el sistema de guiado para buscar vehículos por medio de su placa y dirigir el tráfico vehicular de acuerdo con la información suministrada por los sensores de ocupación. En horas de la noche, el personal de vigilancia realiza la revisión de las novedades reportadas en el transcurso del día, por medio de la grabación de videos generada por el sistema de guiado.

Uno de los problemas observados en la operación del sistema de guiado es su poco uso por parte del personal de monitoreo del centro comercial. Una de las razones que justifica el bajo nivel de uso es la poca área de cobertura de las cámaras, que es ocupada en su mayor parte por el vehículo que ingresa al estacionamiento (ver Figura 2). El monitoreo de una sola cámara no suministra suficiente información para la identificación oportuna de una actividad sospechosa, y dado que la información del estado del parqueadero es suministrada por medio de las 1087 cámaras, la tarea de realizar la revisión de todos los videos resulta ser dispendiosa para el vigilante que hace uso del sistema.

Con el objetivo de identificar las principales actividades sospechosas que pueden ser identificadas en el CCTV, se realizaron entrevistas al coordinador de parqueaderos del centro comercial, y al gerente de la empresa Controles Inteligentes. A continuación, se enumeran las preguntas que se realizaron a cada uno de los entrevistados:

- ¿Cuáles son los robos y las novedades que se presentan comúnmente en el parqueadero, frente al robo y el hurto de vehículos?
- ¿Cuáles son las partes que más se roban dentro de un vehículo?

- ¿Cuál es la acción que debe tomar el personal de monitoreo y vigilancia al momento de presentarse una novedad?
- ¿Cuál es la característica particular de una persona que comete un acto delictivo dentro del parqueadero?

De las entrevistas desarrolladas, se concluye que las principales acciones delictivas que se presentan en la operación de un parqueadero son el robo de partes de los vehículos, el forcejeo de puertas y el rompimiento de vidrios para la extracción de elementos. Entre las partes que más son víctimas de hurto se encuentran las lunas y los espejos, las partes removibles como las antenas, y las llantas. Según el coordinador de parqueaderos del centro comercial, la velocidad con la cual la persona comete el hurto hace difícil su detección en el instante, por lo que las acciones frente a la novedad de robo ocurren una vez ya han sido efectuadas. Según la experiencia del coordinador de parqueaderos, el robo de una luna puede ser ejecutado en un tiempo menor a 10 segundos, mientras que la apertura de una puerta puede ser elaborada en un periodo entre 30 segundos y un minuto.

Adicionalmente, las personas que efectúan los robos son cuidadosas de analizar los patrones de movimiento del personal de vigilancia que se encuentra en campo. Esta condición ocasiona que las acciones sean ejecutadas en el momento en que el vigilante no se encuentre en la zona, lo que implica una mayor importancia de la labor desarrollada por el personal que realiza el monitoreo de las cámaras. Según el gerente de Controles Inteligentes, las acciones que pueden indicar que una persona está cerca de cometer un acto delictivo son estacionar su vehículo en lugares distintos del parqueadero durante un mismo día, permanecer mucho tiempo dentro del vehículo, estar cerca de un vehículo durante mucho tiempo y merodear.

Con el objetivo de identificar las acciones sospechosas que pueden ser ejecutadas en el sistema, a continuación, se definen las actividades objetivo que se desean detectar dentro del modelo. Para realizar la identificación de las actividades sospechosas, se debe definir con claridad cuáles son las actividades no sospechosas que el sistema realiza dentro de su clasificación. La definición de las actividades sospechosas se realizó con base en los resultados de las entrevistas, mientras que la definición de las actividades comunes se desarrolló con base en el análisis de videos capturados por el CCTV del sistema de guiado.

5. DISEÑO DEL SISTEMA ORIENTADO A AGENTES

Durante el análisis del estado del arte se evidenció que una de las principales falencias de los modelos de reconocimiento de actividades es la carencia de un diseño distribuido. El diseño de un modelo escalable es un requerimiento fundamental en la creación de sistemas de apoyo para los CCTV, dado el gran número de cámaras que pueden estar presentes dentro de una instalación. La programación orientada a agentes (AOP por sus siglas en inglés) surge como una alternativa de diseño distribuido, gracias a características como el control de recursos compartidos [5], el uso de operaciones concurrentes [6] y el cumplimiento de metas a partir del uso de estrategias cooperativas [7].

En la siguiente sección se realiza una descripción del paradigma AOP, y se justifica su uso en el diseño del modelo de detección de actividades. A continuación, se desarrolla el modelo de agentes y se definen las metas y roles del sistema, así como las estrategias cooperativas y protocolos de comunicación.

5.1. AOP en el reconocimiento de actividades inusuales

De forma similar a la Programación Orientada a Objetos (OOP), el objetivo de AOP es dividir las responsabilidades del programa en entidades independientes llamadas agentes. Sin embargo, la diferencia principal entre AOP y OOP radica en que los agentes son entidades capaces de funcionar de forma continua y autónoma, en coexistencia con otros procesos y agentes dentro de su entorno. La Programación Orientada a Agentes se encuentra inspirada en los modelos sociales, en las cuales existe una comunicación constante entre los miembros de una sociedad. La comunicación entre miembros lleva a la generación de estructuras organizacionales y modelos de cooperación orientados al cumplimiento de metas [67].

Para evaluar si la programación orientada a agentes es un paradigma apropiado para la detección de actividades inusuales, se identificaron algunos aspectos que deben ser abordados en la solución del problema de investigación:

- Control de recursos distribuidos: En la caracterización del caso de estudio se observó que el procesamiento y el análisis de imágenes es llevado a cabo sobre múltiples equipos de cómputo.
- Cumplimiento de metas: El desarrollo de una arquitectura por niveles define unas metas específicas, que se deben cumplir con el objetivo de garantizar el buen funcionamiento de los niveles superiores.
- Uso de estrategias cooperativas: El modelo debe definir protocolos de comunicación entre los distintos niveles del sistema, que permitan realizar una transmisión eficiente de la información y un buen uso del ancho de banda de la red.
- Uso de técnicas de inteligencia artificial: El desarrollo de sistemas inteligentes posibilita la encapsulación de las funcionalidades del sistema en agentes racionales, que permitan alcanzar de una forma óptima las metas del sistema.

La justificación del uso de AOP en el diseño del modelo de detección de actividades se encuentra en cada uno de los puntos abordados en los numerales anteriores, ya que pueden ser resueltos de una forma natural usando el paradigma de agentes. De forma adicional, algunos Frameworks de AOP como JADE encapsulan todas las operaciones de un agente dentro de un hilo único [68]. Esta característica permite que los agentes de requieren altos recursos de procesamiento puedan operar de forma paralela, optimizando el uso de los procesadores instalados en la máquina. Adicionalmente, el desarrollo de sistemas AOP permite tener libertad de ubicar los agentes en las máquinas del sistema dependiendo de sus capacidades y los criterios del programador, sin requerir modificaciones en el diseño.

5.2. Diseño del sistema orientado a Agentes

El diseño del sistema orientado a agentes se elaboró con base en la metodología AOPOA propuesta por González et al [24]. Para el desarrollo del modelo de agentes, se tomó como base el modelo de identificación de actividades propuesto por Cristiani et al, extendiendo su arquitectura en 3 niveles: nivel bajo, nivel medio y nivel alto.

La metodología AOPOA inicia a partir del levantamiento de requerimientos del sistema. Los requerimientos funcionales y no funcionales se identificaron a partir del análisis del estado del arte y de la caracterización del caso de estudio. Según Rodríguez et al, el análisis de requerimientos debe elaborarse luego de identificar las entidades externas con las cuales el sistema debe interactuar [69]. Rodríguez clasifica las entidades externas en 2 tipos: las entidades externas activas (actores), y las entidades externas pasivas (objetos del ambiente). En el caso de estudio se identificaron 5 entidades externas con los cuales el sistema de agentes debe ser capaz de interactuar: objetos, carros, peatones, vigilantes y administradores. Posterior a identificar las entidades externas, la tabulación de los requerimientos se desarrolla a partir de la identificación de 5 campos: número, nombre, descripción, prioridad y entidades externas involucradas.

Una vez identificados los requerimientos y las entidades externas, se elaboró la tabla de objetivos, habilidades y recursos, según la metodología propuesta en el trabajo de Rodríguez et al [69]. A continuación, los objetivos, las habilidades y los recursos identificados se relacionaron en una única tabla, denominada tabla de tareas. La tabla de tareas fue la base para realizar la descomposición organizacional, a través de un proceso iterativo que dividió los roles complejos en roles simples de acuerdo con las habilidades y recursos requeridos. El Anexo 1 muestra la definición de las tablas de objetivos, habilidades, recursos, tareas y roles elaborada para el sistema de detección de actividades inusuales.

El proceso de descomposición organizacional de la metodología AOPOA generó como resultado la definición de 8 roles en el sistema multiagente. Se realizó un análisis sobre cada uno de los roles para determinar si alguna de las habilidades asociadas requería el uso de algoritmos de inteligencia artificial, con base en las técnicas identificadas en el estado del arte. De acuerdo con el uso de inteligencia artificial, cada uno de los roles se clasificó en 2 grupos: roles inteligentes y normales. Esta clasificación es útil ya que permite identificar los roles que requieren mayores recursos computacionales y que cuentan con mayor complejidad. La Tabla 6 muestra la definición de los 8 roles, así como las metas identificadas para cada uno de ellos.

Tabla 6: Metas definidas para cada uno de los roles - metodología AOPOA

SMA - Detección de Actividades Inusuales		
Tabla de Metas - Actividades		
Código	**Nombre**	**Agente**
1.1.1.1	Capturar imágenes de las cámaras dentro del sistema	A1. Agente Captura
1.1.1.2	Identificar el esqueleto de la persona	A2. Agente Pose
1.1.1.3.1	Extracción de descriptores de la persona	A3. Agente Descriptores
1.1.1.3.2	Clasificación de pose de la persona	A3. Agente Descriptores
1.1.1.3.3	Aplicar algoritmos de reidentificación de personas	A4. Agente Reidentificación
1.1.1.3.4	Separación de la actividad en acciones	A5. Agente Organizador
1.1.1.4	Clasificación de acciones	A6. Agente clasificación
1.1.1.5	Clasificación de actividades	A6. Agente clasificación
1.1.1.6	Ejecución de técnicas de ensamble en la clasificación de la actividad	A7. Agente Ensamble
1.1.2	Identificar actividades como usuales o inusuales	A8. Agente Interfaz
1.2	Gestionar la Interfaz de usuario	A8. Agente Interfaz

5.3. Diseño Organizacional

De acuerdo con la definición de roles, objetivos, habilidades y recursos desarrollado en la sección anterior, se realizó el diseño organizacional del sistema multiagente que se muestra en la Figura 4. De acuerdo con el análisis propuesto por Rodríguez et al, el diseño organizacional debe partir con la identificación de recursos compartidos y el establecimiento de vínculos de cooperación. Se puede observar que el diseño organizacional conserva los niveles propuestos por el modelo de detección de actividades de Cristiani et al: nivel bajo, nivel medio y nivel alto. La Figura 4 muestra que la mayoría de los vínculos de cooperación presentes en el nivel bajo cuentan con un patrón productor-consumidor, permitiendo la implementación de técnicas en procesamiento en paralelo en modo pipeline [8]. Los agentes de un nivel inferior entregan a los agentes de un nivel superior la información recibida, adicionando información adicional. Debido a que existen cámaras que cuentan con capacidad de cómputo, el diseño implementa un grupo de agentes de nivel bajo por cada cámara instalada en el CCTV.

Aunque los agentes de nivel medio se encuentran dentro del grupo de agentes racionales, no requieren la misma capacidad de procesamiento que los agentes de nivel bajo. El diseño del modelo implementa un agente de nivel medio por cada zona del espacio de trabajo. La ubicación de agentes de nivel medio por zonas reduce el número de agentes que requiere el modelo sin afectar su diseño distribuido. Los agentes de nivel medio pertenecientes a cámaras adyacentes cuentan con vínculos de cooperación entre ellos, con el objetivo de poder rastrear una persona en su paso por múltiples zonas.

Los agentes de nivel alto del sistema se encargan de realizar la identificación de las actividades, con base en la secuencia de descriptores entregadas por los agentes del nivel medio. Con el objetivo de abarcar diferentes técnicas identificadas en el estado del arte, se implementarán agentes especializados de clasificación de actividades. Las respuestas provenientes de los agentes de clasificación se someterán a votación, y su respuesta definitiva será enviada a la interfaz en caso de que la actividad sea catalogada como inusual.

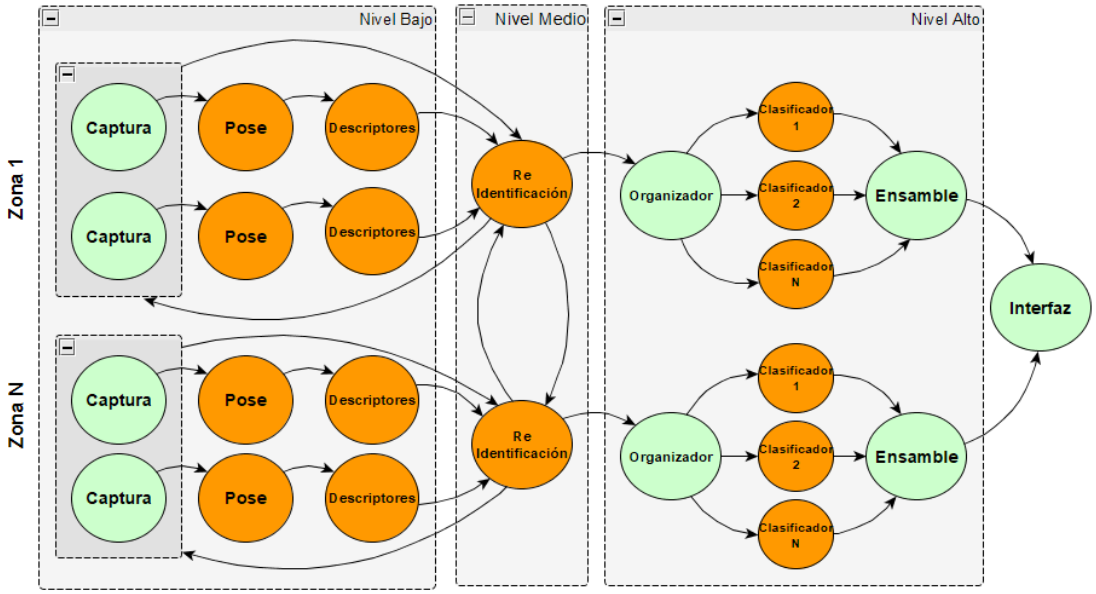

Figura 4: Modelo organizacional del sistema orientado a agentes.

5.4. Estrategias Cooperativas

El modelo organizacional mostrado en la Figura 4 ilustra dos casos en los cuales el sistema multiagente no cuenta con un patrón productor-consumidor. El primer caso corresponde a la sincronización de tiempos requerido por los agentes de captura del nivel bajo, y el segundo caso corresponde al rastreo de una persona cuando realiza un cambio de zona. En las siguientes secciones se describe la metodología y los protocolos de comunicación desarrollados para cada una de las estrategias cooperativas.

Sincronización de Tiempos

Con el objetivo de que garantizar el correcto funcionamiento de los algoritmos de reidentificación, se debe garantizar que la hora utilizada por los de captura para colocar la marca de tiempo en la imagen recibida se encuentre sincronizada. Uno de los protocolos más utilizados para realizar la sincronización de tiempos en sistemas distribuidos es el protocolo de tiempo de red (NTP por sus siglas en inglés). NTP es un protocolo de sincronización de relojes dise-

ñado por la fundación del tiempo de red (Network Time Foundation), que proporciona un estándar para la sincronización de relojes que opera en más de 10 millones de dispositivos a lo largo del mundo [70]. NTP proporciona una arquitectura jerárquica por niveles (stratum), en donde los dispositivos más altos sincronizan los relojes de los dispositivos más bajos.

Dado que uno de los requerimientos del sistema establece que el diseño debe estar en la capacidad de operar sin internet, los agentes deben establecer una estructura jerárquica que permita realizar la sincronización de tiempos por medio del protocolo NTP. Esta estructura jerárquica debe ser autoorganizada y garantizar que la sincronización continúe funcionando en caso de que alguno de los agentes se encuentre fuera de línea. Tomando como base el diseño organizacional por zonas definido en la Figura 4, se ubica a los agentes reidentificación en la parte superior de la jerarquía, los cuales se encuentran encargados de sincronizar los relojes de los agentes de captura asociados a la zona. En caso de que uno de los agentes de reidentificación se encuentre fuera de línea, la sincronización de tiempos solo afectará a los agentes asociados a la zona, y no al sistema completo. El procedimiento de sincronización de relojes por zonas se ilustra en la Figura 5.

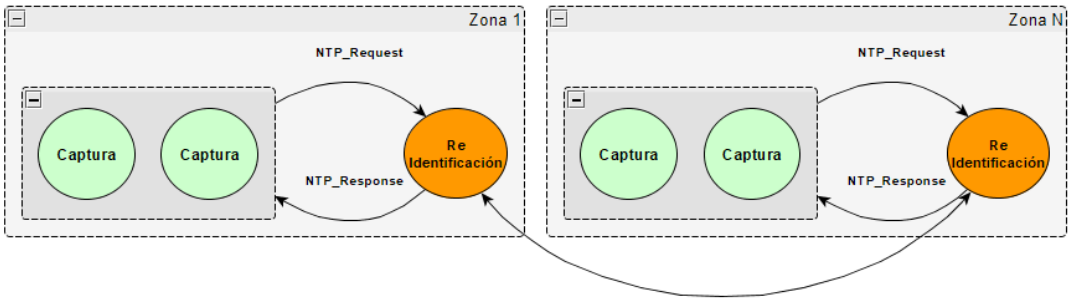

Figura 5: Protocolo de sincronización de relojes en zonas, con base en el protocolo NTP.

Por otra parte, el grupo de agentes de reidentificación ejecutarán la sincronización de relojes a través del algoritmo de Berkeley [71]. El uso del algoritmo de Berkeley como mecanismo de sincronización elimina la dependencia de un único agente para la sincronización de tiempos. A través de la designación de un coordinador, el algoritmo de Berkeley obtiene el tiempo de todos los relojes del grupo y calcula el tiempo promedio entre ellos. A continuación, el tiempo promedio calibra los relojes del sistema, tomando como base el tiempo promedio calculado en el paso anterior. La Figura 6 muestra el protocolo de comunicación utilizado para la sincronización de tiempos, basado en el algoritmo de Berkeley.

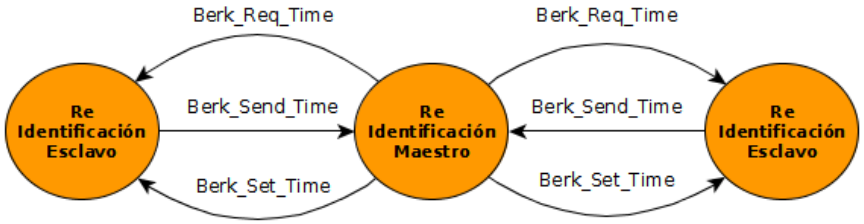

Figura 6: Protocolo de sincronización de relojes entre agentes de reidentificación, basado en el algoritmo de Berkeley.

La selección del agente coordinador que efectuará la sincronización de relojes se efectuará a través del algoritmo Bully [72]. El algoritmo Bully realiza la selección del coordinador a través de un parámetro llamado id de proceso. El agente que tenga el id del proceso más alto será designado como coordinador del sistema. Cada vez que el sistema detecte que el agente coordinador esté fuera de línea, los agentes iniciarán un proceso de envío y recepción de mensajes que culmina con la selección del nuevo agente coordinador. La Figura 7 muestra el protocolo de comunicación efectuado entre los agentes de reidentificación, para la selección del coordinador a través del algoritmo Bully.

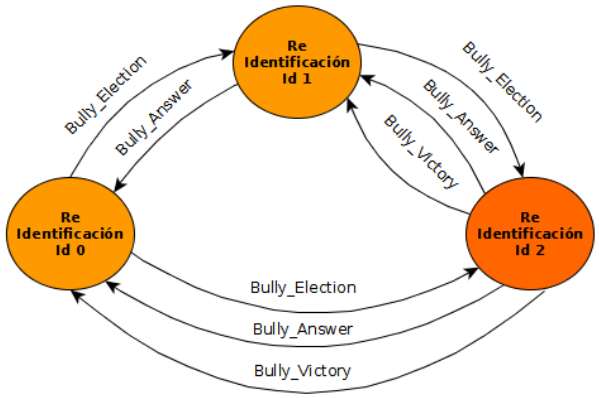

Figura 7: Protocolo de elección del agente coordinador basado en el algoritmo Bully.

Reidentificación en cambios de zona

Con el objetivo de garantizar un diseño distribuido, el modelo realiza una distribución de los agentes de nivel medio a partir de zonas. Si bien la distribución de agentes por zonas permite obtener escalabilidad en el modelo, el sistema debe estar en la capacidad de detectar el paso de una persona a través de múltiples zonas. El desarrollo de esta característica permite realizar el rastreo de actividades que abarcan un gran espacio y cuentan con una gran duración.

Cada vez que una nueva persona es detectada en una zona, el agente reidentificación envía un mensaje a los agentes vecinos. El mensaje contiene la marca de tiempo de la persona identificada, los descriptores de color y la posición global. Los agentes vecinos reciben el mensaje y

comparan los descriptores recibidos con cada una de las personas que han sido detectadas dentro de la zona en los últimos segundos. El agente vecino evalúa entre las dos personas la diferencia en color, distancia y marcas de tiempo. Si el agente determina que las 2 personas son iguales, envía una respuesta afirmativa al agente emisor que contiene la lista de poses almacenadas y una medida de similitud que se genera como resultado de la comparación. Si no hay coincidencia, el agente vecino envía una respuesta negativa al agente vecino.

Si el agente emisor no recibe respuesta afirmativa de alguno de los nuevos vecinos, el sistema genera una nueva lista que almacenará las poses de la persona y sus descriptores. En caso contrario, el agente emisor toma como base la lista de poses enviada por el agente vecino, enviando un mensaje de confirmación para informar al agente vecino que la persona se encuentra en una nueva zona. Si el agente emisor recibe más de una confirmación de un agente vecino, selecciona como base la persona que contenga la mayor medida de similitud. El procedimiento de reidentificación en cambios de zona se encuentra ilustrado en la Figura 8.

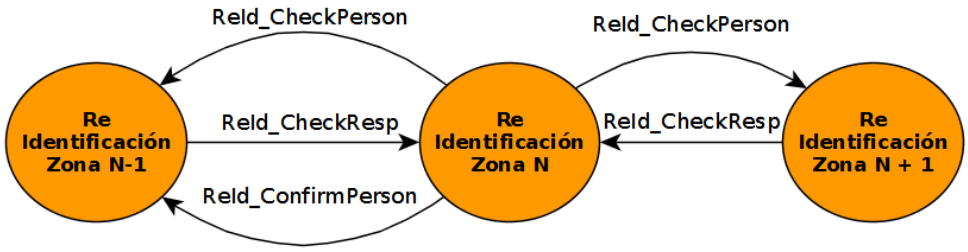

Figura 8: Protocolo de identificación de una persona en su paso por múltiples zonas.

En este capítulo se describió el modelo de agentes de AC-CCTV, desarrollado a partir de la metodología de diseño AOPOA. El modelo desarrollado cuenta con 2 características principales. La primera es el patrón productor-consumidor desarrollado para la mayoría de los agentes, que permite aplicar técnicas de procesamiento en paralelo por medio de modelos pipeline [8]. La segunda característica es su modelo descentralizado a partir de su división por zonas, que permite aumentar el modelo a partir del control de su escalabilidad. El siguiente capítulo desarrolla el modelo de inteligencia de cada uno de los roles descritos en la Tabla 6, y realiza la selección de las técnicas de inteligencia identificadas en el estado del arte.

6. DISEÑO DEL MODELO DE INTELIGENCIA

En el capítulo anterior se realizó el diseño del sistema orientado a agentes utilizando la metodología AOPOA, generando como resultado la definición de 8 roles y la creación de un modelo organizacional que permite obtener un diseño distribuido. Una vez elaborado el modelo organizacional, se definieron los protocolos de comunicación entre los agentes y se desarrollaron estrategias cooperativas, orientados a solucionar problemas como la sincronización de relojes y el rastreo de personas al presentarse un cambio de zona.

Cada uno de los 8 roles identificados en el modelo de agentes fue clasificado como racional o no racional, de acuerdo con su necesidad de implementar técnicas de inteligencia artificial para el cumplimiento de sus metas. En las siguientes secciones se realizará el desarrollo del modelo de inteligencia para cada uno de los 8 roles definido en el modelo de agentes. En el caso de los agentes normales, se definirán los algoritmos utilizados para establecer los mecanismos de cooperación y procesar la información recibida de los demás agentes. Por otro lado, el modelo de inteligencia de los agentes inteligentes se desarrollará con base en las técnicas de reconocimiento de actividades inusuales, identificadas en el estado del arte.

6.1. Agente de Captura

La tabla de habilidades definida en el capítulo anterior determinó que la meta principal del agente de video es realizar la captura del video. La caracterización del caso de estudio mostró que las cámaras instaladas en el CCTV soportan comunicación TCP/IP, en donde la transmisión de videos se ejecuta bajo protocolo MJPEG. La característica principal del protocolo es que la transmisión de video se realiza a través de una secuencia de imágenes. Cada imagen del protocolo es codificada bajo el método de compresión JPEG [73].

La mayoría de los lenguajes de programación cuentan con librerías que permiten obtener la secuencia de imágenes emitida por el protocolo MJPEG. Para el caso de Python existen librerías como PY-MJPEG, que proporciona interfaces para el manejo de cámaras usando el protocolo MJPEG [74]. Por otra parte, el agente de captura debe asignar la marca de tiempo de la imagen capturada, con base en la hora real determinada por el sistema operativo. La asignación de una marca de tiempo al momento de la captura es fundamental para la correcta operación del algoritmo de reidentificación de personas. Además de poder realizar la lectura del reloj del sistema operativo, el agente debe estar en la capacidad de ajustar la hora del sistema. Esta característica es necesaria para implementar la sincronización de relojes, descrita en el desarrollo del modelo de agentes.

6.2. Agente Pose

Una vez el agente de captura realiza la extracción de la imagen asociada a la cámara, la detección de las personas en la imagen se efectúa a partir de los modelos de inteligencia implementados en el agente pose. De acuerdo con el análisis efectuado en el estado del arte, la detección de personas sobre una secuencia de video puede implementarse a partir de 2 méto-

dos: segmentación a partir de técnicas de extracción de fondo, y segmentación a partir de técnicas de análisis sobre imagen única. Teniendo en cuenta que el objetivo del trabajo de investigación no está orientado al desarrollo de técnicas de nivel bajo, el diseño del modelo de inteligencia se basa en el uso de herramientas preexistentes que implementan los algoritmos contenidos en estas 2 categorías.

La evaluación de la identificación de personas a través de las técnicas de extracción de fondo se elaboró a partir de la librería de procesamiento de imágenes OpenCV. La librería OpenCV cuenta con 5 métodos que permiten la implementación de técnicas de extracción de fondo. Las técnicas de extracción de fondo disponibles incluyen algoritmos simples como una resta simple de imágenes (CNT), hasta el uso de modelos adaptativos desarrollados a partir de modelos de mezclas gaussianas (GMM). Por otra parte, la segmentación de personas a partir de técnicas de análisis sobre imagen única se desarrolló a partir de la librería OpenPose. OpenPose permite obtener los esqueletos de cada una de las personas, posibilitando el uso de algoritmos de identificación de actividades a partir de la posición de las articulaciones [10]. Dado que las cámaras instaladas en el sistema de guiado del centro comercial no son estereoscópicas, las coordenadas de cada una de las articulaciones se obtienen en un espacio de dos dimensiones.

Para realizar la evaluación de las técnicas de detección de personas, se tomó una muestra de videos de las cámaras instaladas en el centro comercial. Los 5 algoritmos de extracción de fondo disponibles en OpenCV se pusieron a prueba a partir de un programa desarrollado en Python que procesaba los videos capturados. Los resultados fueron analizados de forma cualitativa a partir de la definición de 5 criterios que incluyen la calidad de las siluetas detectadas, el ruido generado, el control de sombras y el comportamiento frente a cambios de iluminación.

Uno de los problemas recurrentes al momento de aplicar algoritmos de extracción de fondo sobre la secuencia de video es la gran perturbación que producen el paso de los vehículos dentro de la escena. La Figura 9A muestra el resultado de aplicar el algoritmo de extracción de fondo KNN en el momento en que un vehículo se encuentra en proceso de parqueo. De la Figura 9A se puede observar un alto nivel de ruido sobre la imagen, imposibilitando la detección de personas a partir de la extracción de siluetas. Adicionalmente, las personas que usaban tonos grises en sus prendas de vestir generan siluetas parciales en los algoritmos de extracción de fondo, dado que el color de las prendas se mezclaba con el color del asfalto.

De igual forma, la evaluación del desempeño de la librería OpenPose se ejecutó a través del desarrollo de un programa en Python que analizaba los videos capturados. A diferencia de los algoritmos de extracción de fondo, los esqueletos detectados por la librería OpenPose mostraron buenos resultados frente al paso de vehículos dentro de la escena y cambios de iluminación. Dado que el algoritmo realiza el proceso de detección sobre cada imagen, los tiempos de procesamiento requeridos por OpenPose son más altos con respecto a los algoritmos de extracción de fondo implementados en OpenCV. Hidalgo muestra una máquina con una tarjeta gráfica NVIDIA 940MX y un procesador Intel Core I7 puede tardar hasta 0.5 segundos en procesar una única imagen [75]. Una máquina con las mismas configuraciones tarda hasta 50 milisegundos en realizar el procesamiento de una sola imagen, utilizando técnicas de extrac-

ción de fondo. La Figura 9B muestra un ejemplo de la detección efectuada por la librería OpenPose sobre una de las imágenes seleccionadas del caso de estudio.

Figura 9: A: Imagen resultante de aplicar el algoritmo de extracción de fondo KNN en un vehículo en proceso de parqueo. B: Resultados de aplicar la librería OpenPose sobre una de las imágenes pertenecientes a las secuencias de video capturadas.

Si bien la velocidad de respuesta es un requerimiento no funcional en el desarrollo del sistema, el alcance del proyecto de investigación está limitado al desarrollo del módulo de alto nivel a partir de los datos suministrados por el módulo de bajo nivel. Por lo tanto, los tiempos de respuesta no se incluyen como criterio de selección de la herramienta. Para lograr una implementación sobre un sistema distribuido real, se propone como trabajo futuro optimizar los tiempos de respuesta ofrecidos por la librería OpenPose. La Tabla 7 muestra los resultados de aplicar la evaluación cualitativa a cada una de las herramientas. Con base en los resultados obtenidos, se selecciona a OpenPose como la librería que mejor realiza la detección de personas dentro de la secuencia de videos del CCTV.

Tabla 7: Evaluación de criterios - Técnicas de Nivel Bajo

	MOG	MOG2	KNN	GMG	CNT	POSE
Siluetas (3)	2	3	3	3	3	5
Ruido Generado (2)	4	2	2	1	2	4
Tolerancia - Cambios de Iluminación (1)	4	2	2	1	2	5
Manejo de Sombras (2)	3	2	2	2	1	5
TOTAL (PONDERADO)	**24**	**19**	**19**	**16**	**17**	**38**

6.3. Agente Descriptores

Los esqueletos generados por la librería OpenPose cuentan con un total de 25 puntos, en donde cada punto describe una coordenada cartesiana de dos dimensiones sobre la imagen. Los primeros 15 puntos corresponden a posiciones de las articulaciones (codos, rodillas, cadera, entre otros). Los 10 puntos restantes brindan información acerca de la posición del rostro, las orejas, los dedos de las manos, entre otros. La Figura 9B muestra que las cámaras instaladas en el CCTV favorecen la captura en cuerpo completo de la persona y no obtienen mayor detalle del rostro o las manos, se omitieron los últimos 10 puntos del esqueleto ya que son pro-

pensos a generar ruido en el sistema. La Figura 11A muestra la posición de cada uno de los 15 puntos identificados por la librería OpenPose.

En el capítulo de caracterización del caso de estudio se mencionó que una de las propiedades más relevantes del CCTV es la poca área de cobertura de las cámaras. La baja área de cobertura favorece que la silueta de la persona no aparezca completamente en la imagen, generando la aparición de esqueletos parciales. Dado que los esqueletos parciales no aportan mucha información en la identificación de una actividad, el agente descriptor cuenta con un filtro que rechaza las poses parciales que no cumplen algunas características. Para que una pose no sea rechazada por el filtro, debe contener al menos el torso y alguna de las dos piernas, o la parte correspondiente al fémur de ambas piernas. Además de eliminar las poses que no transmiten información, el filtro rechaza las poses parciales que generan ruido en el sistema. La Figura 10 muestra un ejemplo del filtro de poses parciales implementado en el modelo.

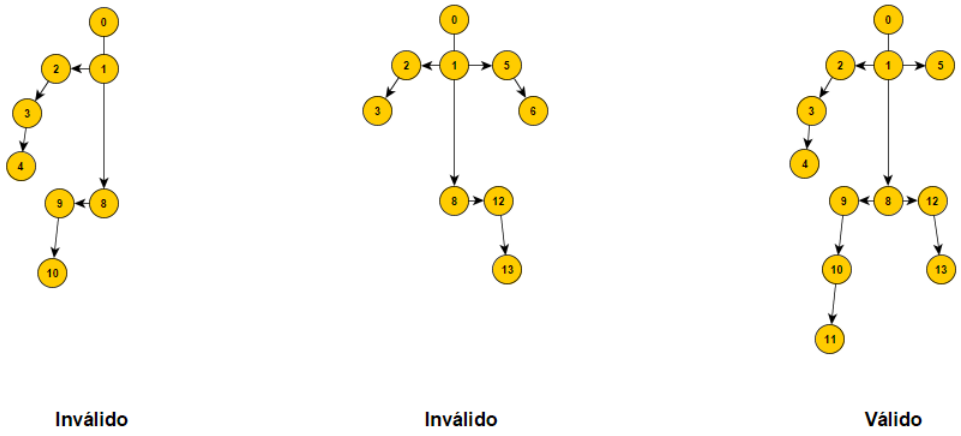

Figura 10: Ejemplo del filtro de poses parciales en el sistema. Las 2 primeras poses se marcan como inválidas, dado que no contiene los 2 fémur o alguna de las piernas completas. La tercera pose se marca como válida, ya que contiene de forma completa una de las 2 piernas.

Una vez las poses parciales válidas son identificadas, el agente descriptor calcula la posición local de cada una de las poses sobre la imagen. Dado que la posición local determina una coordenada de ubicación de la persona sobre el piso de la escena, la primera aproximación es utilizar la información de los pies para determinar la posición local. Aunque esta aproximación es utilizada en trabajos como el de Eshel et al [54] y Khan et al [55], la poca área de cobertura de las cámaras y la oclusión generada por los vehículos genera que la posición de los pies no sea una medida confiable para el cálculo de la posición. El modelo propuesto en este proyecto de investigación propone trazar una línea vertical desde la cadera de la persona al piso, similar al efecto que produce la gravedad sobre una plomada de construcción. La longitud de la línea de la plomada se define en función de alguna medida del esqueleto, que para este caso corresponde a 2 veces el fémur de la persona. La Figura 11B muestra el funcionamiento del efecto plomada aplicado al cálculo de la posición local del esqueleto sobre una imagen.

Luego de calcular la posición local de cada uno de los esqueletos sobre la imagen, el agente descriptor calcula la posición global a través de técnicas de transformación proyectiva. De acuerdo con el análisis desarrollado en el estado del arte, la homografía requiere ejecutar una calibración de 4 puntos en cada una de las cámaras instaladas en el CCTV. Para realizar la calibración de las cámaras, cada uno de los 4 puntos de calibración ubicados en la imagen debe estar relacionado con su posición global. Por lo tanto, la calibración requiere realizar un plano de la escena, en donde se ubique un sistema de coordenadas y se determine, con la mayor precisión posible, la posición global de los puntos de calibración con respecto al origen.

Luego de calcular la posición global del esqueleto, el agente descriptor extrae características de color de la persona que permitan la ejecución de la etapa de reidentificación. El método propuesto en el modelo está inspirado en el trabajo de Jang et al, los cuales realizan una comparación de regiones de color, dividiendo la silueta de la persona en sus partes superior e inferior [50]. Sin embargo, una de las variaciones del modelo propuesto con respecto al trabajo de Jang et al es eliminar el uso del espacio de color HSV, dado que la componente H cuenta con fluctuaciones al realizar el análisis de colores en escala de grises [76]. Adicionalmente, el modelo propuesto utiliza el concepto de colores dominantes, en donde la extracción de colores se ejecuta utilizando el método implementado por Erconalelli [77]. El modelo de Erconalelli utiliza el algoritmo de aprendizaje no supervisado K-Means para realizar la extracción, en donde cada punto de la región de interés es procesado en el espacio de color RGB.

La región utilizada para realizar la extracción de colores corresponde a puntos cercanos al torso y los brazos, dado que estos proporcionan con mayor fiabilidad información acerca del color de ropa de la persona. El modelo evita el uso de descriptores de color correspondientes al pantalón, dada la alta incertidumbre en la posición las piernas generada por la librería OpenPose. Para la implementación del método K-Means se seleccionó un valor de K = 3, permitiendo la detección de un máximo de 3 colores preponderantes por esqueleto. La Figura 11C muestra en color rojo un ejemplo de la región de interés seleccionada en la extracción de colores dominantes de la persona.

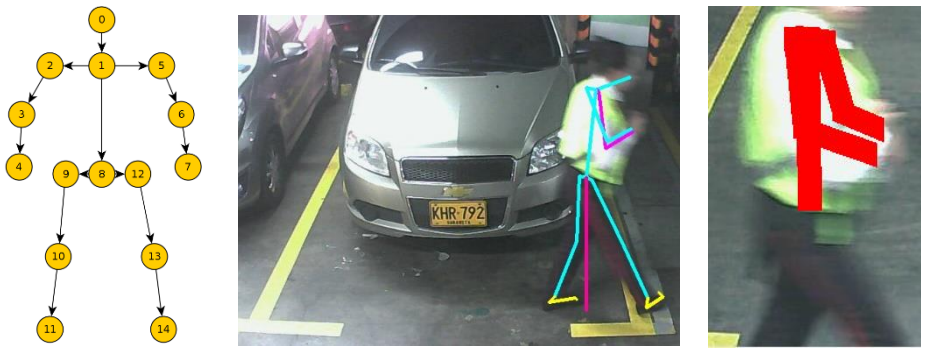

Figura 11: A: Ubicación de la posición de las articulaciones identificadas por la librería OpenPose. B: Ejemplo del cálculo del efecto plomada utilizado para el cálculo de la posición del esqueleto. C:

Ejemplo de la región utilizada por el modelo para la extracción de descriptores de color de la persona.

Posterior a realizar el cálculo de los colores dominantes, el agente procede a realizar el cálculo de los descriptores asociados al esqueleto de la persona. El primer grupo de descriptores usa la posición de cada uno de los puntos de las articulaciones, trasladando el origen del sistema de coordenadas a la posición que ubica el cuello del esqueleto (punto 1, Figura 11A). A continuación, el conjunto de puntos se normaliza con base en alguna medida de proporcionalidad del esqueleto que, al igual que el efecto plomada aplicado en el cálculo de la posición, corresponde a la medida del fémur. Si ignoramos el descriptor de la cabeza y omitimos la posición del cuello dado que se encuentra en el origen, la normalización genera un total de 13 puntos que dan lugar a 26 descriptores.

Inspirado en el trabajo desarrollado por Gedat et al, el segundo grupo de descriptores asociados al esqueleto está relacionado con el ángulo que forman las articulaciones al efectuar la pose. Con base en la Figura 11B, se realiza la medición de los ángulos formados por las siguientes articulaciones: brazo derecho (4-3-2), hombro derecho (3-2-1), codo izquierdo (7-6-5), codo derecho (5-4-1), rodilla derecha (11-10-9), cadera derecha (10-9-8), rodilla izquierda (14-13-12) y cadera izquierda (13-12-8). De forma adicional, las líneas que conforman los húmeros y fémures se extienden hasta intersectarse con el torso (1-8), realizando el cálculo que forma la intersección. En total, el vector de descriptores contiene un total de 12 ángulos.

Para eliminar la variabilidad en el número de descriptores ocasionado por las poses parciales, el agente descriptor intenta completar las coordenadas de los puntos faltantes, con base en la información proporcionada por las partes del esqueleto que fueron detectadas. Por ejemplo, si el agente detecta que los puntos correspondientes a los pies no se encuentran en el esqueleto, utiliza la distancia del fémur para proyectar su posición. De igual forma, el sistema utiliza la medida del húmero para proyectar la posición de las manos, en caso de que estas no estén presentes en el esqueleto. Si los puntos correspondientes a los brazos no son detectados (posición 3, 4, 6, 7, Figura 11A), el sistema ignora la pose parcial detectada. Este criterio define una nueva característica para la implementación del filtro de poses parciales, basado en la completitud de los puntos de los brazos. Al implementar un filtro de poses parciales en dos etapas, algunos esqueletos que no cumplen los criterios de la segunda etapa cuentan con descriptores de posición y colores dominantes, siempre y cuando cumplan con los criterios definidos en la primera etapa del filtro.

Con base en los descriptores calculados en la etapa anterior, el agente procede a realizar la identificación de la pose de la persona. Aunque el estado del arte determinó que la identificación de poses es una responsabilidad de alto nivel, los parámetros calculados por el agente descriptores permiten realizar la identificación de la pose en este nivel. Inspirado en los trabajos de Du et al [31] y Gedat et al [32], la identificación de poses se ejecuta alimentando modelos de clasificación basados en redes neuronales y SVM con los descriptores de ángulos y posición desarrollados en la etapa anterior.

6.4. Agente Re-Identificación

Una vez que se calculan los descriptores y se categorizan las poses de cada uno de los esqueletos, el agente reidentificación realiza el seguimiento del paso de la persona por múltiples cámaras y organiza los descriptores identificados en secuencia. El modelo de inteligencia del agente se desarrolla en 2 etapas: identificación de personas en cámaras sobrelapadas, y reidentificación de personas a partir de análisis de color y posición.

La primera etapa del modelo de inteligencia se encarga identificar las personas que se encuentran en regiones con sobrelapamiento. A través de la posición global calculada por el agente descriptores, se puede identificar si una persona se encuentra en una región con sobrelapamiento si la distancia entre los esqueletos es muy cercana. Para el caso en que dos o más personas se encuentren en la escena, esta heurística se puede complementar con una métrica de diferencia de color, tomando como base los colores dominantes identificados por el agente descriptores. Para realizar la comparación de color, se utiliza la fórmula CIEDE2000 desarrollado por la Comisión Internacional de la Iluminación (CIE por sus siglas en francés). Una de las ventajas de utilizar la fórmula CIEDE2000 es que proporciona una escala que asemeja la diferencia de color percibida por el ojo humano [78]. Dado que el descriptor de color cuenta con la información de 3 colores dominantes, el cálculo de la diferencia de color entre dos imágenes se ejecuta a través del Algoritmo 1.

El Algoritmo 1 permite tener robustez frente a la posición de la persona con respecto a la cámara, dado que los colores de la ropa permanecerán, en mayor o menor medida, entre los 3 colores dominantes identificados en la generación de los descriptores. Una vez aplicada la comparación de color entre los esqueletos, se aplica la condición generada por la Ecuación 1 para determinar si 2 esqueletos corresponden a la misma persona:

$$si\ Dis < DTH\ y\ K < KTH: 1, sino: 0 \qquad \textbf{Ecuación 1}$$

Algoritmo 1: Calcular Diferencia de Color

```
1  function DiferenciaColor (a: arreglo, b: arreglo):
2    da = Dominante(a);
3    min_diff = 1
4    for i = 0; i < 3; i++:
5      diferencia = CIEDE2000(da, b[i])
6      if diferencia < min_diff:
7        min_diff = diferencia
8  return min_diff
```

En donde Dis corresponde a la distancia entre la posición global de los esqueletos ($Dis \geq 0$), y K corresponde a la diferencia de colores obtenida por medio del Algoritmo 1. Los umbrales de cada uno de los criterios se seleccionaron por medio de la comparación de poses se seleccionaron de acuerdo con pruebas realizadas sobre videos capturados del centro comercial, y deben ser ajustados de acuerdo con el caso de estudio seleccionado.

Una vez la etapa de identificación en cámaras sobrelapadas se encuentra efectuada, el agente procede a realizar la etapa de reidentificación. De forma similar, la etapa de reidentificación utiliza los descriptores de color y posición para determinar si una persona presente en el instante de tiempo $t-1$ es igual a una persona que se encuentra en el instante de tiempo t. Con el objetivo de generar un indicador que permita validar el nivel de certeza de que dos esqueletos tomados en diferentes instantes de tiempo sean similares, el agente reidentificación utiliza el valor proporcionado por la Ecuación 2.

$$D = \alpha * DScore + (1 - \alpha) * CScore \quad \textbf{Ecuación 2}$$

En donde D corresponde a la certeza de similitud, que varía en un rango entre $[0, 1]$. El parámetro α es un factor de proporcionalidad que brinda mayor importancia a la comparación de color o posición al momento de realizar la reidentificación. Los valores de $DScore$ y $CScore$ corresponden a la similitud de color y posición calculado para los esqueletos, y se calculan por medio de la Ecuación 3 y la Ecuación 4.

$$CScore = \frac{CIEDE2000\big(C(t), C(t-1)\big)}{100} \quad \textbf{Ecuación 3}$$

$$DScore = 1 - \frac{Dis\big(P(t), P(t-1)\big)}{UmbralPos} \quad \textbf{Ecuación 4}$$

La función $CIEDE2000$ contenida en la Ecuación 3 es la fórmula de comparación de color aplicado a los descriptores de color de los esqueletos $C(t),\ C(t-1)$. Dado que la fórmula de comparación de color $CIEDE2000$ genera un valor que se encuentra en el rango $[0, 100]$, la Ecuación 3 realiza una normalización para que el valor de $CScore$ se encuentre entre el rango $[0, 1]$. Por otra parte, la ecuación para el cálculo de $DScore$ utiliza la distancia euclidiana Dis entre las posiciones de los esqueletos $P(t),\ P(t-1)$. La distancia euclidiana entre los puntos es normalizada por medio del parámetro $UmbralPos$, que define la distancia máxima de separación entre los esqueletos. Para evitar la aparición de valores negativos, el cálculo del parámetro $DScore$ se complementa por medio de la Ecuación 5.

$$Si\ DScore < 0 : DScore = 0 \quad \textbf{Ecuación 5}$$

La Ecuación 5 limita los valores que puede tomar el parámetro $DScore$, en caso de que la distancia euclidiana entre los 2 puntos sea mayor al valor de $UmbralPos$. Una vez se haya calculado la certeza de similitud D, el agente de reidentificación define si dos esqueletos, tomados en distintos instantes de tiempo, pertenecen a la misma persona por medio de la Ecuación 6:

$$S = Si\ D > DUmbral: 1, Sino: 0 \qquad \textbf{Ecuación 6}$$

Por cada persona que ejecuta una acción, el agente reidentificación genera una lista de descriptores. Si la Ecuación 6 determina que el nuevo esqueleto pertenece a uno de los esqueletos almacenados, el agente adiciona sus descriptores al arreglo. En caso contrario, el agente genera una nueva lista con los descriptores del esqueleto, indicando que una nueva persona ha aparecido en escena. Si un esqueleto no es adicionado a una lista durante un tiempo determinado, el agente asume que la persona ha desaparecido de escena y envía la lista de descriptores al agente organizador. El agente captura una muestra de los descriptores de la persona a una tasa definida por un intervalo de tiempo Tp.

Dado que existen puntos ciegos que ocasionan la desaparición de la persona de la escena, el agente descriptor realiza una interpolación con base en la información que se tiene certeza. Para interpolar el descriptor de posición, se asume que la persona se desplazó en línea recta a una velocidad constante durante el tiempo que desapareció de escena. Por otra parte, la información correspondiente a la pose del esqueleto (posición de las articulaciones, ángulos) se rellena extendiendo la información de los descriptores sobre los cuales el agente tiene certeza. Por ejemplo, si la persona sobre la cual se realiza el análisis desaparece durante 4 intervalos $(t0, t1, t2, t3)$, los descriptores de los intervalos $t0, t1$serán rellenados con los descriptores de la persona en el intervalo $t - 1$. De igual forma, los descriptores de la persona en t2, t3serán equivalentes al descriptor de la persona en el intervalo t4. De esta forma, la información temporal de la actividad se mantiene, dado que cada muestra de descriptores se captura en tiempos regulares.

6.5. Agente Organizador

Una vez los descriptores de la actividad ejecutada por una persona son organizados en secuencia, son enviados al agente organizador. De acuerdo con la asignación de habilidades definida en el diseño del modelo de agentes, el agente organizador debe contar con la habilidad de dividir una lista de descriptores en una secuencia de acciones. Aunque Saad et al introduce el concepto de acción dentro del modelo de identificación, no existe una definición clara de los límites que existen entre una acción y una actividad. En general, los trabajos publicados en el estado del arte no realizan una distinción clara entre los conceptos acción y actividad. Por ejemplo, Cippitelli et al proponen un modelo de clasificación orientado al reconocimiento de actividades, pero en la validación del modelo utiliza un set de datos que contiene un conjunto de acciones [79].

El modelo desarrollado en este trabajo de investigación propone delimitar las acciones ejecutadas por una persona en el desarrollo de una actividad, a partir de su desplazamiento. Gracias

a que el agente organizador cuenta con la posición global de los esqueletos, se puede identificar las franjas de tiempo en que la persona permanece estática en una misma posición. Al clasificar las acciones de una persona de acuerdo con su movimiento, se pueden implementar por separado clasificadores que identifican acciones que incluyen desplazamiento en la posición, y clasificadores en donde la persona permanece estática. Dado que la mayoría de las acciones asociadas al desplazamiento están relacionadas con correr o caminar, el clasificador de acciones de desplazamiento puede ser implementado con algoritmos como árboles de decisión [57]. Por otra parte, la clasificación de acciones estáticas debe realizarse con los métodos evaluados en el estado del arte, los cuales incluyen redes neuronales y cadenas ocultas de Markov [31] [32].

Para realizar la separación de la secuencia de descriptores en acciones, se recorre la lista de descriptores y se analiza la posición global del esqueleto actual con respecto a las siguientes posiciones de la lista. Si se detecta que uno de los esqueletos se encuentra en una posición mayor a un umbral determinado, el agente detecta que la persona comenzó a moverse y envía la lista guardada hasta el momento al clasificador de acciones estáticas. De igual forma, si la persona se encuentra en movimiento y el agente detecta que la persona se mantiene por debajo de un umbral determinado en los siguientes instantes de tiempo, el agente envía la lista de descriptores recopilada en el momento al clasificador de acciones de desplazamiento. Esta estrategia desarrolla el modelo propuesto por Saad et al, en donde la identificación de la actividad se realiza a partir de la identificación de sus acciones.

Si bien la segmentación de una actividad con base en el movimiento permite realizar la segmentación de una actividad, la información proporcionada por las acciones con movimiento dificulta la detección de actividades sospechosas como merodear. Por lo general, los trabajos relacionados en el estado del arte realizan la detección de la actividad merodear, a partir de conteo de cambios de la trayectoria desarrollada por una persona [80]. Inspirado en el método de Koo et al, el modelo segmenta las acciones con movimiento a partir de los cambios de trayectoria detectados durante una actividad. Esta estrategia permite generar descriptores que identifican las veces que una persona realizó cambios en trayectoria, aumentando la precisión del clasificador de actividades en la detección de actividades como merodear.

6.6. Agente Clasificador

Una vez la secuencia de descriptores es definida por el agente organizador, el agente clasificador se encarga de identificar las acciones, y posteriormente, las actividades ejecutadas por una persona. Con base en la división de descriptores efectuada por el agente organizador, el agente clasificador debe contar con 2 modelos de identificación de acciones: clasificador de acciones estáticas y clasificador de acciones de desplazamiento.

El modelo del clasificador de acciones estáticas se encuentra inspirado en el trabajo de Du et al, en donde las acciones son detectadas a partir de un sistema de detección basado en redes neuronales convolucionales (CNN por sus siglas en inglés). De acuerdo con el análisis desarrollado en el estado del arte, los modelos de clasificación basados en CNN han cobrado gran relevancia en el desarrollo de clasificadores, gracias a los altos porcentajes de precisión re-

portados por diversos autores en la implementación de modelos de reconocimiento de imágenes [59]. A diferencia de las redes neuronales convencionales, los descriptores de entrada de las CNN deben estar organizados en un arreglo de 1 dimensión, de una forma similar a la organización de los píxeles sobre una imagen en escala de grises. En su trabajo, Du et al implementó un algoritmo que genera una imagen a color de 64x64 píxeles, a partir de la posición de las articulaciones del esqueleto de la persona. Las columnas de la matriz contienen la información espacial de la pose, mientras que las filas de la matriz codifican la información temporal de los descriptores.

Una de las variantes propuestas en el modelo con respecto al trabajo de Du et al es aplicar la redimensión de las imágenes únicamente sobre el eje horizontal (temporal) más no sobre el eje vertical (espacial) en un espacio de 28x28 pixeles. La redimensión sobre un único eje permite reducir el daño de la estructura interna de los pixeles, ocasionado los algoritmos de redimensión de imágenes [81]. Las zonas de la imagen que se encuentran por fuera del rango de descriptores se rellenan de un valor nulo. La Figura 12 muestra un ejemplo de las imágenes obtenidas para el entrenamiento del CNN, para distintas acciones ejecutadas dentro del caso de estudio seleccionado. Aunque la redimensión de descriptores se aplica para la generación de las imágenes base para las CNN, esta estrategia se puede aplicar de igual forma para el desarrollo de otros modelos de clasificación como las neuronales o SVM. Adicionalmente, a partir de la extracción de la secuencia de poses se pueden desarrollar clasificadores basados en Cadenas Ocultas de Markov, que eliminan la dependencia de un tamaño fijo de descriptores.

Figura 12: Ejemplo de imágenes generadas como descriptores de entrada para el entrenamiento de una red CNN.

En contraste al clasificador de acciones estáticas, el clasificador de acciones de desplazamiento basa su operación en el análisis de la posición global de la persona. Dado que la mayoría de las acciones que involucran desplazamiento se encuentran asociadas con correr o caminar, la adición de la información del contexto a la acción es una estrategia que aportará información significativa en el desarrollo del clasificador de actividades. La información del contexto dependerá en gran medida del caso de estudio seleccionado, y puede categorizarse en función del sitio en el cual la persona inició, ejecutó y finalizó la acción de desplazamiento. En el contexto del caso de estudio seleccionado, la información de las acciones girará en torno a los objetos de interés monitoreados, que, para este caso, son los vehículos. De acuerdo con la posición inicial y final de la acción, se pueden generar categorías de acciones como acercarse o alejarse de un vehículo, merodear entre vehículos, caminar sobre la calle, entre otros.

De acuerdo con el modelo propuesto por Saad et al, el último nivel del modelo de inteligencia corresponderá al clasificador de poses a partir de acciones identificadas en la etapa anterior. Al igual que los modelos de reconocimiento utilizados en la detección de acciones, cadenas ocultas de Markov son utilizadas para la clasificación de una secuencia de acciones de longitud variable. Otros trabajos como el desarrollado por el desarrollado por Liu et al utilizan modelos basados en BoW (Bolsa de Palabras en sus siglas en inglés) en la clasificación de actividades a partir de acciones [62]. El diccionario de palabras que compone el modelo de Liu et al está compuesto de patrones de secuencias de acciones (o n-gramas), que se encuentran definidos a partir de 1 o 4 elementos. A partir de los descriptores generados por el modelo BoW, se aplican técnicas de clasificación clásicas como redes neuronales o SVM.

La combinación de clasificadores de acciones y actividades da lugar a varios conjuntos de clasificadores, cuya respuesta es enviada para ser evaluada por el agente ensamble. La combinación de clasificadores permite obtener una organización virtual de agentes de clasificación, en la que se puede aplicar mecanismos de votación en la identificación de la actividad ejecutada por la persona.

6.7. Agente Ensamble

Una vez el conjunto de agentes de clasificación identifica de forma individual la actividad ejecutada por la persona, el agente ensamble analiza las respuestas de cada uno de los agentes, y con base en esta información, determina la respuesta global del modelo de clasificación de actividades. El uso de mecanismos de votación ha sido abordado en trabajos como el de Pit et al, en donde se definen roles y protocolos de comunicación orientados a llevar a cabo el proceso de votación en un sistema multiagente [82]. Dado que el modelo propuesto se encuentra orientado al diseño de un sistema distribuido, el uso de un mecanismo de votación genera robustez frente al fallo o la manipulación externa de alguno de los agentes de la organización.

En el momento en que uno de los agentes de clasificación envía su resultado, el agente de clasificación espera la respuesta de los demás clasificadores en un tiempo Tc. El mecanismo de votación utilizado en el modelo corresponde al voto de pluralidad, en donde cada agente tiene derecho a un único voto y la actividad con más votos es la seleccionada como ganadora [83]. En caso de presentarse un empate, se selecciona la actividad que cuente con mayor confianza reportada por los algoritmos de clasificación. En caso de continuar el empate, o de no contar con una medida de confianza por parte de los algoritmos de clasificación, el agente de votación selecciona la actividad que tenga el menor valor categórico.

6.8. Agente Interfaz

Una vez la actividad ejecutada por la persona es detectada por el modelo de clasificación, la respuesta es enviada al agente interfaz. Además de gestionar la respuesta obtenida por el sistema de detección de actividades, el agente interfaz se encuentra encargado de la gestión de la interfaz de usuario operada por el personal de seguridad y vigilancia. La interfaz de usuario debe contar con características como la visualización de las cámaras en tiempo real, la repro-

ducción y grabación de videos históricos, la administración de la configuración de las cámaras (formato de video, FPS, resolución) y la detección de actividades inusuales.

La detección de una actividad inusual inicia con la respuesta generada por el agente de votación al momento de clasificar una actividad. El modelo de inteligencia del agente interfaz cuenta con una tabla que relaciona cada una de las actividades identificadas por el modelo de acuerdo con el comportamiento de la persona (normal, inusual). La tabla de relación es generada a partir del conocimiento de expertos, y depende en gran medida del contexto sobre el cual se encuentre el caso de estudio seleccionado. En caso de que el agente determine que una de las actividades detectadas corresponde a una actividad inusual, la interfaz genera una alarma sonora y reproduce en pantalla la secuencia de video correspondiente a la actividad identificada. La ubicación y la hora de la acción es mostrada en pantalla, con el objetivo de que el vigilante pueda tomar acciones oportunas frente a la novedad.

En los siguientes capítulos se aterrizará el desarrollo del modelo de agentes y el modelo de inteligencia al caso de estudio seleccionado. Con el objetivo de evaluar el desempeño de cada uno de los clasificadores, se diseñará un protocolo experimental que definirá las actividades objetivo a identificar, la toma de muestras y la recopilación de resultados. Dadas las características de las librerías utilizadas en el desarrollo del módulo de nivel bajo, se utiliza Python como plataforma de desarrollo.

6.9. Aplicación del Modelo de Detección a Otros Contextos

Aunque el desarrollo del modelo de clasificación de actividades se encuentra inspirado en el caso de estudio seleccionado, el uso del método de descomposición de actividades propuesto por Saad et al permite su aplicación en otros contextos. Esto se ve representado en trabajos de autores como Charaaoui et al y Gedat et al, que utilizan la implementación parcial del método de Saad et al para la identificación de acciones cotidianas [52] y actividades relacionadas con deportes [32]. En los modelos de Charaaoui et al y Gedat et al, el elemento base de la detección de la acción es la pose, que es detectada a partir de métodos de aprendizaje supervisado por medio de la definición de poses clave [32], y algoritmos de métodos no supervisado mediante la definición de poses clave identificadas a través de algoritmos de agrupación [52].

Por otra parte, el modelo de inteligencia desarrollado para el agente reidentificación, y en general, el modelo de cooperación desarrollado en el sistema multiagente, permite extrapolar la aplicabilidad del clasificador en sistemas CCTV que se encuentren instalados en recintos cerrados. Al utilizar descriptores de color en conjunto con la información de la posición global, el sistema de agentes puede identificar actividades en CCTVs que cuenten con cámaras sobrelapadas y no sobrelapadas. Por otra parte, el desarrollo de un modelo orientado al uso de sistemas distribuidos favorece la escalabilidad del modelo y permite su implementación en sistemas CCTVs que contengan una gran cantidad de cámaras.

Una aplicación directa de AC-CCTV en la detección de actividades inusuales se encuentra en los edificios de oficinas. Los edificios de oficinas cuentan con una información topológica que puede ser reconstruida para posición global de las personas y evaluar su trayectoria. El contexto de las acciones ejecutadas por la persona se encontraría relacionados al sitio del

edificio (comedor, recepción, oficina), su cercanía con objetos de valor o que contienen información sensible y la franja horaria en que se ejecutó la acción (día, noche). Existe una ventaja evidente en el CCTV instalado en las oficinas contra el CCTV instalado en el caso de estudio, relacionada con la calidad de las imágenes obtenidas por las cámaras. En su artículo. Bharathi et al muestra que en la actualidad, el mercado ofrece resoluciones de las cámaras que varían desde los 720p hasta los 1080p [84]. Estas resoluciones ofrecen una mayor información y calidad que las cámaras instaladas en el centro comercial Oviedo, lo que permite aplicar técnicas avanzadas de reidentificación de personas como el reconocimiento facial [85].

El uso de información topológica y la identificación del contexto de la actividad, permite de igual forma, extrapolar AC-CCTV en edificios como hospitales, en donde el contexto está asociado a identificar los pacientes y las personas que los rodean, o en conjuntos residenciales, en donde el contexto está asociado con los puntos de acceso de las habitaciones y apartamentos. Aunque el proceso general de detección de actividades se mantiene para todos los contextos, el contexto particular de cada instalación dará las pautas para el uso de técnicas específicas, que mejorarán los resultados finales del modelo de clasificación.

6.10. Algoritmo Macro

El proceso mostrado en el Algoritmo 2 describe el conjunto de pasos ejecutados por cada uno de los modelos de inteligencia, desde la captura de la imagen por parte del agente de captura, hasta la emisión de la alerta por parte del agente interfaz. El ciclo que se ejecuta desde el agente de captura hasta el agente reidentificación se realiza por cada imagen emitida por la cámara, mientras que el ciclo que se cumple desde el agente reidentificación hasta el agente interfaz se realiza por cada actividad detectada. Cabe resaltar que el proceso lineal que se muestra en el Algoritmo 2 solo se usa para fines representativos, dado que la arquitectura orientada a agentes desarrollada en el capítulo anterior permite la ejecución de los procesos en paralelo mediante mecanismos pipeline [8].

En este capítulo se desarrolló el modelo de inteligencia para cada uno de los agentes descritos en la arquitectura de AC-CCTV. En el siguiente capítulo se describe los detalles técnicos de la implementación del modelo, la elaboración del protocolo experimental, la ejecución de pruebas y la recopilación de resultados.

Algoritmo 2 : Algoritmo macro – modelo de inteligencia clasificación de actividades

```
1  function ClasificarActividad():
2    imagen = agente_captura.capturar_imagen()
3    lista_poses = agente_pose.detectar_poses(imagen)
4    lista_poses_validas = agente_descriptor.filtrar_poses(lista_poses)
5    lista_descriptores = agente_descriptor.calcular_descriptores(imagen,
                          lista_poses_validas)
6    agente_reidentificacion = procesar_lista_descriptores(lista_descriptores)
7    if agente_reidentificacion.actividad_detectada:
8      for lista_poses in agente_reidentificacion.lista_actividades_detectadas:
9        lista_acciones = agente_organizador.segmentar_acciones(lista_poses)
10       acciones_detectadas = list()
11       for accion in lista_acciones:
12         if accion.movimiento:
13           accion = agente_clasificador.clasificar_accion_estatica(accion)
14         else:
15           accion = agente_clasificador.clasificar_accion_movimiento(acción)
16         acciones_detectadas.append(accion)
17       actividades_detectadas = agente_clasificador.
                                  clasificar_actividad(acciones_detectadas)
18       actividad = agente_ensamble.realizar_votacion(actividades_detectadas)
19       if actividad.sospechosa:
20         agente_interfaz.mostar_alerta()
```

7. IMPLEMENTACIÓN Y RESULTADOS

En este capítulo se describe la implementación del modelo de detección de actividades inusuales desarrollado en los capítulos anteriores. Más que buscar el desarrollo de un producto de software que pueda funcionar en un ambiente de producción, la implementación del modelo tiene como objetivo de evaluar el desempeño del modelo de inteligencia desarrollado en los niveles medio y alto del sistema de detección de actividades. La primera sección del capítulo describe detalles técnicos acerca de la implementación del modelo en componentes de software. Las siguientes secciones del capítulo describe el protocolo experimental desarrollado para la ejecución de pruebas, la recopilación y el análisis de resultados.

7.1. Implementación del modelo de detección de actividades

En el desarrollo del modelo de inteligencia orientado a agentes se seleccionó OpenPose como librería para la implementación de la etapa de bajo nivel. La versión de OpenPose publicada en 2018 cuenta con integraciones para los lenguajes de programación C++ y Python. Aunque se ha demostrado que lenguajes compilados como C++ cuentan con mejor desempeño que lenguajes interpretados como Python [86], se selecciona a Python como lenguaje de programación gracias a que su facilidad de uso permite reducir los tiempos de implementación del software. Adicionalmente, Python cuenta con integraciones a librerías de procesamiento de imágenes e inteligencia artificial como OpenCV, TensorFlow y Caffe.

De acuerdo con lo mencionado en el diseño del modelo de inteligencia, la captura de las imágenes de las cámaras del CCTV se realiza por medio de la implementación de una clase en Python con base en la librería PY-MJPEG [74]. Debido a que las cámaras del CCTV del centro comercial cuentan con el protocolo MJPEG, la clase de captura realiza la detección de cuadros a partir de las marcas de inicio y finalización presentes en una imagen JPEG. Una vez la clase realiza la captura de una imagen, asigna una marca de tiempo que dependerá del reloj del sistema operativo. La captura de las imágenes en el sistema se realiza a una resolución de 640x480 y a una tasa de 2 cuadros por segundo.

La identificación de poses a partir de la librería OpenPose se realiza por medio del modelo BODY_25, que permite detectar los 15 puntos del esqueleto requeridos en el modelo de inteligencia del agente pose. Durante la compilación e instalación de la librería se activó el soporte GPU disponible para las tarjetas gráficas NVIDIA. El uso de GPUs en el procesamiento de las imágenes permite reducir hasta 8 veces el tiempo de procesamiento por imagen [75]. Cabe resaltar que BODY_25 es un modelo pre-entrenado que hace parte del conjunto de herramientas que proporciona la librería OpenPose. En las pruebas desarrolladas con BODY_25, se observó que el modelo detecta de forma precisa los videos capturados del centro comercial, incluso en ejemplos con presencia de oclusiones y baja iluminación. Dado que el enfoque del proyecto de investigación está orientado al análisis de las librerías de alto nivel, se propone como trabajo futuro el análisis del desempeño de la librería OpenPose al entrenar un nuevo modelo con ejemplos del caso de estudio seleccionado.

El cálculo de los descriptores asociados a la pose de la persona se desarrolla por medio de las librerías de computación numérica de Python, y la extracción de los descriptores de color se realiza con ayuda de la librería de procesamiento de imágenes OpenCV. La extracción de los colores dominantes de la persona se efectúa a través de la librería scikit-learn, que cuenta con funciones para aplicar métodos de clusterización por medio del algoritmo K-Means [87]. El cálculo de los descriptores de color a partir de la fórmula CIEDE2000 se realiza por medio de la librería python-colormath, que implementa métodos para realizar conversión de pixeles entre espacios de color y funciones para el cálculo de diferencias de color [88].

El modelo de inteligencia define que el agente de descriptores se encuentra encargado de realizar la clasificación de la pose de la persona. Los modelos de clasificación basados en redes neuronales se desarrollaron a partir de la librería de aprendizaje de máquina TensorFlow. TensorFlow cuenta con APIs de alto nivel que permite el desarrollo de modelos de inteligencia artificial como redes neuronales, CNN (redes neuronales convolucionales por sus siglas en inglés), RNN (redes neuronales recurrentes), entre otros. Adicionalmente, TensorFlow cuenta con integración a las tarjetas de procesamiento gráfico NVIDIA. Al igual que la librería OpenPose, el uso de las GPU en TensorFlow permite la reducción de los tiempos de entrenamiento de los modelos.

Por otra parte, los modelos basados en SVM (máquinas de soporte vectorial) se implementaron por medio de la librería scikit-learn, que cuenta de igual forma con librerías para la implementación de SVM a partir de múltiples funciones de Kernel y permite la implementación de modelos multi-clase a partir del paradigma de entrenamiento uno a muchos [87]. Así mismo, los modelos que requieren el uso de HMM (cadenas ocultas de Markov), se implementaron a partir de la librería HMM-Learn. HMM-Learn es un conjunto de funciones para el uso de modelos basados en HMM, con capacidad de manejar funciones de emisión multinomiales, gaussianas, y modelos de mezcla gaussianas (GMM por sus siglas en inglés) [89].

7.2. Definición de actividades, acciones y poses

De acuerdo con el análisis desarrollado en el capítulo de análisis del caso de estudio, en la Tabla 8 realiza una categorización de las actividades a identificar dentro del sistema CCTV. La categorización define el nombre de la actividad en el sistema, el comportamiento de la persona (usual, sospechoso), y una descripción de las acciones ejecutadas. Hay que resaltar que, de acuerdo con el análisis efectuado en el desarrollo del modelo de inteligencia, se define las acciones de acuerdo con el desplazamiento de la persona en la ejecución de la actividad. Con base en las actividades definidas en la Tabla 8, la Tabla 9 muestra la categorización de las acciones que alimentan el modelo de clasificación de actividades. Cada una de las acciones se clasifica como acción estática o de desplazamiento, con base en la variación de la posición global de la persona al momento de realizar su ejecución.

De acuerdo con el análisis de las actividades efectuado en la Tabla 9, la Tabla 10 definen las poses que alimentarán el modelo de clasificación de poses. La selección de poses base para el modelo de identificación de actividades se encuentra inspirada en el trabajo de Gedat et al, en donde se realiza de forma manual las poses más representativas de acuerdo con la definición de las acciones [32]. De una forma similar al trabajo de Gedat et al, la definición de las cate-

gorías de pose descrita en la Tabla 10 se realizó considerando la orientación de la persona durante su ejecución. Esta distinción se realiza con el objetivo de optimizar el desempeño del clasificador, al momento de usar descriptores asociados a la posición de las articulaciones.

Aunque la definición de las acciones, actividades y poses definen los eventos más relevantes que ocupa un parqueadero, existe la posibilidad de que las actividades de la persona no pertenezcan a alguna de las categorías. Para las categorías relacionadas en la Tabla 9 se adiciona una clase nula, que abarca todos los ejemplos que no clasifican en alguna de las demás clases identificadas. Esta estrategia motiva la aplicación de la estrategia de clasificación uno-a-muchos, en donde se entrenan clasificadores especializados para cada una de las clases identificadas y se obtiene una respuesta conjunta a partir de métodos de ensamble [90].

Tabla 8: Definición de actividades

ID	Nombre	Descripción	Comportamiento
1	Subirse Vehículo	Acercarse al vehículo en un desplazamiento constante. Se puede o no acercar la mano para realizar el movimiento de apertura de la puerta. Abrir la puerta, subirse al vehículo.	Usual
2	Bajarse Vehículo	Descender del vehículo. Cerrar la puerta. Se puede o no realizar el movimiento de cierre de la puerta. Retirarse de la escena en un desplazamiento constante.	Usual
3	Caminar	Realizar el desplazamiento sobre la escena, sin volver al mismo punto y sin realizar numerosos cambios en la dirección de trayectoria.	Usual
4	Forcejear Puerta	Acercarse al vehículo en un desplazamiento constante. Acercar la mano para intentar abrir la puerta. Durar en esa posición más de 10 segundos. Abrir la puerta. Extraer un objeto del interior del vehículo. Cerrar la puerta. Retirarse de la escena a una velocidad constante.	Sospechoso
5	Forcejear Llanta	Acercarse al vehículo en un desplazamiento constante, a un lado de la llanta. Agacharse. Acercar las manos en repetidas ocasiones, intentando extraer elementos de la llanta. Permanecer en esta acción más de 15 segundos. Levantarse. Retirarse del vehículo.	Sospechoso
6	Forcejear Plumillas	Acercarse al vehículo. Estirarse hacia el frente con una o las 2 manos en posición de extraer un elemento. Permanecer en esta misma acción 5 segundos. Retirarse del vehículo.	Sospechoso
7	Merodear	Desplazarse a lo largo del parqueadero, realizando numerosos cambios de dirección.	Sospechoso

7.3. Captura de datos

Cada una de las actividades definidas en la Tabla 8 se desarrollaron por 3 personas, en una de las zonas monitoreada por el CCTV del centro comercial. Antes de realizar la ejecución de las actividades, se realizó la caracterización de la zona en donde se ejecutó la captura de los videos para la validación experimental del modelo. El objetivo de la caracterización de la zona es determinar la posición de los puntos de calibración para cada una de las cámaras, en miras de poder aplicar las técnicas de transformación proyectiva necesarias para el cálculo de la posición global. La Figura 13A y Figura 13B muestra la posición de los puntos tomados para la calibración de cámaras, que se encuentra asociado a las líneas de delimitación entre bahías y la posición de los tope llantas.

Tabla 9: Definición de acciones

ID	Nombre	Categoría
1	Acercarse al lado del vehículo	Desplazamiento
2	Acercarse al frente del vehículo	Desplazamiento
3	Acercarse detrás del vehículo	Desplazamiento
4	Acercarse a zona despejada	Desplazamiento
5	Agacharse	Estático
6	Quedarse quieto	Estático
7	Manipular puerta	Estático
8	Estirar brazos	Estático
9	Nulo	Estático

Tabla 10: Definición de poses

Id	Nombre	Subcategoría	Ejemplo
1	De Pie	Frente	
2		Detrás	
3		Izquierda	
4		Derecha	
5	Brazos al frente	Izquierda	
6		Derecha	
7	Agacharse	Izquierda	
8		Derecha	
9	Brazo Estirado	Izquierda	
10		Derecha	

Cada uno de los actores realizó cada una de las actividades definidas en la Tabla 8, para un total de 90 ejemplos por cada una de las clases. Para completar el entrenamiento del modelo, se tomaron un total de 50 secuencias de video de los registros capturados por el centro comercial. Cada uno de los actores fue inducido a ejecutar las acciones de forma distinta (puerta de entrada del vehículo, velocidad de desplazamiento, lado del vehículo por el cual se ejecuta la acción), con el objetivo de abarcar una muestra diversa para el correcto entrenamiento del modelo.

7.4. Elaboración del Protocolo Experimental

Una vez ejecutada la implementación del modelo y realizada la captura de los videos, se ejecutan los protocolos experimentales para las etapas de nivel medio y nivel alto. El objetivo de obtener la precisión del sistema al clasificar el conjunto de actividades definido en la Tabla 8. Dado que el modelo de inteligencia contempla la implementación de múltiples clasificadores en cada uno de los niveles, la evaluación de desempeño de cada uno de los clasificadores se medirá en función del desempeño del sistema global.

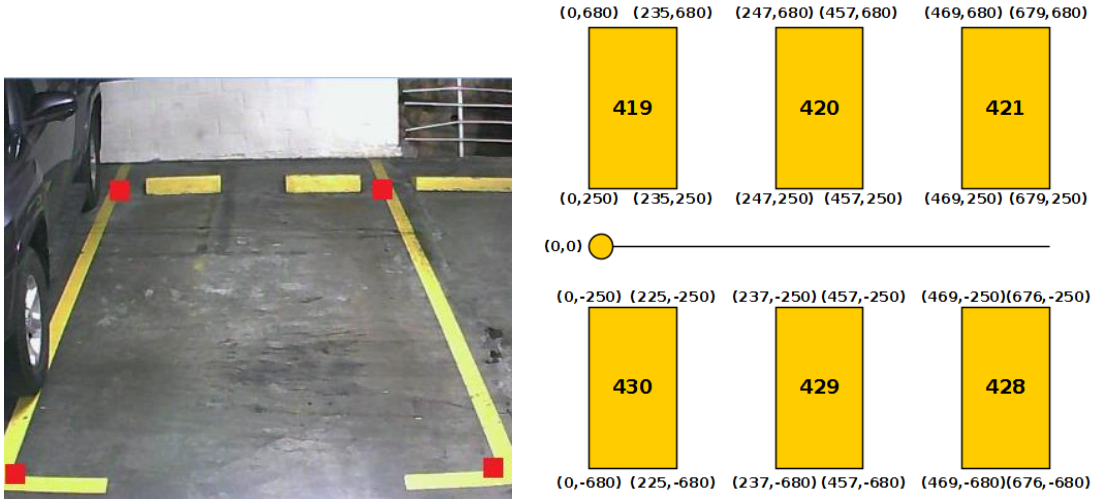

Figura 13: A**: Puntos de calibración seleccionados para el cálculo de la matriz proyectiva, efectuado para una de las cámaras que monitorean la zona en que se hicieron las muestras. B: Coordenadas de los puntos de calibración de cada una de las bahías en que se realizaron la captura de muestras. Los números internos representan la identificación de cada cámara dentro del CCTV. Cada una de las coordenadas se encuentra expresada en centímetros. La línea del centro representa el espacio entre bahías por el cual transitan los vehículos, y sobre el cual se encuentran instaladas las cámaras.**

El protocolo experimental realiza un análisis especial en el clasificador de poses, debido a que el modelo de inteligencia plantea el uso de descriptores de diferentes tipos para la implementación de los clasificadores (poses, ángulos). El desarrollo de los experimentos incluye pruebas del clasificador de poses utilizando diferentes tipos de descriptores, con el objetivo de seleccionar el que demuestre un valor más alto de precisión. El tipo de descriptor seleccionado será la base para el entrenamiento de los clasificadores que participarán en las pruebas globales del sistema.

Aunque el desarrollo del protocolo experimental se centra en el análisis de la etapa de alto nivel, las pruebas ejecutadas en los experimentos también evalúan el desempeño de técnicas como la transformación proyectiva o los algoritmos de reidentificación utilizados en la etapa de nivel medio. La siguiente sección muestra el protocolo de los 3 experimentos ejecutados a partir de los datos capturados en la sección 7.3. En cada uno de los experimentos, se definen las variables independientes, dependientes e intervinientes, su tipo, y su rango de valores.

El proceso experimental se ejecutó en un computador Intel Core I7 6500U con memoria RAM de 8GB y tarjeta gráfica NVIDIA 940MX. El computador de pruebas cuenta con un sistema operativo Linux de 64 bits. La versión de Python que se utilizó la implementación del modelo y la ejecución de pruebas es la 3.6 [91]. Dado que el protocolo experimental se encuentra orientado a la evaluación de niveles medio y alto, se realiza un preprocesamiento

sobre los videos capturados, para extraer las poses generadas a través de la librería OpenPose. El objetivo del preprocesamiento es reducir los tiempos de entrenamiento de los modelos de clasificación, al adicionar información a la imagen que es calculada por las etapas de bajo nivel.

Experimento 1: Medición del error obtenido en la ejecución de la transformación proyectiva

OBJETIVO DEL EXPERIMENTO	Obtener una medida del error al realizar el cambio del espacio de coordenadas a partir de técnicas de homografía.
PROCEDIMIENTO	Seleccionar una de las bahías que componen el parqueadero del centro comercial. Realizar las marcas sobre el piso de 20 puntos de prueba, divididos en 4 grupos. 2 de los grupos de puntos tendrán variaciones en el eje Y mientras su valor en X permanece constante. Por otra parte, los otros grupos restantes tendrán variaciones en el eje X mientras el eje Y permanece constante. Aplicar la transformación proyectiva sobre cada uno de los puntos marcados en la imagen.
V. INDEPENDIENTES	**(x, y)** Coordenadas de la posición del punto sobre la imagen • Tipo de variable: numérica • Unidad: Pixeles • Valores Posibles en el experimento: 20
V. DEPENDIENTES	**e**: Error en la posición entre el píxel proyectado por la transformación proyectiva y el valor real • Tipo de variable: numérica • Unidad: Centímetros
V. INTERVINIENTES	• Distorsión producida por el lente de la cámara. El control de la variable se realiza ejecutando el experimento sobre una misma cámara.

Experimento 2: Evaluación del desempeño del clasificador de poses

OBJETIVO DEL EXPERIMENTO	Definir el tipo de descriptor que permita obtener el mayor porcentaje de precisión en el clasificador de poses.
PROCEDIMIENTO	Del conjunto de videos recopilado en la sección 7.3, generar una base de datos que contenga 70 muestras de cada una de las poses definidas en la Tabla 3. Realizar el entrenamiento de los clasificadores evaluados en el modelo de inteligencia (KNN, Redes Neuronales), recalculando las entradas del modelo según el tipo de descriptor. Obtener la precisión de cada uno de los modelos.
V. INDEPENDIENTES	**des:** Tipo de descriptor aplicado al modelo de clasificación Tipo de variable: categórica • Valores posibles en el experimento: ángulos, poses, transformación de cosenos de ángulos, mezcla entre poses y ángulos. **mod:** Modelo de clasificación: • Tipo de variable: categórica • Valores posibles en el experimento: KNN, SVM
V. DEPENDIENTES	**p**: Precisión del descriptor • Tipo de variable: numérica • Unidad: Porcentaje **h:** Matriz de confusión: • Tipo de variable: Matriz • Unidad: Números enteros.
V. INTERVINIENTES	• Iluminación de la escena: La ejecución de las escenas se realiza dentro de un

	ambiente controlado con iluminación artificial, evitando los cambios de iluminación generados por la luz natural. • Características intrínsecas de la cámara: El mismo conjunto de cámaras se utilizaron para realizar la captura de las secuencias de video. • Tipo de clasificador utilizado para el protocolo experimental: Se selecciona una red neuronal para la ejecución de las pruebas.

Experimento 3: Evaluación de desempeño del sistema de clasificación de actividades

OBJETIVO DEL EXPERIMENTO	Obtener el nivel de precisión del modelo de clasificación de actividades.
PROCEDIMIENTO	Del conjunto de videos recopilado en la sección 7.3, dividir los ejemplos capturados en sets de entrenamiento y validación. De los videos capturados en el set de entrenamiento, generar 3 bases de datos que funcionarán como base para el entrenamiento de los modelos. Las bases de datos deben contar con la siguiente relación. • Poses: 70 ejemplos por clase (Tabla 10) • Acciones: 63 ejemplos por clase (Tabla 9) • Actividades: 63 ejemplos por clase (Tabla 8) Entrenar los modelos del sistema usando las bases de datos de entrenamiento. El entrenamiento se debe realizar de forma escalonada, dado que los clasificadores de niveles superiores requieren información de los niveles inferiores. Una vez entrenados los clasificadores, evaluar la precisión del sistema global utilizando los videos de validación.
V. INDEPENDIENTES	**modPose:** Modelo de clasificación de poses. • Tipo de variable: categórica • Valores posibles en el experimento: Redes Neuronales, SVM **modAcciones:** Modelo de clasificación de acciones. • Tipo de variable: categórica • Valores posibles en el experimento: CNN, SVM, HMM **modActividades:** Modelo de clasificación de actividades. • Tipo de variable: categórica • Valores posibles en el experimento: BoW, HMM **modEnsamble:** Método de ensamble utilizado en el experimento • Tipo de variable: categórica • Valores posibles en el experimento: Votación
V. DEPENDIENTES	**p**: Precisión del descriptor • Tipo de variable: numérica • Unidad: Porcentaje **h:** Matriz de confusión: • Tipo de variable: Matriz • Unidad: Números enteros
V. INTERVINIENTES	Iluminación de la escena: La ejecución de las escenas se realiza dentro de un ambiente controlado con iluminación artificial, evitando los cambios de iluminación generados por la luz natural. Características intrínsecas de la cámara: La escena se desarrolló utilizando el mismo conjunto de cámaras.

7.5. Resultados

A continuación, se describen los resultados obtenidos para cada uno de los experimentos relacionados en la sección anterior. El objetivo de los experimentos es evaluar la precisión y el tiempo de respuesta de cada uno de los componentes del modelo, con el objetivo de evaluar su uso potencial en el caso de estudio seleccionado.

Ejecución del Experimento 1: Medición del error obtenido en la ejecución de la transformación proyectiva

La Tabla 11 muestra los resultados obtenidos después de aplicar el protocolo descrito en el Experimento 1. Se puede observar que la diferencia más significativa del conjunto de puntos con respecto a su referencia es de 8.3 cm, que es aceptable en comparación al rango de movimientos que puede tener una persona al ejecutar una acción. Además, se puede observar que a medida que aumenta el valor en Y aumenta la imprecisión en el cálculo de la posición global. Este comportamiento se debe a la perspectiva dentro de la imagen, dado que variaciones pequeñas en la posición local generan mayores variaciones en la posición global, a medida que aumenta el valor de Y.

Aunque estos resultados muestran que el error producido por la transformación proyectiva no es significativo, el error en el cálculo de la posición de la persona puede variar por errores en la detección de la silueta, generados por oclusiones o bajos niveles de iluminación en la imagen. Por lo tanto, es importante que los algoritmos de reidentificación sean tolerantes frente a variaciones en la posición de la persona, con el objetivo de lograr un óptimo funcionamiento.

Tabla 11: Resultados comparativos del desempeño de la homografía para el cálculo de la posición global

Experimento 1.A y 1.B - Variación en Y					**Experimento 1.C y 1.D. - Variación en X**				
	Eje X Constante (X = 192 cm)		**Eje X Constante (X = 26 cm)**			**Eje Y Constante (Y = 322 cm)**		**Eje Y Constante (Y = 199 cm)**	
Medición Y (cm)	Proyección Y (cm)	Diferencia (cm)	Proyección Y (cm)	Diferencia (cm)	**Medición X (cm)**	Proyección X (cm)	Diferencia (cm)	Proyección X (cm)	Diferencia (cm)
0.00	-0.62	0.62	0.86	-0.86	**0**	0.23	-0.23	-1.91	1.91
105.00	106.98	-1.98	107.52	-2.52	**57**	55	2.00	55.83	1.17
210.00	217.70	-7.70	218.6	-8.60	**114**	109	5.00	114.43	-0.43
315.00	321.00	-6.00	323.3	-8.30	**171**	165	6.00	172.47	-1.47

Ejecución del

Experimento 2: Evaluación del desempeño del clasificador de poses

Con el objetivo de limitar el espacio de posibilidades descrito en el estado del arte para el diseño del clasificador de poses, el objetivo del

Experimento 2 es seleccionar el tipo de descriptores que generen mejores resultados en la clasificación de poses. De acuerdo con los trabajos de Gedat et al [32] y Du et al [31], se seleccionan la posición de las articulaciones y los ángulos como los descriptores principales para realizar la clasificación de poses. Una consideración importante es que los ángulos, al ser descriptores simétricos, no pueden realizar la clasificación de las subcategorías definidas en la Tabla 10. Por lo tanto, el entrenamiento de los clasificadores cuyos descriptores se encuentra basados exclusivamente en ángulos se realiza mediante la definición de 5 poses extrapoladas en las poses de la Tabla 10: de pie frontal, de pie lateral, brazos al frente, agacharse y brazo estirado.

Por otra parte, es importante resaltar que el uso directo de descriptores asociados en ángulos cuenta con algunos problemas de convergencia, dado que su naturaleza periódica genera que valores como 0° y 359° sean detectados como distantes por el clasificador de poses. Para solucionar esta novedad, autores como Potavov et al proponen la división del ángulo en 2 descriptores: $\mathbf{cos}(\boldsymbol{\alpha})$ y $\mathbf{sin}(\boldsymbol{\alpha})$, con el objetivo de eliminar los inconvenientes generados por la periodicidad del ángulo en los clasificadores [92]. El protocolo descrito en el Experimento 2 analiza el desempeño del clasificador al realizar la transformación del descriptor de ángulos, en combinación con el descriptor de posición de las articulaciones. La Tabla 12 muestra los resultados de la ejecución del protocolo experimental.

Los resultados contenidos en la Tabla 12 evidencian que el descriptor óptimo para la clasificación de poses es la posición de las articulaciones. Aunque se evidencia que la transformación de los ángulos mejora en todas las ocasiones el porcentaje de precisión del modelo se concluye que la información de los ángulos se encuentra implícita en la posición de las articulaciones, y su adición en el vector de descriptores no genera más que ruido en el sistema.

Tabla 12: Resultados del clasificador de poses al realizar variaciones en el tipo de descriptor.

Tipo de Descriptor	Loss	Precisión
Ángulos	0.4668	82.30%
Transformación de ángulos	0.347	86.60%
Posición de articulaciones	0.424	90.50%
Posición de articulaciones + Ángulos	0.4025	88.40%
Posición de articulaciones + Transformación de ángulos	0.3732	89.70%

La Tabla 13 muestra la matriz de confusión del clasificador de poses que presentó mayor precisión en el experimento. La matriz de confusión evidencia que el mayor error de clasificación se encuentra al momento de presentarse la pose de brazo al frente hacia la derecha. De acuerdo con la definición de poses elaborada en la Tabla 10, este resultado es de esperarse dado la similitud que existen entre las poses de pie y brazo estirado. Por otra parte, se observa que las poses con las cuales el sistema cuenta con mayor precisión son de pie y agachado.

Tabla 13: Matriz de confusión del clasificador de poses

	DPD	BFI	BFD	DPF	DPI	DPR	AD	AI	ED	EI	
De Pie Detrás	25	0	0	0	0	1	0	0	0	0	96.15%
Brazo Frente Izquierda	0	24	0	2	0	0	0	0	1	0	88.89%
Brazo Frente Derecha	1	0	18	3	0	1	0	0	0	0	78.26%
De Pie Frente	1	0	0	23	0	1	0	0	0	0	92.00%
De Pie Izquierda	1	1	0	0	19	0	0	0	0	0	90.48%
De Pie Derecha	0	0	2	1	0	18	0	0	0	0	85.71%
Agachado Derecha	0	0	0	0	0	1	25	0	0	0	96.15%
Agachado Izquierda	0	0	0	0	0	1	0	27	0	1	93.10%
Estirado Derecha	0	2	0	0	0	0	0	0	10	0	83.33%
Estirado Izquierda	0	0	2	0	0	0	0	0	0	20	90.91%

Ejecución del Experimento 3: Evaluación de desempeño del sistema de clasificación de actividades

La evaluación de la precisión de los clasificadores de poses, acciones y actividades se evalúan en conjunto por medio del protocolo definido en el Experimento 3. El protocolo experimental tiene como objetivo seleccionar el mejor modelo de clasificación de actividades, tomando como base el paradigma de clasificación por etapas (poses, acciones y actividades) definido por Saad et al. El protocolo definido en el Experimento 3 cuenta con 2 particularidades. La primera particularidad es el clasificador CNN utilizado para la detección de acciones en el experimento. El clasificador CNN se encuentra inspirado en el trabajo de Du et al, cuyo modelo realiza la clasificación de acciones usando directamente la posición de las articulaciones, sin realizar una etapa previa que realice la detección de las poses [31]. A partir de los resultados del CNN con los otros clasificadores, se puede discutir la posibilidad de juntar la etapa de detección de poses y acciones en una sola, argumentando que la posición de los puntos del esqueleto contiene de forma implícita la información de la pose de la persona.

La segunda particularidad del experimento es el uso o no de métodos de ensamble como variable independiente en la clasificación de actividades. Según el trabajo elaborado por Dietterich et al, el uso de métodos de ensamble en un conjunto de clasificadores que cuentan con un grado de diversidad produce mejores resultados que el uso de cada uno de los clasificadores de forma individual [12]. Dado que protocolo experimental efectúa la combinación de diferentes clasificadores para la detección de actividades, los 10 clasificadores del experimento proporcionan un conjunto diverso para la ejecución de métodos de ensamble.

El set de datos utilizado para la clasificación de actividades se divide en 3 partes. La primera parte corresponde al 60%, el cual es utilizado para el entrenamiento de los clasificadores en cada una de las etapas. El 20% del set corresponde al set de validación, que permite obtener los valores de ponderación que son utilizados en el método de ensamble, que dependen del nivel de precisión de cada uno de los 10 modelos. Además, el set de validación es utilizado en el entrenamiento de los clasificadores basados en HMM, dado que su entrenamiento se realiza en varias iteraciones y se selecciona el modelo con mayor porcentaje de precisión. Finalmente, el 20% restante corresponde al set de Prueba, que permite medir el desempeño real de cada uno de los clasificadores. La Tabla 14 muestra los resultados de precisión de cada combinación de clasificadores.

Tabla 14: Resultados del modelo de clasificación de actividades.

Experimento	Modelo Poses	Modelo Acciones	Modelo Actividades	Precisión
1	NN	NN	HMM	68.10%
2			BoW	88.90%
3		HMM	HMM	68.10%
4			BoW	87.93%
5	SVM	NN	HMM	69.80%
6			BoW	90.51%
7		HMM	HMM	64.65%
8			BoW	83.86%
9	CNN		HMM	72.40%
10			BoW	93.10%
11	Ensamble			91.37%

Los resultados obtenidos en la Tabla 14 muestra que el mejor desempeño se obtuvo utilizando el modelo CNN en la clasificación de acciones, en conjunto con el modelo BoW en la clasificación de actividades. A pesar de que Dietterich et al afirmaron que el uso de métodos de ensamble proporciona mejores resultados que el uso de clasificadores en conjunto [12], los resultados del experimento manifiestan que no se logró una mejora con respecto al resultado contenido en el experimento 10. El fracaso del método de ensamble en el experimento se puede justificar cuando el error en la predicción de los modelos que intervienen en el ensamble se encuentra correlacionado.

Aunque los resultados obtenidos al usar métodos de ensamble no supera el resultado de usar el mejor clasificador de forma individual, la Tabla 14 muestra que la precisión del clasificador ensamble se encuentra aproximadamente 2% por debajo del mejor resultado. Teniendo en cuenta que en el conjunto de 10 clasificadores existen 5 que cuentan con niveles de precisión inferiores al 75%, se puede concluir que el resultado del clasificador ensamble se encuentra por encima del promedio de los clasificadores. Adicionalmente, se evidencia que el uso de técnicas de ensamble reduce en alguna medida el riesgo de seleccionar modelos que generen

buenos resultados en la etapa de validación, pero que cuenten con mal desempeño durante la etapa de pruebas.

Utilizando los resultados contenidos en la Tabla 14, la Tabla 15 evalúa el desempeño individual de cada uno de los clasificadores dentro del modelo de reconocimiento. Para cada uno de los resultados en donde el clasificador participa en el experimento, se calcula la media y la desviación estándar del conjunto de registros. Este método de evaluación es válido únicamente en el contexto del experimento, y evalúa el desempeño del clasificador en conjunto con los clasificadores de las demás etapas.

Tabla 15: Desempeño individual de cada uno de los clasificadores, con base los resultados de desempeño del sistema global.

Poses			Acciones			Actividades		
Clasificador	**Media**	**Desviación**	**Clasificador**	**Media**	**Desviación**	**Clasificador**	**Media**	**Desviación**
NN	78.26%	11.74%	NN	79.33%	12.02%	HMM	68.61%	2.83%
SVM	77.21%	12.02%	HMM	76.14%	11.48%	BoW	88.86%	3.41%
			CNN	82.75%	14.64%	Ensamble	91.37%	0.00%

La Tabla 15 muestra que el porcentaje de clasificación de actividades aumenta notablemente al usar modelos tipo BoW sobre HMM en la última etapa. Este resultado se puede justificar debido a que las cadenas ocultas de Markov no almacenan información sobre los estados anteriores producidos por una persona durante la ejecución de una actividad. Esta característica ocasiona que las HMM omite información relevante en la discriminación de actividades como merodear o caminar, generando el aumento del porcentaje de error en la clasificación

Por otra parte, el clasificador que presentó mejor desempeño en la etapa de identificación de acciones fue el modelo basado en CNN. Hay que resaltar que el modelo basado en CNN está inspirado en el trabajo de Du et al, en donde la detección de acciones se realiza de forma directa a partir de la posición de las articulaciones [31]. Estos resultados muestran que, a partir del uso de CNN se puede unificar en un solo nivel la identificación de poses y acciones, eliminando la necesidad de definir las poses base para cada una de las actividades de forma similar a lo desarrollado en la Tabla 10. Por otra parte, hay que resaltar que el uso del CNN involucra un mayor costo computacional que los otros métodos (NN, HMM), lo cual puede impactar de forma negativa en los tiempos de respuesta del sistema.

Adicionalmente, la Tabla 15 muestra que el porcentaje de precisión clasificador HMM alcanzó un nivel aceptable con respecto al desempeño de los demás clasificadores. Sin embargo, el desempeño aceptable no fue suficiente para que HMM presentara el peor desempeño del clasificador de acciones. El bajo desempeño del HMM se puede asociar, de igual forma con la incapacidad de procesar información sobre los estados anteriores, una desventaja que impacta el resultado del sistema global de clasificación de actividades.

La Tabla 16 muestra los resultados de clasificación del clasificador de poses que presentó mayor porcentaje de precisión en la clasificación de actividades (CNN-BoW). A pesar de que

el porcentaje de clasificación supera el 90%, se puede observar que el mayor error genera al momento de clasificar la actividad bajarse del vehículo, que se confunde en mayor medida a la actividad subirse al vehículo. Esto se puede justificar por la forma como se diseñó el clasificador BoW, dado que la extracción del vector de clasificación se extrajo a partir del conteo de unigramas, ignorando la información secuencial contenida en el vector de acciones. Una forma de introducir información secuencial en el vector de acciones es realizar el conteo de bigramas, el cual representa una secuencia de dos acciones.

Tabla 16: Matriz de confusión del clasificador de acciones CNN-BoW

	FPu	**BV**	**ME**	**FPl**	**FL**	**SV**	**CA**	**Precisión**
Forcejear Puerta	**16**	0	0	0	1	0	0	94.12%
Bajarse vehículo	1	**12**	0	0	0	4	0	**70.59%**
Merodear	0	0	**16**	0	0	0	0	100.00%
Forcejear Plumillas	0	0	0	**17**	0	0	0	100.00%
Forcejear Llantas	0	0	0	0	**15**	0	0	100.00%
Subirse al vehículo	0	0	0	0	0	**16**	1	94.12%
Caminar	0	1	0	0	0	0	**16**	94.12%

Medición del desempeño del modelo de inteligencia de los agentes

Con el objetivo de medir el desempeño del sistema para validar su potencial aplicación en el caso de estudio seleccionado, la Tabla 17 muestra el tiempo requerido por cada uno de los agentes que requieren mayor cantidad de recursos, para la ejecución de las funciones descritas en el Algoritmo 2. La captura de tiempos se realizó en un computador Intel Core I7 6500U con memoria RAM de 8GB y tarjeta gráfica NVIDIA 940MX.

Los resultados contenidos en la Tabla 17 muestran que el agente que más recursos computacionales requiere en la clasificación de actividades es el agente pose (433ms en el procesamiento de una única imagen). Este valor se encuentra acorde a la tabla de tiempos publicado en la página oficial de la librería [75]. Cabe resaltar que el procesamiento de las poses se realiza con ayuda de la tarjeta gráfica NVIDIA con la cual cuenta la máquina. Sin ayuda de la tarjeta gráfica, el procesamiento que requiere la librería por cada una de las imágenes alcanzaría un valor cercano a los 4 segundos [75].

Por otra parte, aunque el desempeño del agente reidentificación se encuentra por encima del agente pose, el tiempo que requiere para el procesamiento de las imágenes aún sigue siendo considerable con respecto a los otros componentes del modelo. El tiempo que consume el agente reidentificación se justifica con el uso del algoritmo K-Means al momento de realizar la extracción de los colores dominantes de la persona. Además, los recursos computacionales requeridos por el agente reidentificación aumentarán a medida que aumente el número de personas en la escena, dado que aumenta el número de veces que el sistema debe aplicar el algoritmo sobre la imagen.

Tabla 17: Mediciones de tiempos en procesos internos de los agentes

Muestra	Poses (ms)	Descriptores (ms)	Reidentificación (ms)	NN Poses (ms)	SVM-Poses (ms)	HMM-Acciones (ms)	CNN-Acciones (ms)
1	429	16.5	125.8	0.877	0.117	1.149	6.001
2	443	10.3	172.5	0.531	0.116	1.284	6.003
3	437	9.1	154.9	0.554	0.116	0.968	7.54
4	431	12.0	156	0.584	0.117	0.373	7.173
5	427	7.4	149.2	0.577	0.117	1.113	8.619
6	436	12.3	169.9	0.694	0.117	0.707	10.809
7	434	12.0	147.5	0.052	0.117	1.05	7.17
8	438	9.6	173.5	1.05	0.163	0.617	7.249
9	427	10.2	164.1	0.655	0.116	0.441	6.859
10	430	9.6	161.1	0.55	0.117	1.142	5.917
Media	**433.2**	**10.9**	**157.5**	**0.6**	**0.1**	**0.9**	**7.3**
Desviación	**5.28**	**2.5**	**14.4**	**0.3**	**0.0**	**0.3**	**1.5**

Adicionalmente, se puede observar que el tiempo de ejecución de los algoritmos de clasificación es muy bajo, dado que utilizan modelos que ya se encuentran entrenados. Como se comentó en el análisis de los experimentos, el CNN es el algoritmo que más recursos computacionales requiere para su operación. Si bien los resultados de las mediciones mostraron un desempeño aceptable del CNN con respecto a los otros módulos, cabe resaltar que la librería utilizada para la implementación de las CNN cuenta con integración para las tarjetas gráficas NVIDIA. Entre los modelos de clasificación seleccionados para la medición, el SVN fue el que presentó el mejor desempeño.

Dado que el objetivo del modelo de detección de actividades es realizar la implementación del modelo en un sistema distribuido, es importante reducir los requerimientos computacionales de cada uno de los módulos con el objetivo de que puedan trabajar sobre sistemas embebidos. Los tiempos computacionales de la Tabla 17 muestra que es fundamental optimizar la detección de poses en el sistema, ya sea adicionando componentes de hardware que aceleren el procesamiento (por ejemplo, VPU o unidades de procesamiento de video por sus siglas en inglés), o por medio del uso de técnicas que requieran menores recursos computacionales.

Los resultados de los experimentos muestran un funcionamiento sobresaliente en el modelo de detección de actividades, al utilizar una combinación de CNN en la detección de acciones en conjunto con BoW en la detección de actividades. De acuerdo con la matriz de confusión mostrada en la Tabla 16, se observó que el modelo de detección de actividades puede ser mejorado al incluir el conteo de bigramas en el vector de características del modelo BoW. La siguiente sección contiene las conclusiones del proceso de diseño, implementación y validación de AC-CCTV, así como el trabajo futuro.

8. CONCLUSIONES

En este trabajo de grado se presentó AC-CCTV, un sistema de reconocimiento de actividades orientado a optimizar las labores del personal de seguridad y vigilancia en la detección oportuna de actividades inusuales. El desarrollo de AC-CCTV se efectuó en función del objetivo general del trabajo de investigación, buscando proporcionar un modelo de detección de actividades inusuales que pueda ser aplicado en recintos cerrados. Aunque la idea de investigación y el caso de estudio seleccionado se encuentran asociados a sistemas CCTV instalados en parqueaderos, el modelo desarrollado para AC-CCTV puede ser aplicado a otros contextos, gracias que las actividades son definidas a partir de elementos básicos como la trayectoria de la persona y las poses ejecutadas.

A partir de la definición de los objetivos específicos, AC-CCTV proporciona un grado de novedad al basar su arquitectura en una organización de agentes racionales. Como se mencionó en el capítulo de diseño del sistema orientado a agentes, la naturaleza distribuida de los sistemas CCTV hacen que los modelos orientados a agentes sean idóneos para el desarrollo de sistemas de clasificación de actividades. Los sistemas orientados a agentes permiten el desarrollo de estrategias cooperativas que explotan la información contextual proporcionada por múltiples cámaras. Además, la implementación de propiedades el control de recursos compartidos [5] y el uso de operaciones concurrentes [6], aumentan la escalabilidad del sistema y mejoran su rendimiento al realizar procesamiento paralelo de tareas en modo pipeline [8].

Con base en los modelos de detección de actividades propuesto por Cristiani et al [9], AC-CCTV cuenta con un diseño dividido en 3 etapas. El desarrollo de la etapa de bajo nivel describió los retos principales que deben superarse al momento de realizar la detección de las personas en la escena, que corresponden al control de oclusiones, el área de cobertura y la variación en los niveles de iluminación. Aunque el estado del arte mostró que los algoritmos de extracción de fondo son candidatos válidos en el diseño de sistemas de detección de personas, los resultados de la evaluación recopilados en la Tabla 7 comprobaron que estos algoritmos cuentan con un bajo nivel de desempeño en entornos con alto nivel de movimiento. Por otra parte, la extracción de esqueletos utilizada a través de la librería OpenPose mostró buenos resultados, con la desventaja de que su uso requiere altos recursos computacionales que penalizan el desempeño del sistema.

La etapa de nivel medio se centró en el desarrollo del módulo de reidentificación de personas, cuya parte se centró en el desarrollo del agente de reidentificación a partir de técnicas de comparación de color y transformación proyectiva. Los resultados recopilados en la Tabla 13 muestran que la transformación proyectiva muestra buenos resultados al momento de aplicarse en cámaras que no cuentan con un alto grado de proyección, como lo son las cámaras que componen el caso de estudio seleccionado. Por otra parte, una de las diferencias del modelo de AC-CCTV con respecto a otros modelos como el Jang et al es el uso de la fórmula de comparación CIEDE2000. Además de proporcionar una escala de color que se asemeja al ojo

humano, la ecuación CIEDE2000 elimina la ambigüedad en la componente H del espacio HSV al realizar el análisis de un color en escala de grises [50]. Finalmente, la extracción de los colores dominantes de una persona a partir del algoritmo K-Means, genera un grado de robustez en la posición de la persona con respecto a la cámara al momento de realizar la reidentificación.

El desarrollo del módulo de alto nivel se elaboró con base en el modelo de actividades propuesto por Saad et al, en donde una actividad se define mediante acciones y una acción se define mediante poses [11]. De acuerdo con los resultados recopilados en la Tabla 15, el uso de modelos CNN en la clasificación de acciones permite unificar los niveles de pose y acción, eliminando la necesidad de definir las poses base para cada una de las actividades de forma similar a lo desarrollado en la Tabla 10. Adicionalmente, a pesar de que las técnicas de ensamble se introducen en el modelo con el objetivo de obtener un mayor nivel de precisión en la clasificación [12], los resultados contenidos en la Tabla 14 no muestran una mejora al aplicar esta estrategia en el modelo.

A partir de los criterios estipulados en los objetivos específicos, el Experimento 3 realizó la evaluación de la precisión global del sistema, cuyos resultados se resumen en la Tabla 10. El mejor resultado del experimento muestra una precisión del 93.9%, al combinar CNN en la detección de acciones con BoW en la detección de actividades. Es importante resaltar que el sistema actúa únicamente de forma informativa en la detección de actividades inusuales, y que el vigilante es el responsable de realizar la evaluación de la actividad y de ejecutar las acciones correspondientes.

Por medio de las mediciones recopiladas en la Tabla 17, se obtuvo una medición del desempeño de cada uno de los procesos que consumen mayor cantidad de recursos. De acuerdo con los resultados presentados, se identificó que los 2 agentes que se deben optimizar para lograr la implementación de AC-CCTV en un sistema de producción es el agente de poses y el agente de reidentificación. La optimización del agente reidentificación se puede lograr con la utilización de algoritmos alternos al K-Means que permitan la extracción de los colores dominantes. Por otra parte, la optimización de la extracción puede alcanzarse mediante el uso de modelos pre entrenados que ofrezcan un menor nivel de precisión, pero que favorezcan los tiempos de ejecución [93]. Los tiempos obtenidos en la Tabla 17 muestra la imposibilidad de ejecutar el sistema en tiempo real a una tasa de 2FPS, valor utilizado en el CCTV instalado en el Centro Comercial Oviedo. Además de requerir grandes recursos computacionales para el procesamiento de las cámaras, disminuir la tasa de FPS de las cámaras afectaría la fluidez del video percibida por el personal de monitoreo, afectando la usabilidad del sistema.

8.1. Trabajo Futuro

Con el objetivo de lograr la implementación de AC-CCTV en el caso de estudio seleccionado, es fundamental optimizar los recursos computacionales requeridos las librerías de detección de poses utilizadas en las etapas de bajo nivel. Dado que los sistemas CCTV instalados por Controles Inteligentes cuentan con sistemas embebidos instalados de forma anexa a cada una de las cámaras, es deseable que los algoritmos que componen el modelo de AC-CCTV puedan ser ejecutados en estos dispositivos. Para acelerar el cálculo de los procesos compu-

tacionales ejecutados por OpenPose, se validará la integración de la librería al dispositivo de computación gráfica Movidius desarrollado por Intel. Movidius es un dispositivo externo compatible con sistemas embebidos como Raspberry Pi, que acelera el entrenamiento y la ejecución de modelos basados en Deep Learning [13].

Para generar un valor agregado, el modelo de AC-CCTV se integrará al sistema de guiado proporcionado por Controles Inteligentes. El sistema de guiado cuenta con una interfaz gráfica desarrollada en .NET, la cual recibirá las alertas generadas por el sistema de agentes al momento de realizar la detección de una actividad sospechosa. En caso de que el operador no se encuentre presente para responder la alerta, AC-CCTV integrará funcionalidades como la grabación de registros de video de la actividad y la elaboración de un histórico de alertas, que permitan al vigilante revisar las novedades reportadas del sistema en un periodo de tiempo.

Buscando validar la aplicabilidad de AC-CCTV en otros contextos diferentes al caso de estudio seleccionado, es importante desarrollar un protocolo experimental en una empresa que cuente con un CCTV instalado en un recinto cerrado y que no se encuentre asociado al contexto de parqueaderos. De acuerdo con el análisis desarrollado en el diseño del modelo, se implementará el protocolo en instalaciones que cuenten con oficinas. El protocolo experimental tendrá como objetivo de validar aspectos como el seguimiento de la persona a partir de su posición global, el uso de los descriptores de color, y la identificación del contexto dentro de la etapa de clasificación de acciones. Con base en la calidad de las imágenes obtenidas en el CCTV, se evaluará la posibilidad de implementar técnicas avanzadas de rastreo de personas, como el reconocimiento facial [85].

REFERENCIAS

[1] SuperVigilancia, «Estado del Sector de Vigilancia y Seguridad Privada en Colombia,» 2015.

[2] S. Leman-Langlois, «The Myopic Panopticon: The Social Consequences of Policing Through the Lens,» *Policing and Society,* pp. 44-58, 2003.

[3] J. Ferenbok y A. Clement, «Hidden Changes: from CCTV to "Smart" video surveillance,» 2015. [En línea]. Available: https://www.yumpu.com/en/document/view/34561291/pdf-hidden-changes-from-cctv-to-smart-video-surveillance. [Último acceso: 11 noviembre 2018].

[4] N. Ejaz, U. Manzoor, S. Nefti y S. Wook, «A collaborative multi-agent framework for abnormal activity detection in crowded areas,» *International Journal of Innovative Computing,* vol. 8, nº 6, pp. 4219-4234, 2012.

[5] A. Cicortas y V. Iordan, «Multi-Agent Systems for Resource Allocation,» *Technology and Economics of Smart Grids and Sustainable Energy,* pp. 3-15, 2018.

[6] F. Maturana, W. Shen, M. Hong y D. Norrie, «Multi-agent Architectures for Concurrent Design and Manufacturing,» 2004.

[7] L. Panait y S. Luke, «Cooperative Multi-Agent Learning: The State of the Art,» *Autonomous Agents and Multi-Agent Systems,* pp. 387-434, 2004.

[8] C. Ramamoorthy, «Pipeline Architecture,» *Petitioner Apple Inc,* 1997.

[9] M. Cristiani, R. Raghavendra, A. Del Blue y V. Murino, «Human Behavior Analysis in Video Surveillance: a Social Signal Processing Perspective,» *Neurocomputing Journal,* 10 Diciembre 2011.

[10] CMU-Perceptual-Computing-Lab, «OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation,» 2018. [En línea]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose. [Último acceso: 11 noviembre 2018].

[11] M. Saad, A. Hussain y W. Kong, «SESRG-InViSS : Image and Video Data Set For Human Pose, Action, Activity and Behaviour,» *2013 IEEE 3rd International Conference on System*

Engineering and Technology,, pp. 37-42, 2013.

[12] T. G. Dietterich, «Ensemble Methods in Machine Learning,» *International workshop on multiple classifier systems*, 2000.

[13] Intel, «Intel Movidius Neural Compute Stick,» [En línea]. Available: https://software.intel.com/en-us/neural-compute-stick. [Último acceso: 12 noviembre 2018].

[14] DANE, «Encuesta de convivencia y seguridad ciudadana,» 2015.

[15] Veeduría Distrital, «Estado de la política pública de seguridad y convivencia de Bogotá D.C.,» 2017.

[16] C. Vivien, «Valoración del CCTV como una Herramienta efectiva de manejo y seguridad para la resolución, prevención y reducción de crímenes,» Montreal, 2018.

[17] SIE, «El importante rol del sistema de CCTV en la Seguridad Privada,» 6 julio 2015. [En línea]. Available: https://siesa.com.ar/el-importante-rol-del-sistema-de-cctv-en-la-seguridad-privada/. [Último acceso: 11 noviembre 2018].

[18] P. Turaga, R. Chellapa, V. Subrahmanian y O. Udrea, «Machine Recognition of Human Activities: A Survey,» *IEEE Transactions on circuits and systems for Video Technology*, vol. 18, nº 11, pp. 1473-1488, 2008.

[19] J. Kooij, M. Liem, J. Krijnders, T. Andringa y D. Gavrilaa, «Multi-modal human aggression detection,» *Computer Vision and Image Understanding*, 2015.

[20] X. Fang, Z. Xia, C. Su, T. Xu, Y. Tian, Y. Wang y T. Huang, «A system based on sequence learning for event detection in surveillance video,» *ICIP*, pp. 3587-3591, 2013.

[21] D. Weinland, R. Ronfard y E. Boyer, «Free viewpoint action recognition using motion history volumes,» *Computer Vision and Image Understanding*, nº 104, pp. 249-257, 2006.

[22] N. R. Jennings, «On agent-based software engineering,» *Artificial Intelligence*, nº 117, p. 277–296, 2000.

[23] Controles Inteligentes, «Sistema de Guiado Vehicular,» [En línea]. Available: http://www.ci24.com/automatizacion-de-parqueaderos/sistema-de-guiado-con-camaras/. [Último

acceso: 12 noviembre 2018].

[24] E. González y M. Torres, «AOPOA: Organizational Approach for Agent Oriented Programming,» *Proceedings of the Eighth International Conference on Enterprise Information Systems,* 2006.

[25] F. Á. Bravo, «Interactive Robotics Drama for Educational Purposes,» *Ph.D. Research Proposal,* 2017.

[26] IHS Market, «Top Video Surveillance Trends for 2016,» 2016. [En línea]. Available: https://technology.ihs.com/api/binary/572252. [Último acceso: 12 noviembre 2018].

[27] D. Barret, «One surveillance camera for every 11 people in Britain, says CCTV survey,» 10 julio 2013. [En línea]. Available: http://w3.salemstate.edu/~pglasser/One_surveillance_camera_for_every_11_people_in_Britain__says_CCTV_survey_-_Telegraph.pdf. [Último acceso: 11 noviembre 2018].

[28] A. Cuevas, «Peñalosa anuncia la instalación de 1.500 cámaras nuevas de seguridad en toda Bogotá,» *Alcaldía Mayor de Bogotá,* 14 febrero 2017.

[29] Y. Kong y Y. Fu, «Human Action Recognition and Prediction: A Survey,» *JOURNAL OF LATEX CLASS FILES,* vol. 13, nº 9, 2018.

[30] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara y N. Tishby, «Detecting anomalies in people's trajectories using spectral graph analysis,» *Computer Vision and Image Understanding,* pp. 1099-1111, 2011.

[31] Y. Du, Y. Fu y L. Wang, «Skeleton Based Action Recognition with Convolutional Neural Network,» *3rd IAPR Asian Conference on Pattern Recognition,* 2015.

[32] E. Gedat, P. Fechner, R. Fiebelkorn y R. Vandenhouten, «Human Action Recognition with Hidden Markov Models and Neural Network Derived Poses,» *IEEE 15th International Symposium on Intelligent Systems and Informatics,* pp. 157-162, 2017.

[33] Fauziah, E. Wibowo, S. Madenda y Hustinawati, «Identification of hand motion using background subtraction method and extraction of image binary with backpropagation neural network on skeleton model,» *2nd International Conference on Computing and Applied Informatics* , nº 978, pp. 1-14, 2017.

[34] P. KaewTraKulPong y R. Bowden, «An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection,» *2nd European Workshop on Advanced Video Based*

Surveillance Systems, nº 02, 2001.

[35] T. Chateau, A. Vacavant y J. M. Lavest, «Skin detection and tracking by monocular vision,» *IEEE Xplore,* 2015.

[36] OpenCV, «About,» 2018. [En línea]. Available: https://opencv.org/about.html. [Último acceso: 11 noviembre 2018].

[37] A. Magar y J. Shinde, «A New Approach of Human Segmentation from Photo Images,» *International Journal of Scientific and Research Publications,* vol. 5, nº 1, 2015 .

[38] K. Kim, T. H. Chalidabhongse, D. Harwood y L. Davis, «Real-time foreground–background segmentation using codebook model,» *Real-Time Imaging,* vol. 11, nº 3, pp. 172-185, 2005.

[39] P. Viola y M. Jones, «Rapid Object Detection Using a Boosted Cascade of Simple Features,» *COMPUTER VISION AND PATTERN RECOGNITION,* 2001.

[40] Y. Peng, M. Xu, J. S. Jin, S. Luio y G. Zhao, «Cascade-based License Plate Localization with Line Segment Features and Haar-like Features,» *Sixth International Conference on Image and Graphics,* pp. 1023-1028, 2011.

[41] M. Siala, N. Khlifa, F. Bremond y K. Hamrouni, «People detection in complex scene using a cascade of Boosted classifiers based on Haar-like-features,» 2009.

[42] S. Zeevi, «The BackgroundSubtractorCNT project (CNT stands for 'CouNT),» 2016. [En línea]. Available: https://github.com/sagi-z/BackgroundSubtractorCNT. [Último acceso: 24 noviembre 2018].

[43] A. Angelova, A. Krizhevsky, A. Ogale y D. Ferguson, «Real-Time Pedestrian Detection With Deep Network Cascades,» p. 12, 2015.

[44] Microsoft, «Kinect for Windows SDK 2.0,» 20 octubre 2014. [En línea]. Available: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn782025(v%3dieb.10). [Último acceso: 11 noviembre 2018].

[45] T.-L. Le, M.-Q. Nguyen y T.-T.-M. Nguyen, «Human posture recognition using human skeleton provided by Kinect,» pp. 340-345, 2013.

[46] S. Qiao, Y. Wang y J. Li, «Real-Time Human Gesture Grading Based on OpenPose,» *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics,* 2017.

[47] X. Li, K. Wang, W. Wang y Y. Li, «A Multiple Object Tracking Method Using Kalman Filter,» *IEEE International Conference on Information and Automation,* 2010.

[48] M. Mirabi y S. Javadi, «People Tracking in Outdoor Environment Using Kalman Filter,» *Third International Conference on Intelligent Systems Modelling and Simulation,* pp. 303-307, 2012.

[49] Y. Cai y M. Pietikainen, «Person Re-identification Based on Global Color Context,» *Computer Vision – ACCV Workshops,* pp. 205-215, 2010.

[50] K. Jang, S. Han y I. Kim, «Person Re-identification Based on Color Histogram and Spatial Configuration of Dominant Color Regions,» 2014.

[51] K.-E. Aziz, D. Merad y B. Fertil, «People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation,» 2011.

[52] A. A. Chaaraoui, P. Climent-Pérez y F. Flórez-Revuelta, «Silhouette-based human action recognition using sequences of key poses,» *Pattern Recognition Letters,* nº 34, pp. 1799-1807, 2013.

[53] Mestreacasa, «Transformaciones geométricas - Proyectividad y homografía, homología y afinidad, inversión,» mestreacasa.gva.es, [En línea]. Available: http://mestreacasa.gva.es/c/document_library/get_file?folderId=500006275129&name=DLFE-1000233.pdf. [Último acceso: 11 noviembre 2018].

[54] R. Eshel y Y. Moses, «Homography Based Multiple Camera Detection and Tracking of People in a Dense Crowd,» *EEE Conference on Computer Vision and Pattern Recognition,* 2008.

[55] S. M. Khan y M. Shah, «A Multiview Approach to Tracking People in Crowded Scenes using a Planar Homography Constraint,» *European Conference on Computer Vision,* 2006.

[56] Z. Zhu, V. Branzoi, M. Sizintsev, N. Vitovitch y T. Oskiper, «AR-Weapon: Live Augmented Reality based First-Person Shooting System,» *IEEE's and PAMITC's premier meeting on applications of computer vision,* 2015.

[57] L. L. Ng y H. S. Chua, «Vision-Based Activities Recognition by Trajectory Analysis for Parking Lot Surveillance,» *IEEE International Conference on Circuits and Systems (ICCAS),* pp. 137-

142, 2012.

[58] L. Hao-zhe, H. Kui-hua y L. Guo-hui, «A surveillance activity recognition model based on Hidden Markov Model,» *International Conference on Automatic Control and Artificial Intelligence*, pp. 305-308, 2012.

[59] P. Ballester y R. Matsumura, «On the Performance of GoogLeNet and AlexNet Applied to Sketches,» *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1124-1128, 2016.

[60] ChaLearn, «ChaLearn Looking at People,» [En línea]. Available: http://gesture.chalearn.org/mmdata#Track2. [Último acceso: 11 noviembre 2018].

[61] L. Xia, C.-C. Chen y J. K. Aggarwal, «View Invariant Human Action Recognition Using Histograms of 3D Joints,» *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012.

[62] Y. Liu, L. Nie, L. Liu y D. S. Rosenblum, «From action to activity: Sensor-based activity recognition,» *Neurocomputing*, vol. 191, pp. 108-115, 2016.

[63] X. Wang, «Intelligent multi-camera video surveillance: A review,» *Pattern Recognition Letters*, 2012.

[64] P. Remagnino, A. Shihab y G. Jones, «Distributed intelligence for multi-camera visual surveillance,» *Pattern Recognition*, p. 14, 2003.

[65] M. Panda y M. R. Patra, «Ensemble Voting System for Anomaly Based Network Intrusion Detection,» *International Journal of Recent Trends in Engineering*, vol. 2, nº 5, pp. 8-13, 2009.

[66] Foscam, «IP Camera CGI V1.27,» [En línea]. Available: https://www.foscam.es/descarga/ipcam_cgi_sdk.pdf. [Último acceso: 11 noviembre 2018].

[67] P. Stone y M. Veloso, «Multiagent Systems: A Survey from a Machine Learning Perspective,» *Autonomous Robots*, vol. 8, nº 3, p. 345–383, 2000.

[68] G. Caire, «JADE Tutorial - JADE programming for beginners,» 30 junio 2009. [En línea]. Available: http://jade.tilab.com/doc/tutorials/JADEProgramming-Tutorial-for-beginners.pdf. [Último acceso: 11 noviembre 2018].

[69] J. Rodríguez, M. Torres y E. González, «The AOPOA Methodology,» *Revista Avances en Sistemas e Informática,*, vol. 4, nº 2, 2007.

[70] Network Time Foundation, «NTP Project,» [En línea]. Available: https://www.nwtime.org/projects/ntp/. [Último acceso: 11 noviembre 2018].

[71] J. Kangasharju, «Chapter 3: Distributed Systems: Synchronization,» 2013. [En línea]. Available: https://www.cs.helsinki.fi/webfm_send/1232. [Último acceso: 11 noviembre 2018].

[72] Iowa State University, «Election Algorithms,» [En línea]. Available: http://web.cs.iastate.edu/~cs554/NOTES/Ch6-Election.pdf. [Último acceso: 11 noviembre 2018].

[73] R. Al-Tayeb, «Motion JPEG Streaming Server,» [En línea]. Available: https://www.codeproject.com/Articles/371955/Motion-JPEG-Streaming-Server. [Último acceso: 12 noviembre 2018].

[74] Janajk, «MJPEG Streaming Utilities for Python 3.x,» GitHub, [En línea]. Available: https://github.com/janakj/py-mjpeg/. [Último acceso: 12 noviembre 2018].

[75] CMU-Perceptual-Computing-Lab, «OpenPose 1.1.0 benchmark,» [En línea]. Available: https://docs.google.com/spreadsheets/d/1-DynFGvoScvfWDA1P4jDInCkbD4lg0IKOYbXgEq0sK0/edit#gid=0. [Último acceso: 12 noviembre 2018].

[76] New Mexico Tech Computer Center, «The hue-saturation-value (HSV) color model,» [En línea]. Available: http://infohost.nmt.edu/tcc/help/pubs/colortheory/web/hsv.html. [Último acceso: 11 noviembre 2018].

[77] L. Ercolanelli, «Color Extractor,» [En línea]. Available: https://github.com/algolia/color-extractor. [Último acceso: 11 noviembre 2018].

[78] G. M. Johnson y M. D. Fairchild, «A Top Down Description of S-CIELAB and CIEDE2000,» *Wiley InterScience,* vol. 28, nº 6, pp. 425-435, 2003.

[79] E. Cippitelli, E. Gambi, S. Spinsante y F. Flórez-Revuelta, «Evaluation of a skeleton-based method for human activity recognition on a large-scale RGB-D dataset,» *2nd IET International Conference on Technologies for Active and Assisted Living,* 2016.

[80] J.-G. Ko y J.-H. Yoo, «Rectified Trajectory Analysis based Abnormal Loitering Detection for Video Surveillance,» *First International Conference on Artificial Intelligence, Modelling &*

Simulation, pp. 254-258, 2013.

[81] W. Dong, N. Zhou, J.-C. Paul y X. Zhang, «Optimized Image Resizing Using Seam Carving and Scaling,» *ACM Transactions on Graphics, Association for Computing Machinery*, vol. 29, nº 5, p. 10 p, 2009.

[82] J. Pitt, L. Kamara, M. Sergot y A. Artikis, «Voting in Multi-Agent Systems,» *The Computer Journal*, vol. 49, nº 2, pp. 156-170, 2006.

[83] Y. Chen, «Voting Protocols,» 14 septiembre 2011. [En línea]. Available: http://www.eecs.harvard.edu/cs286r/courses/fall11/slides/lec4-w.pdf. [Último acceso: 12 noviembre 2018].

[84] B. Bharathi y S. M. Vaniya, «Multi-Camera Human Activity Recognition in Visual Surveillance — Recent Advances,» *Jour of Adv Research in Dynamical & Control Systems*, vol. 10, nº 9, 2018.

[85] B. Ma, Y. Su y F. Jurie, «BiCov: a novel image representation for person re-identification and face verification,» *British Machive Vision Conference*, 2012.

[86] T. H. Romer, D. Lee, G. M. Voelker, A. Wolman y W. A. Wong, «The Structure and Performance of Interpreters,» *ASPLOS VII Proceedings of the seventh international conference on Architectural support for programming languages and operating systems*, 1996.

[87] Scikit-learn, «Machine Learning in Python,» [En línea]. Available: https://scikit-learn.org/stable/. [Último acceso: 12 noviembre 2018].

[88] Python-colormath, readthedocs.io, [En línea]. Available: https://python-colormath.readthedocs.io/en/latest/. [Último acceso: 12 noviembre 2018].

[89] HMMLearn, «readthedocs.io,» [En línea]. Available: https://hmmlearn.readthedocs.io/en/latest/. [Último acceso: 12 noviembre 2018].

[90] M. Galar, A. Fernández, E. Barrenechea, H. Bustince y F. Herrera, «An overview of ensemble methods for binary classifiers in multi-class,» *Pattern Recognition*, vol. 44, pp. 1761-1776, 2011.

[91] Python Software Foundation, «Python 3.6.0,» 23 diciembre 2016. [En línea]. Available: https://www.python.org/downloads/release/python-360/. [Último acceso: 12 noviembre 2018].

[92] A. Potavov, «SingularityNET,» 14 mayo 2018. [En línea]. Available: https://blog.singularitynet.io/can-deep-networks-learn-invariants-1e06a5052555. [Último acceso: 11 noviembre 2018].

[93] A. Rosebrock, «Real-time object detection on the Raspberry Pi with the Movidius NCS,» 19 febrero 2018. [En línea]. Available: https://www.pyimagesearch.com/2018/02/19/real-time-object-detection-on-the-raspberry-pi-with-the-movidius-ncs/. [Último acceso: 12 noviembre 2018].

[94] D. D. Bloisi, A. Grillo y A. Pennisi, «Multi-modal Background Model Initialization,» *Scene Background Modeling and Initialization Workshop,* 2015.