Universitat d'Alacant
Universidad de Alicante

Arquitectura de visión y aprendizaje para el reconocimiento de actividades de grupos usando descriptores de movimiento

Luis Felipe Borja Borja

Tesis Doctorales
UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

Universitat d'Alacant
Universidad de Alicante

Instituto Universitario de Investigación en Informática

Escuela Politécnica Superior

Arquitectura de visión y aprendizaje para el reconocimiento de actividades de grupos usando descriptores de movimiento

Luis Felipe Borja Borja

Tesis presentada para aspirar al grado de DOCTOR POR LA UNIVERSIDAD DE ALICANTE

DOCTORADO EN INFORMÁTICA

Dirigida por:

Dr. Jorge Azorín López
Dr. Marcelo Saval Calvo

Universitat d'Alacant
Universidad de Alicante

"Per aspera ad astra."

Anónimo

"La ciencia se compone de errores, que a su vez son los pasos hacia la verdad."

Jules Verne

"La ciencia no sabe de países, porque el conocimiento le pertenece a la humanidad y es la antorcha que ilumina al mundo. La ciencia es el alma de la prosperidad de las naciones y la fuente de todo progreso."

Louis Pasteur

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Agradecimientos

A la memoria de mis queridos padres Ángel y Amelia, quienes siempre creyeron en mí sabiendo que la educación es la mejor fuente de respuestas y la luz más brillante que guía el camino hacia el progreso.

A mi amada esposa Paola, con su gran dedicación, esfuerzo y paciencia ha sido el pilar fundamental y mi complemento para la construcción de este sueño. A nuestra querida hija Isabella que pronto estará con nosotros, siendo el motor para superar cualquier adversidad. A mis hermanos, sobrinos y amigos que nunca han dejado de apoyarme para seguir siempre hacia adelante.

Además, quiero expresar un especial agradecimiento a la Universidad Central del Ecuador, a la Universidad de Alicante de España, a mis maestros y amigos el Dr. Jorge Azorín y el Dr. Marcelo Saval, quienes con su sabiduría y paciencia me iniciaron en el mundo de la investigación en Visión por Computador.

Alicante, 11 de marzo de 2020
Luis Felipe Borja Borja

Universitat d'Alacant
Universidad de Alicante

Resumen

Según los últimos censos, nuestro planeta tiene cerca de 7.000 millones de habitantes principalmente concentrados en zonas urbanas. Consecuencia de esto las multitudes de personas se congregan en estos sitios, complicando la tarea de supervisión y vigilancia para mantener la seguridad pública en calles, plazas, avenidas y demás. Esto motiva el estudio y mejora de métodos de análisis automático del comportamiento humano. A esta área de investigación se le denomina Análisis del Comportamiento Humano, o Reconocimiento de Actividades Humanas. Gran parte de los trabajos dedicados a este problema se basan en técnicas de visión por computador junto con algoritmos de Machine Learning y, más recientemente, en Deep Learning.

En este proyecto de tesis, se ha hecho inicialmente una revisión del estado del arte respecto al tema del análisis y reconocimiento de actividades y comportamientos humanos. En este estudio se han analizado los principales trabajos de machine learning tradicional y deep learning para el tema de la tesis, así como los principales datasets. Se ha visto que no existe un estándar o arquitectura que proponga solución genérica. Por otro lado, la mayoría de trabajos se centran en un determinado rango de individuos, habiendo propuestas para personas individuales, para pequeños grupos, grandes grupos o multitudes. Además, no existe un consenso en la nomenclatura respecto a los grados de complejidad, niveles de comportamiento o, como aquí se denomina, nivel de semántica de las acciones que se realizan. Tras este estudio, se ha propuesto una taxonomía bidimensional que permite clasificar las propuestas en el espacio "número de personas/nivel de semántica", siendo más descriptivo respecto al actual estado del arte y

Universitat d'Alacant
Universidad de Alicante

permitiendo ver donde se concentran mayormente los trabajos y cuales los retos aun no resueltos.

Tras el estudio del estado del arte, en este trabajo se ha propuesto una arquitectura de visión y aprendizaje para reconocer actividades de grupos usando descriptores de movimiento. Se compone de dos bloques principales, el descriptor de movimiento y el clasificador de actividad. Las arquitecturas de red profunda que se estudian actualmente tienen la bondad de, dados unos datos en crudo (imágenes, secuencias, etc.) tratarlos internamente de forma que devuelvan un resultado, sin necesidad de preprocesarlos primero. Sin embargo, esto los hace dependientes de los datos de entrenamiento y necesitan grandes datasets para que el entrenamiento sea suficiente. El hecho de introducir un descriptor hace que el espacio de búsqueda se reduzca, y por lo tanto se pueda entrenar con menor número de datos, y además, se pueda independizar la escena (número de individuos, localización de la actividad en el espacio, etc.) del comportamiento en sí. Para el descriptor de la arquitectura se propone en esta tesis como una variante del descriptor Activity Descriptor Vector (ADV), que se denomina D-ADV, y que obtiene dos imágenes del movimiento local acumulado, una UDF (de los movimientos arriba, Up, abajo, Down, y Frecuencia) y otra LRF (de los movimientos Left, izquierda, Right, derecha y Frecuencia).

Por otro lado, como instancias de la arquitectura haciendo uso del D-ADV, se proponen el D-ADV-MultiClass para clasificación de múltiples clases. Esta propuesta se basa en utilizar los dos streams UDF y LRF, junto con una red profunda y transfer learning, para reconocer la actividad del grupo. Además, se ha propuesto otra instancia, llamada D-ADV-OneClass, que añade a los dos streams anteriores, otro con información de contexto. Esta última instancia da solución a problemas en los que solo se conoce una clase durante el entrenamiento, y por lo tanto se utilizan técnicas de one-class classification. En la experimentación se ha validado la arquitectura con las dos instancias D-ADV-MultiClass y D-ADV-OneClass utilizando los datasets públicos ampliamente conocidos, como son BEHAVE, INRIA y CAVIAR para multi-class, y para one-class los datasets Ped 1, Ped 2 y Avenue. Los resultados experimentales muestran la capacidad de la arquitectura para clasificar las actividades de los grupos presentados en los

datasets. Además, se demuestra que la arquitectura es capaz de tener buenos resultados utilizando datasets con poca cantidad de datos. En este caso, no a partir de la imagen sino de la representación del movimiento. Por último se plantean como trabajos futuros experimentar con otros datasets de mayor tamaño o con otro tipo de datos (peleas callejeras y en rings de boxeo para ver como afecta el contexto en estas situaciones). A medio o largo plazo se realizarán mejoras aumentando y comprobando otras instancias de la arquitectura utilizando múltiples streams de entrada que puedan permitir detectar otros comportamientos.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice general

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice de figuras

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice de tablas

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 1

Introducción

Este primer capítulo introductorio resume el planteamiento del trabajo de tesis doctoral incluyendo el contexto y la motivación en la que se enmarca el trabajo, un estado del arte general del área, los objetivos del trabajo y finalmente una propuesta genérica de solución que se instanciará en capítulos posteriores. La propuesta para solucionar el problema de detección automática de acciones de grupos de personas se plantea como una arquitectura capaz de analizar el flujo de movimiento de grupos, y clasifica la actividad. Se plantea, además, dos casos de clasificación, una clasificación binaria (normal y anormal) llevada a cabo mediante One Class Classification (OCC) y una clasificación multi-clase con una propuesta Multi-Class Classification (MCC). Además, hace uso del descriptor Activity Description Vector (ADV), que consiste en un método de representación del movimiento, y lo representa mediante un vector de características (Up, Down, Left, Right, Frecuency).

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

1.1. Motivación y Contexto

Según algunos autores la motivación es el énfasis que se descubre en una persona para lograr un objetivo que satisface alguna necesidad. En mi caso personal, el motivo de investigar en el área de Visión por Computador viene del interés por comprender el complejo proceso que representa darle un sentido lógico al entorno gracias a los rayos de luz que penetran en una cámara, simulando la visión humana y el cerebro. Además, enfrentarme a problemas todavía no resueltos en este campo, es otro factor motivante para desarrollar este proyecto de investigación.

Entender y emular algunas de las características del sentido de la vista mediante el uso de hardware y software es un reto gigantesco, el cual ha llevado muchas décadas de investigación, y donde aún existen muchos problemas todavía no resueltos.

De esta investigación se esperan resultados a distintos niveles que deben ser difundidos en congresos y revistas especializadas en el tema, además de beneficiar a la sociedad con el conocimiento que se ha generado. Por otro lado, a nivel personal, obtener satisfacción del quehacer científico aportando soluciones a los problemas abiertos de Análisis de Comportamiento Humano (HBA del inglés Human Behaviour Analysis), es otro de los motivos que me ha empujado a realizar esta este trabajo de investigación. Además, han sido varios retos los que se han debido superar para llegar a esta etapa que finalizaría con la presentación mi tesis doctoral, siendo un inicio de una nueva carrera como investigador en el área de Visión por Computador.

Este marco de trabajo y motivación personal, y gracias a un convenio entre la Universidad Central de Ecuador y la Universidad de Alicante, pude incorporarme al grupo de investigación de mis directores Jorge Azorín y Marcelo Saval, que tiene larga experiencia en visión por computador y aprendizaje máquina. El grupo ha tenido y varios proyectos de investigación con financiación pública centrados en visión por computador y decenas de publicaciones en revistas y congresos internacionales, también en temáticas de inteligencia artificial.

Actualmente, en el planeta tierra vivimos aproximadamente 7.000 mi-

llones de habitantes. El rápido crecimiento de la población mundial observado en el último siglo, en particular en las zonas urbanas, genera situaciones de congestión que se han convertido en una parte de la vida cotidiana y representa un gran reto para mantener la seguridad pública y para el diseño de trazados urbanos.

En las últimas décadas, la vigilancia de las personas y grupos mediante el uso de los sistemas de video vigilancia ha aumentado en todo el mundo y se ha convertido en una herramienta ampliamente utilizada para monitoreo y control del comportamiento humano. Gracias a ello, el estudio del comportamiento de grupos y multitudes mediante video vigilancia se ha convertido en una necesidad para la sociedad, ya que puede ayudar en muchos problemas comentados anteriormente.

En la mayoría de los sistemas tradicionales de video vigilancia existentes se requiere la presencia de un operador para analizar, detectar o alertar acerca de cualquier contenido de interés que puede darse [220, 226, 232, 272, 281, 398]. Conociendo las limitaciones actuales, es necesario diseñar sistemas de video vigilancia que permitan la identificación automática del comportamiento humano con mayor fiabilidad. Este tipo de sistemas se pueden realizar mediante técnicas de visión por computador junto con métodos de aprendizaje automático, ya que permiten identificar patrones, que en este caso serán de comportamiento de personas de forma no supervisada, tales como acciones, actividades, trayectorias, entre otros.

En términos generales, los investigadores han hecho un gran esfuerzo para modelar el comportamiento de individuos, grupos y multitudes en espacios abiertos o cerrados como universidades, centros comerciales, parques o calles, y luego analizarlos utilizando métodos de Aprendizaje Automático (ML del inglés Machine Learning),[86, 87, 161, 240, 292, 301, 329, 335, 376, 408].

Dado este contexto general, este proyecto abordará el problema del HBA mediante técnicas de visión por computador, ML y DL . Más concretamente, y como se describirá después en detalle, se estudia reconocimiento de actividades de grupo usando descriptores de movimientos locales utilizando técnicas de Deep Learning (DL).

Centrándonos en el área en la que se enmarca este trabajo, existen

muchos problemas abiertos en el área de visión por computador (Computer Vision en inglés). La investigación en esta área aborda problemas tanto con datos en dos dimensiones (una imagen o secuencia de imágenes tipo RGB), hasta el tratamiento de datos tridimensionales como imágenes RGBD. Aquí se trata el análisis de secuencia de imágenes enfocado al Análisis de Comportamiento Humano (HBA) [3, 78, 267] de grupos de personas. En general, el proceso para el HBA detectado en secuencias de imágenes se suele dividir en tres partes, como se muestra en la figura 1.1:

1. **Adquisición:** fase en la que se obtienen los datos mediante cámaras y estos son pre-procesados para eliminar ruido u otros factores que puedan tener las imágenes. En esta fase suelen utilizarse métodos para filtrado de ruido, filtros de mediana, o correcciones de la imagen como distorsiones o alineamientos.

2. **Procesado:** en esta etapa se toman las imágenes ya tratadas para extraer características que permitan, posteriormente, analizar la escena. Aquí se pueden incluir técnicas de segmentación para aislar el elemento relevante en la escena, o técnicas de flujo óptico para determinar la dirección del movimiento en la escena, etc.

3. **Análisis:** en esta última fase se utilizan diferentes técnicas para extraer características del Comportamiento Humano en base a los datos procesados anteriormente. Esta fase incluye la aplicación combinada o individual de diversas técnicas, desde métodos clásicos como Support Vector Machines (SVM), hasta las actuales técnicas de aprendizaje profundo, como las Redes Neuronales Convolucionales (CNNs). Cabe destacar que con DL no es siempre necesario pre-procesar los datos.

En el estudio de HBA existen muchas sub-áreas o problemas por resolver: reconocer acciones, actividades o comportamiento de una persona, grupo o multitud en función de variables como el movimiento, tiempo de duración y contexto; por ejemplo, saber si esta persona está corriendo, caminando, o agachándose. Además, se pueden analizar actividades en las

Universitat d'Alacant
Universidad de Alicante

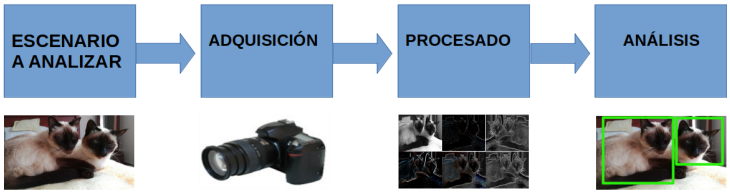


Figura 1.1: La figura muestra las tres etapas de procesamiento de imágenes en una escena (adquisición, procesado y análisis).

que interactúan personas con objetos como cocinar, comprar, ver la televisión [34, 65, 106]; otra sub-área de HBA es identificación de gestos del rostro de una persona [85, 144, 144, 319]; el reconocimiento de lenguaje de señas o gestos con la mano [75, 98, 157, 190]; y muchas más. Cada una de estas sub-áreas de investigación incluye innumerables problemas que requieren propuestas de soluciones específicas. En este proyecto de tesis doctoral me centro en el análisis de acciones en grupos de personas mediante una arquitectura que permite clasificar utilizando características del movimiento generado por las personas que aparecen en una escena.

1.1.1. Inteligencia Artificial(AI)

Puesto que esta tesis tiene como núcleo una propuesta arquitectural basada el técnicas de inteligencia artificial, en esta subsección se explica de forma general la evolución y principales aspectos de la misma respecto al aprendizaje máquina. Como parte de la fundamentación teórica de esta investigación definiremos brevemente el término Inteligencia Artificial (IA) y otros temas asociados. En la década de los años 60 el matemático inglés Alan Turing propuso el test de Turing, que consiste en una prueba para medir la habilidad que posee una máquina para simular un comportamiento inteligente parecido al de una persona. El autor realizó la siguiente pregunta: ¿Las máquinas pueden pensar? [363]. El objetivo principal de la IA es proporcionar métodos que permitan a las máquinas actuar de distintas formas en función de la información que reciben como entradas. La IA incluye multitud de áreas, como Computer Vision, Machine Learning, Natural Language Processing. Si nos centramos solo en el área de Machine

Learning, la Figura 1.2 muestra la relación entre IA, ML y DL de acuerdo a contenido y evolución desde la década de los años 1960 [17].

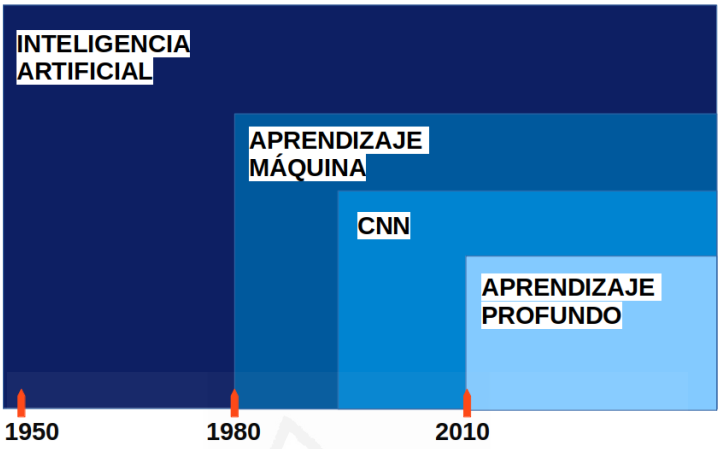


Figura 1.2: La figura muestra la evolución de los diferentes métodos de inteligencia artificial, desde el Aprendizaje Máquina hasta el Deep Learning.

El Aprendizaje Máquina (ML del inglés Machine Learning) es una subárea de la Inteligencia Artificial (IA) que incluye métodos y algoritmos con la capacidad de aprender a partir de ejemplos. Estos métodos y algoritmos pueden generar modelos de problemas complejos en circunstancias específicas, teniendo la capacidad de generalizar o adaptarse a nuevas situaciones para responder en forma eficiente. Un ejemplo clásico de uso de Aprendizaje Máquina es diferenciar en una imagen entre motocicletas y aviones, luego de haber aprendido con muchos ejemplos, qué es una motocicleta o un avión, el computador puede identificar en una imagen que no ha procesado antes si lo que se ve es una motocicleta o un avión, sobre de este ejemplo existen algunos trabajos en donde se muestra en la práctica el proceso antes descrito [8, 332, 344]. Actualmente el desarrollo del Machine Learning ha permitido solucionar problemas en muchas áreas del conocimiento, como por ejemplo, reconocimiento de patrones [135, 173, 373], clasificación, codificación [182, 353, 359], Internet of the Things (IoT) [81, 168, 236, 388], minería de datos [150, 261, 350], reglas de asociación[76, 179, 287, 387], etc. Sin embargo, pese a todo el avance realizado hasta el momento, aun existen problemas sin resolver, como por ejemplo; el desarrollo de sistemas

eficientes de video vigilancia automática para la detección de anomalías de forma genérica y con independencia del número de personas, que es en lo que se centra esta tesis [276].

El Deep Learning (DL) o Aprendizaje Profundo ejecuta el proceso de ML a través de una Red Neuronal Artificial (NN) compuesta por múltiples capas jerárquicas. En las capas iniciales de esta jerarquía, la red neuronal aprende características sencillas y luego transmite este aprendizaje las siguientes capas. Las capas superiores que son normalmente ocultas, es decir, su entrada y salida quedan dentro de la red, procesan los datos convirtiéndolos en más complejos y lo vuelven a pasar hacia las capas superiores para repetir nuevamente el proceso. Las últimas capas son las que producen la salida esperada, ya sea una clasificación, una codificación o aquello para lo que se haya entrenado [9].

Si aplicamos el proceso de funcionamiento descrito en el párrafo anterior en el ejemplo para identificar motocicletas y aviones; en las capas iniciales de la red neuronal se realizan convoluciones aprendidas durante el proceso de aprendizaje, que permiten obtener características básicas como encontrar bordes o esquinas. La capa inicial pasa esta información hacia la segunda capa, que combina los bordes detectados para construir formas sencillas como líneas, curvas o figuras. La siguiente capa combina las formas simples e identifica cosas más complejas cómo figuras tales como círculos, óvalos o cuadrados. Este proceso se repite muchas veces hasta que se pueda identificar la mayor cantidad de características que permitan a la red completa identificar en forma eficiente dentro de una imagen si encuentra una motocicleta o avión [321].

1.2. Descripción del problema a resolver

Las detección automática del comportamiento humano de grupos en secuencias de video es un problema de Visión por Computador y Machine Learning que tiene todavía muchos retos que los investigadores deben superar. Existen algunos trabajos sobre este tema que se centran en actividades o acciones de corta duración y sin tener en cuenta el tipo de actividad o si esta es normal o anormal.

En esta sección se describe la problemática del HBA con ML/DL, donde se verán las principales áreas de estudio, unas nociones sobre niveles semánticos tratados, y el análisis de anomalías en comportamiento. En el Capítulo 2 se describe con más detalle el estado del arte de la tesis.

1.2.1. Detección de Anomalías

La detección de anomalías es un problema que a diario se presenta en nuestras actividades. Este problema ha sido estudiado durante décadas en diferentes áreas de investigación y sus dominios de aplicación, abordado el tema de forma general y específica. Las soluciones propuestas han sido planteadas para diferentes problemas relacionados con la vida real, siendo útiles para varios problemas del mismo tipo, o generalistas, ha sido específicas para problemas complejos. Las soluciones planteadas tienen ventajas y desventajas, sin embargo aun existen casos en lo que todavía no ha sido posible plantear soluciones debido a la complejidad de ciertos temas.

A continuación se describen varios aspectos relacionados con la detección de anomalías para comprender de qué forma los investigadores han abordado este tema, tanto de manera general, como específica.[56]

¿Qué es una anomalía?

La palabra anomalía puede tener diferentes significados dependiendo del campo de investigación en que se la utilice, en este estudio se aplica a conceptos matemáticos aplicados a la informática y específicamente al campo de Visión por Computador. En este caso estamos analizando comportamientos anómalos de personas a partir de secuencias de imágenes obtenidas con cámaras de video.

Según la aclaración anterior una posible descripción adecuada para el campo de investigación abordado en este trabajo, una anomalía se refiere a un desvío de la norma, incompatibilidad, divergencia, discordancia, de una práctica que sea habitual, es decir, un patrón que no se ajusta al comportamiento esperado. Esto se resume como algo fuera de lo común, o atípico. La detección de anomalías tiene aplicación en muchas situaciones relacionadas con las personas como la detección de ataques cibernéticos, fraudes en tarjetas de crédito, sistemas de seguridad entre otros similares.

Chandola et al. en 2007 [56] plantea varios retos acerca de la definición

y detección de anomalías:

- Definir una región que contenga todos los posibles comportamientos normales es una tarea complicada debido a que el límite entre comportamientos normales y anormales en muchas ocasiones no es preciso, sobre todo si están cerca del límite.

- Cuando los comportamientos anómalos son producto de acciones premeditadas estos suelen tener características muy parecidas a los normales. Hablamos, por ejemplo, del caso de los ataques fraudulentos que son acciones que intentan semejarse al comportamiento normal para pasar desapercibidos.

- Es importante mantener actualizados los dominios de comportamientos normales debido a que en un futuro pueden cambiar y no ser lo suficientemente representativos.

- La definición exacta de anomalía puede cambiar de acuerdo al dominio y área de aplicación. Por ejemplo, lo que en una determinada situación es normal en otra puede considerarse anormal. Una pelea en la calle es una actividad anormal, mientras que una pelea en un ring de boxeo es normal.

- Para un entrenamiento adecuado de los modelos deberían existir los suficientes datos etiquetados para entrenamiento y validación. Esta posibilidad es difícil de hacerse realidad, en la mayoría de casos se tienen muchos más datos de actividades normales que anormales. Este problema motiva, como veremos posteriormente, al uso de técnicas llamadas **one-class**.

- Los datos utilizados para probar los modelos pueden contener ruido, por lo que a veces se confunden con anomalías reales, esto puede generar errores en el proceso de seleccionar y eliminar los datos correctos.

Como resultado, el trabajo [56] presenta una clasificación técnicas y aplicaciones desarrolladas para tratar el tema de anomalías mediante ML:

- **Técnicas:** basadas en clasificación, basadas el clusterización, basadas en el vecino más cercano, estadísticas, información teórica, espectrales.

- **Aplicaciones:** detección de ciber-intrusiones, detección de fraudes, detección de de anomalías médicas, detección de daños industriales, procesamiento de imágenes, detección de textos anómalos, sensores de red.

En 2019 el trabajo [196], presentó un estudio acerca de las técnicas basadas en Deep Learning desarrolladas para el mismo fin que el caso de las anteriores. Como resumen del estudio realizado se tiene la siguiente clasificación:

- **Métodos:** supervisados, no supervisados, modelos híbridos, redes neuronales one-class.

- **Aplicaciones:** detección de fraudes, detección de ciber-intrusos, detección de anomalías médicas, sensores de detección de anomalías en redes, internet de las cosas(IoT), detección de anomalías en big-data, video vigilancia, detección de daños industriales.

Universitat d'Alacant
Universidad de Alicante

En este estudio sobre el tema también se plantean varios retos sobre las técnicas para detectar anomalías con redes profundas:

- El rendimiento de algoritmos tradicionales no es óptimo cuando se analiza imágenes. Siendo las técnicas de DL las que mejores resultados consiguen hasta ese momento.

- La detección de datos anómalos a gran escala se dificulta a medida que el tamaño de datos crece, debido a que se requiere mayor capacidad de cómputo, o es necesario ajustar el entrenamiento para más clases de salida.

- Las técnicas de detección de anomalías profundas DAD, (del inglés Deep Anomaly Detection) aprenden características discriminatorias jerárquicas de los datos, por lo que ya no es necesario la intervención manual para la definición de dominios.

- El límite entre comportamiento normal y anormal no es preciso en muchos dominios y están en constante evolución. La ausencia de límites bien definidos es un reto para los algoritmos tradicionales y los basados en aprendizaje profundo.

Aspectos del problema de Detección de Anomalías

Antes habíamos definido una anomalía como una desviación o valor atípico de un conjunto de datos que están plenamente identificados, tanto para normal o anormal. Sin embargo, puede existir casos en los que aparezcan datos que no han sido definidos, es decir, con diferentes características a los que ya se conocen. A este tipo de datos se les llama novedades. A estos se les debe considerar como un nuevo patrón y no se consideran como datos anómalos, sino que se aplican al modelo de datos regular. Se puede asignar una puntuación de novedad para estos puntos de datos nunca antes detectados, utilizando una puntuación de umbral de decisión [52]. Cuando el valor de puntuación alcanza dicho umbral, este pasa a ser un dato anormal.

Los autores de los trabajos [52] y [56] han abordado el tema del tratamiento de anomalías, tomado como referencia estas técnicas propuestas se ha realizado un resumen en la tabla 1.1, en donde se muestra una propuesta de organización de dichas técnicas.

Universitat d'Alacant
Universidad de Alicante

Tabla 1.1: Detección de anomalías basado en Aprendizaje Profundo

1. Naturaleza de los datos	
2. Basado en disponibilidad de etiquetas	2.1 Detección profunda supervisada
	2.2 Detección profunda semi-supervisada
	2.3 Detección profunda no supervisada
3. Basado en el objetivo de entrenamiento	3.1 Modelos híbridos
	3.2 Redes neuronales Una-Clase(OC-NN)
4. Tipos de anomalías	4.1 Puntos de anomalías
	4.2 Detección contextual
	4.3 Detección colectiva o grupal
5. Salidas de las técnicas DAD	5.1 Puntuación de anomalías
	5.2 Etiquetas

- Naturaleza de los datos de entrada: Los datos de entrada contienen un número de características o atributos que se pueden clasificar

en datos de baja o alta dimensión. Las técnicas DAD aprenden las complejas relaciones jerárquicas de características de los datos de entrada. El número de capas utilizadas en técnicas DAD está relacionado por la dimensión de los datos de entrada, las redes más profundas demuestran que producen un mejor rendimiento en datos de alta dimensión[205].

- Basado en la disponibilidad de etiquetas: Las etiquetas de los datos indican si una instancia es normal o anormal. Sin embargo, con el tiempo una etiqueta puede cambiar su valor, es decir un comportamiento considerado normal con el tiempo puede volverse anormal. Los modelos de Detección de Anomalías Profundas (DAD) se pueden clasificar en tres categorías según el grado de disponibilidad de las etiquetas. (1) detección de anomalías profundas supervisada. (2) detección de anomalías profundas semi-supervisadas. (3) detección de anomalías profundas no supervisadas[290].

- Basado en el objetivo de entrenamiento: Según el trabajo [52] existen dos nuevas categorías de técnicas de detección de anomalías profundas (DAD) basadas en los objetivos de entrenamiento empleados: 1) Modelos Híbridos Profundos (DHM). 2) Redes Neuronales basadas en clasificación Una-Clase (OC-NN).

- Tipos de anomalías: Los métodos de detección de anomalías profundas (DAD) han demostrado detectar los tres tipos de anomalías con gran éxito: anomalías puntuales, anomalías contextuales y anomalías colectivas [54].

- Salidas de las técnicas DAD: Un aspecto crítico para un método de detección de anomalías es determinar cuáles son normales o anormales, por lo general las salidas son de etiquetas de tipo binario. La puntuación de anomalías se establece a partir de umbrales, es decir, todo valor que está dentro de un umbral es normal, y todo lo que está fuera es anormal. Otra forma de explicar esto es, cuando encerramos los valores en una esfera y definimos que todo lo que está dentro de la esfera es normal, y todo lo que está fuera, es anormal[306].

Universitat d'Alacant
Universidad de Alicante

Las diferentes técnicas para Detección de Anomalías Profundas (DAD) tienen ventajas y desventajas, es importante entender qué técnica de detección de anomalías es la más adecuada para un determinado problema de detección de anomalías, según los investigadores esta área es muy activa y con el avance en las investigaciones está sujeta a muchos cambios.

Las técnicas de DAD aprendizaje supervisado basadas en la clasificación que se ilustran en la mejores opciones en un escenario que consiste en la misma cantidad de etiquetas tanto para casos normales y anormales. La complejidad de cálculo computacional en técnicas DAD supervisadas es un aspecto importante, especialmente cuando se aplica a un dominio real. Mientras que las técnicas basadas en la clasificación, supervisadas o semi-supervisadas tienen altos tiempos de entrenamiento, las técnicas no supervisadas de DAD están siendo utilizadas, considerando que su principal ventaja sería ahorrar tiempo, especialmente en la adquisición de etiquetas.

Cada una de las técnicas de Detección de Anomalías Profundas (DAD) discutidas en las secciones anteriores tiene sus fortalezas y debilidades únicas. Es fundamental comprender qué técnica de detección de anomalías es la más adecuada para un determinado contexto de problema de detección de anomalías. Dado que el DAD es un área de investigación activa, no es factible proporcionar tal comprensión para cada problema de detección de anomalías. Por lo tanto, en esta sección, analizamos las fortalezas y debilidades relativas de las diferentes categorías de técnicas para unos pocos problemas simples.

Adicionalmente existen modelos híbridos que extraen características robustas en las capas ocultas de la red neuronal profunda y se alimentan de los algoritmos clásicos de detección de anomalías más eficaces. Las Redes Neuronales basadas en falsificación One-class (OC-NN) son capaces de combinar redes profundas para extraer características de los datos para separar todos los puntos de datos normales y anormales.

Como se puede notar en un periodo de diez años de investigación, iniciando con [56], hasta [52], algunos retos siguen vigentes, como el problema de definir un límite más preciso entre un comportamiento normal y anormal. Además aparecen nuevos retos respecto al tratamiento de imágenes, ya que cada día tenemos más videos e imágenes con detalles adicionales, es

decir, con mejor resolución, y por lo tanto, con mayor necesidad de procesamiento. Adicionalmente se tiene un incremento constante en la cantidad de datos recopilados, lo que provoca dificultades para analizar grandes cantidades de información en la que se deben identificar los comportamientos.

1.3. Objetivos

Actualmente existen numerosas propuestas que abordan el tema del HBA tanto desde la perspectiva tradicional del ML como el actual DL. Sin embargo, aun existen retos que necesitar resolverse por lo que es necesario llevar a cabo nuevas propuestas. Es por ello que en este proyecto de tesis doctoral planteo una arquitectura para el reconocimiento de actividades genérica que posteriormente se instanciará con redes DL, pero que permitiría utilizar otras técnicas.

Para abordar el objetivo acerca de detección de comportamiento humano, en este trabajo se ha planteado una propuesta genérica que permita abordar el problema desde el caso de una clasificación Multi-clase con técnicas Multi-Class Classification (MCC); y también desde el punto de vista de One-Class Classification (OCC) en las que solo se disponga de una clase "normalz no de muestras de lo que se consideraría anormal. Estos términos se explicarán en detalle en secciones posteriores.

Dada la situación actual que se tiene del conocimiento en el ámbito del HBA, se pueden extraer varias conclusiones generales: no existen métodos desarrollados que analicen de forma general y automática el comportamiento humano independientemente del número de personas en la escena; actualmente el estudio de grupos y multitudes está en auge, aunque aún hay grandes problemas por resolver; a pesar de que varios autores han propuesto diferentes clasificaciones de los niveles de comportamiento, no existe una clara diferenciación entre ellos; por último, concluir que no hay trabajos que tengan en cuenta, a parte del movimiento o trayectoria de los individuos, y el contexto como variables de selección.

Por lo tanto, la hipótesis de partida de este trabajo es que se puede mejorar la clasificación de comportamientos uniendo al análisis de los individuos el contexto en el que se encuentran. Además, el uso de descriptores

Universitat d'Alacant
Universidad de Alicante

de movimiento o flujo permitirá independizar el sistema del número de personas en la escena. De esta forma, se define el objetivo principal de este trabajo como la **propuesta de una arquitectura para el reconocimiento de actividades de grupo usando descriptores de movimientos locales.** Se han definido los siguientes objetivos para este este proyecto de tesis doctoral:

- Estudiar los trabajos existentes acerca de Análisis de Comportamiento Humano (HBA), Machine Learning (ML) y Deep Learning (DL) para conocer el estado actual del problema que aquí se aborda y proponer soluciones a problemas no resueltos.
- Proponer una solución genérica basada en técnicas de aprendizaje (DL/ML) para la clasificación de actividades utilizando descriptores de movimientos locales. Además, esta propuesta permite otras entradas de información para robustecer la clasificación, como podría ser el contexto.
- Validar la propuesta tanto con la estrategia One-Class Clasification como con datos Multi-Clase. Además, los resultados se compararán con propuestas del estado del arte para comprobar la eficacia del sistema.

Universitat d'Alacant
Universidad de Alicante

1.4. Propuesta de solución

En este trabajo se va a proponer una arquitectura de visión y aprendizaje para el reconocimiento de actividades de grupos usando descriptores de movimiento. Como se ha visto, la hipótesis de partida es que analizando el movimiento de los individuos mediante movimientos locales podemos mejorar la clasificación e independizar el número de personas de la escena. Además, alimentar métodos de machine learning o deep learning con descriptores reduce el espacio de búsqueda y nos permite trabajar con conjuntos de datos más pequeños. La idea principal es definir una arquitectura basada en técnicas de machine learning y un descriptor de movimiento, junto con otros posibles datos como el contexto, para dar solución problemas sobre el comportamiento y actividades humanas. Por otro lado, se ha visto

que otro problema es el desequilibrio que hay en los datos en la mayoría de casos, habiendo un mayor número de casos normales que anormales. Por eso, la propuesta, se plantea de forma genérica para que permita su implementación con la metodología One-class classification, es decir, entrenar una con una única clase, y o con múltiples clases mediante Multi-class classification. Explicadas en la sección 3.3.

Para entender el comportamiento de un grupo de personas desde el enfoque de este trabajo es necesario observar los movimientos que realizan, el sitio o escena donde lo hacen, y la frecuencia con que se repiten estos comportamientos, como se verá en mayor detalle en el capítulo 3.

1.5. Aportaciones

Como aporte a la investigación en el tema de Visión por Computador y puntualmente en el Análisis del Comportamiento Humano en este documento se han desarrollado los siguientes temas:

Taxonomía del comportamiento humano: En base al análisis de estado del arte se ha resumido y propuesto una clasificación del comportamiento humano para grupos que incluye descripciones del actuar de las personas de acuerdo al grado de semántica, es decir la complejidad que tiene un acto, de menor a mayor grado y el tiempo aproximado que cada una tarda en desarrollarse. Adicionalmente se incluye el número de personas que cada grupo de personas puede tener.

D-ADV: El Vector Descriptor de Actividad Profundo (D-ADV) es un descriptor de movimiento basado en el Vector Descriptor de Actividad (ADV), que analiza y predice comportamientos con un segmentos de frames de una secuencia de imágenes.

Instancia de arquitectura One-Class Classification (OCC): En esta instancia la arquitectura se entrena a través de aprendizaje supervisado con imágenes de comportamiento considerado como normal.

Instancia de arquitectura One-Class Classification (OCC): La arquitectura es instanciada para un entrenamiento supervisado con imágenes de varios comportamientos definidos de grupo.

En resumen la arquitectura propuesta pata detectar comportamientos

definidos dentro de la taxonomía, está conformada con los elementos antes descritos: Un descriptor de movimiento D-ADV, dos instancias de clasificación, una para OCC, y otra para MCC. El proceso de aprendizaje es de tipo supervisado y se han utilizado datasets con pocas imágenes debido a que se aprovecha el conocimiento previo de segmentos de la arquitectura de actividades para entrenar solamente algunas capas definidas, aprovechando de esto modo el concepto de transferencia de aprendizaje en redes neuronales.

1.6. Estructura del documento

Este documento se compone de cuatro capítulos. Este primer capítulo 1 aborda la definición del problema, los objetivos de la tesis, el fundamento teórico utilizado para la propuesta de solución y un resumen de las aportaciones hechas sobre el tema en la publicación de artículos.

El capítulo 2 contiene un análisis del estado del arte sobre el Análisis de Comportamiento y Actividades de grupo de personas donde se incluyen redes neuronales, tanto en Machine Learning como en Deep Learning utilizadas para detección de comportamiento humano; y también se estudian los principales conjuntos de datos que se pueden utilizar en este tipo de investigaciones, así como los retos y conclusiones para el tipo de problema abordado en este trabajo.

En el capítulo 3 se describe la propuesta general de solución con sus diferentes módulos, y una instancia concreta de la propuesta general para multi-class classification (D-ADV-MC) y otra para one-class classification (D-ADV-OC).

En el capítulo 4 se detallan los experimentos realizados, tanto para el D-ADV-MC como para el D-ADV-OC. Además, se exponen los resultados obtenidos.

En el capítulo 5 se presentan las principales conclusiones de este trabajo de tesis doctoral, los trabajos futuros que se pueden realizar a corto, medio y largo plazo sobre el tema, y las publicaciones surgidas de esta investigación.

Universitat d'Alacant
Universidad de Alicante

Capítulo 2

Estado del arte

En este capítulo se hace una revisión del estado del arte del problema del Análisis dede Actividades y Comportamiento Humano. En primer lugar se estudian diferentes propuestas de clasificación del tipo de comportamiento en niveles semánticos desde el más sencillo hasta el más complejo. A continuación, se estudian los trabajos relacionados con técnicas de Machine Learning (ML) y Deep Learning (DL), enfocándose principalmente en los trabajos de DL para el problema concreto de esta tesis. también se hace una revisión de los datasets más importantes utilizados para experimentación tanto en ML como en DL. Por último, se propone una taxonomía bidimensional que permite clasificar los problemas tanto desde la perspectiva del número de personas involucadas en la actividad como el nivel de semántica o complejidad de la acción en términos de tiempo.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

2.1. Análisis de Comportamiento Humano (HBA)

Uno de los factores claves cuando se analiza comportamiento humano es el contexto en donde se desarrollan las actividades, en el caso se ha propuesto una arquitectura multi-stream para la detección de comportamiento de grupos y multitudes. En general las escenas donde se desarrollan las actividades pueden ser en ambientes en interiores o exteriores, en el primer caso se entiende por ambientes que son parte de una edificación, por ejemplo un restaurante, oficina, aula de clase, cocina, otros; en el segundo caso son ambientes en campo abierto, fuera de las edificaciones o al aire libre como por ejemplo un estadio de fútbol, cancha de baloncesto, plaza, parque, playa, centros comerciales, otros.

En ambientes como los antes descritos es en donde generalmente se desarrollan actividades de grupos y multitudes, se analiza el o los comportamientos según el sitio en donde se desarrollan para establecer si son normales o anormales. La idea principal del presente trabajo es desarrollar una arquitectura que combine varios tipos de redes neuronales que en conjunto sean capaces de detectar en forma independiente el tipo de comportamiento de grupos o multitudes en secuencias de video.

En el campo de la investigación de Visión por Computador existen varias áreas como son el reconocimiento facial, análisis de imágenes de diferente naturaleza como: médicas, aéreas, mapas de calor, entre otras, y una que abordamos en esta tesis, el Análisis de Comportamiento Humano o conocido en inglés como Human Behavior Analysis (HBA)[48, 49, 51, 84, 194, 268]. En este campo los investigadores han realizado numerosos trabajos en los que se incluyen como principal herramienta para solucionar problemas de Aprendizaje Máquina y, en los últimos tiempos el, Aprendizaje Profundo. A groso modo, el principal objetivo de este trabajo es reconocer comportamiento humano en una imagen o secuencia de imágenes almacenadas o en línea.

Actualmente existen retos sobre Análisis de Comportamiento Humano utilizando técnicas de Aprendizaje Automático y Visión por Computador, por ejemplo, cómo predecir una pelea callejera, analizar una evacuación

Universitat d'Alacant
Universidad de Alicante

de emergencia en un lugar concurrido, o detectar un comportamiento anómalo de un individuo entre muchas personas. Para abordar este tipo de problemas, los investigadores han usado diversas estrategias como, por ejemplo, realizar una separación de acuerdo al número de personas que aparecen en una escena. Usualmente en las investigaciones se diferencian tres categorías en función del número de personas: individual (una persona), grupal (2 a 50 personas), y multitud (más de 50 personas). Sin embargo, no existe actualmente una definición clara en el área de Visión por Computador de qué se considera grupo o multitud. El estudio de comportamiento humano analizando secuencias de video es un tema con gran trayectoria en investigación, donde existen aportaciones mayormente en comportamientos individuales, centrándose en acciones de corta duración o baja componente semántica, como caminar, saltar, aplaudir y otros similares [32, 46, 256, 289, 364, 364]. En la actualidad, como tema relacionado con el comportamiento humano, existen estudios relacionados con los Ambientes de Vida Asistidos (AAL, del inglés Ambient Assisted Living). En estos casos la investigación se centra en los comportamientos individuales que tengan mayor tiempo de duración o componente semántica como los trabajos de [159, 160, 386, 413], que proponen aplicaciones para ofrecer una mejor calidad de vida monitoreando el comportamiento de personas adultas.

En el caso del estudio de grupos de personas, los avances para analizar comportamiento se limitan a actividades o acciones muy concretas y sencillas, habitualmente, de poco tiempo de duración (baja componente semántica) como una jugada de fútbol [16, 59, 166, 212, 283], jugadas de volleyball [24, 36, 215, 286, 325, 335, 385], grupos de personas caminando en un solo sentido o grupos de personas conversando[12, 63, 86, 376], detección de peatones en una calle [5, 112, 134, 210, 274, 275, 335, 360, 392, 400], violencia entre grupos [13, 133, 201, 213, 348, 365], entre otras. Finalmente, los trabajos relacionados con multitudes se limitan específicamente a tareas como contar personas y calcular densidad de una multitud [38, 177, 204, 340, 341, 403, 414], o detectar movimientos de una masa de personas o colisiones de multitudes [139, 237, 260, 329, 333, 409].

Como se ha descrito en el capítulo 1, esta tesis se enmarca en el pro-

blema del HBA con imágenes o secuencias, y haciendo uso de técnicas de Aprendizaje Profundo. Por ello es fundamental definir una clasificación o taxonomía de los problemas en función de la complejidad que pretenden abordar. En este trabajo se denominará a la complejidad, *nivel de semántica* en el HBA. Este espacio es continuo, ya que no se pueden definir umbrales fijos para dividir los niveles. Sin embargo, sí se definen regiones o áreas en la taxonomía que se propone en la sección 2.2 dividida en bajo(movimiento), medio(acciones y actividades con varios movimientos), y alto(comportamiento que involucra varias acciones continuas).

El nivel de semántica puede clasificarse en función de multitud de parámetros, el tiempo, el significado de la acción, las interacciones entre personas y/o objetos, etc. En este proyecto de tesis doctoral se utilizará el tiempo como parámetro por el cual se definirán los niveles semánticos del comportamiento humano. Por lo general una actividad es considerada compleja si tiene una duración más larga e implique más movimientos o interacción con objetos.

Tras la definición de los niveles en los que se ha estudiado el Análisis del Comportamiento Humano, es fundamental saber las principales técnicas utilizadas, tanto en Aprendizaje Máquina (ML) tradicional como en el actual Aprendizaje Profundo (DL).

Actualmente sigue siendo una utopía superar el funcionamiento del cerebro humano mediante una máquina, sin embargo, los investigadores cada día superan retos relacionados con este complejo proceso. En las últimas décadas se han dado muchos avances en el uso de ML y DL, específicamente desde hace pocos años en el campo de DL. Los dos términos ML y DL son un subconjunto de la IA que fue concebida con el objetivo de las máquinas emulen y de ser posible, superen a los humanos.

El ML en forma resumida se describe como el proceso de usar algoritmos para analizar datos y aprender de ellos, y en lo posterior como producto de este análisis realizar predicciones. Con el paso del tiempo los algoritmos han evolucionado y se han afinado para conseguir mejores resultados. Como resultado de la evolución del ML en los últimos años ha surgido con mucha fuerza la técnica conocida como DL, que es un subconjunto del ML. Su principal característica es que aprende con ejemplos, en

vez de enseñarle al computador reglas para solucionar un problema específico, se le proporciona como entrada algunas instrucciones para ajustar el modelo cuando existan errores, se espera que con la evolución de estas técnicas, se obtengan mejores resultados, ya que una de sus principales ventajas es que tienen la capacidad de extraer características de los datos que analizan. En las sección 2.3 se abordan estos aspectos en profundidad.

2.2. Taxonomías del Análisis de Comportamiento Humano

Para entender el comportamiento humano es necesario clasificarlo de alguna manera en función de algún parámetro que sea de interés para el estudio, la complejidad en base al tiempo en este caso. Deberíamos poder clasificar desde las formas más elementales de comportamiento de una personas que puede durar desde pocos segundos hasta situaciones mucho más complejas con horas e inclusive días de duración. La clasificación de diferentes tipos de comportamientos de personas ha sido motivo de estudio de muchos investigadores dando lugar a distintas definiciones y clasificaciones. Para abordar el problema de HBA intervienen dos variables principales como son el comportamiento, y el grado de semántica.

A continuación describiremos los principales elementos que los investigadores han analizado para clasificar el comportamiento humano, resaltando las principales argumentaciones que se han hecho para proponer, posteriormente, una clasificación adecuada que contribuya y sustente nuestra investigación.

Moeslund et al.[251] definieron una taxonomía de comportamiento humano de menor a mayor grado de abstracción en tres niveles: 1) reconocimiento básico de movimiento al que lo definen como primitivas de acción o motoras que representan entidades atómicas(indivisibles), las cuales son el punto de partida para cualquier otro tipo de comportamiento con un mayor grado de semántica o complejidad. Un ejemplo de estas entidades sería el movimiento de las extremidades superiores o inferiores; 2) en este nivel se incluyen un conjunto de primitivas de acción que se repiten una o varias veces para conformar una acción. Un ejemplo de acción sería el saludo de

dos personas; 3) en este nivel los autores consideran una escala mayor de eventos en la que intervienen contexto, objetos y las primitivas de acción y acciones antes descritas. Un ejemplo de este nivel sería preparar una taza de te y tomarla sentado en una silla frente a una mesa.

En [384] los autores consideran que las actividades pueden caracterizarse por un conjunto de "verbos", i.e, acciones realizadas por un actor humano, y un conjunto de "sustantivos", i.e, los objetos o lugares que son el objetivo de la acción. Es decir, asumen que las actividades son un conjunto de acciones desarrolladas en un contexto interactuando con objetos.

Turaga et al. [362] analizan dos niveles de semántica, denominados como: 1) las acciones y 2) las actividades. Las acciones están representadas por patrones de movimiento simples, por lo general realizadas por una sola persona. Las actividades son más complejas y se componen de acciones coordinadas, realizadas por un grupo pequeño de personas. Adicionalmente también incluyen en su definición las acciones atómicas, que son parte del primer nivel de esta clasificación.

En [251] los autores utilizan una jerarquía de acciones, considerando el nivel de semántica desde la más simple, hasta la más complejo de la siguiente forma: primero están las primitivas de acción, acciones y actividades. Las primitivas de acción o primitivas motoras se utilizarán para las entidades atómicas a partir de las cuales se construyen acciones. Así mismo, las actividades se dividen en acciones.

En[371] los autores describen cuatro niveles relacionados con la semántica para una persona y grupos de personas:

- **Nivel 1 (Gestos):** Movimientos básicos de partes del cuerpo que duran un tiempo muy corto. Ejemplos de gestos pueden ser movimientos de la mano, brazos, pies o cabeza entre otros.

- **Nivel 2 (Acciones):** También llamado atómico, consiste en acciones realizadas por una sola persona, su duración es mayor que un gesto. Un ejemplo de acciones podría ser caminar, correr, saltar.

- **Nivel 3 (Interacción):** En esta categoría se realizan actividades de interacción hombre-hombre u objeto-humano. Ejemplos de estas

interacciones pueden ser dos personas bailando, corriendo una detrás de la otra, niños jugando, gente conduciendo una bicicleta.

- **Nivel 4 (Actividad de grupo):** En este nivel de descripción se ajusta a dos o más grupos de personas, uno o más objetos pueden intervenir en la escena. Una carrera atlética, un equipo de baloncesto, peatones cruzando una calle, un partido de fútbol, una pelea en un estadio pueden ser ejemplos de actividades de grupo.

En[49] se propone otra taxonomía del comportamiento humano que lo clasifica según la complejidad y el tiempo de duración. En este enfoque, el análisis se clasifica según el grado de semántica en cuatro niveles:

- **Nivel 1 (Movimiento):** Detección de movimientos básicos en segundos o frames.
- **Nivel 2 (Acción):** Detección de tareas simples en términos de segundos. El ser humano puede interactuar con objetos, o estar sentado, de pie, caminando.
- **Nivel 3 (Actividad):** Estas son tareas de minutos a horas de duración. Constituyen la secuencia de acciones, como limpiar una habitación, lavar un vehículo.
- **Nivel 4 (Comportamiento):** Este es el nivel más alto de comprensión, ya que su duración puede ser de horas y días. Ejemplo de comportamiento puede ser las rutinas diarias de una persona, la mezcla de dos actividades en secuencia lógica.

La figura 2.1 muestra resumen de cada uno de los niveles de semántica y los ejemplos de cada uno descritos en [49].

Las taxonomías expuestas en los trabajos [49, 251, 251, 362, 371, 384] coinciden en varios aspectos. Como más relevante, es de clasificar el comportamiento humano desde el punto de vista de nivel de semántica en función de su duración.

Adicionalmente todas incluyen en sus definiciones las acciones atómicas, es decir, que no se pueden dividir en otras más sencillas. Las taxonomías descritas anteriormente se basan en las actividades diarias de

Universitat d'Alacant
Universidad de Alicante

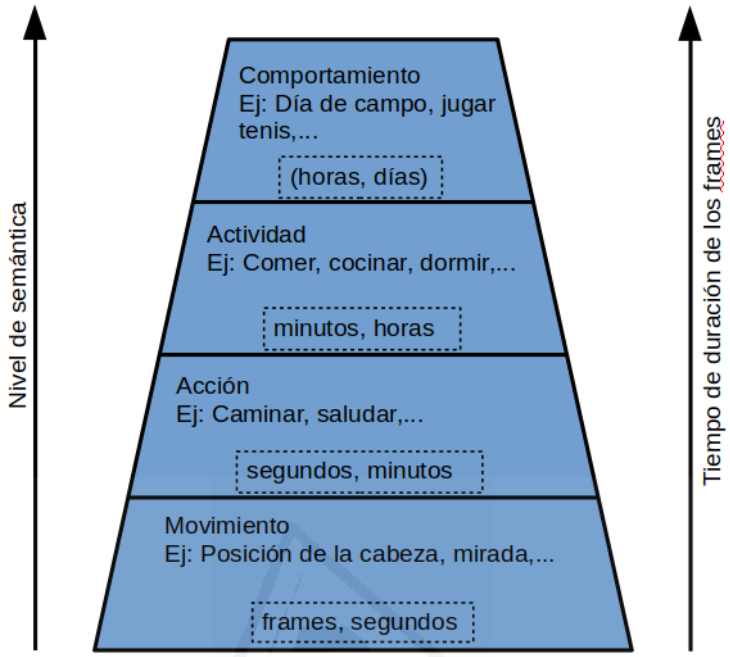


Figura 2.1: La figura muestra las cuatro niveles de semántica definidos en [49] con el tiempo de duración de cada uno.

Universitat d'Alacant
Universidad de Alicante

las personas, tomando en cuenta factores importantes como duración y las actividades compuestas por otras partes más simples como los movimientos y acciones. Describen los niveles/órdenes de comportamiento de los movimientos simples que duran segundos, hasta actividades complejas realizadas por personas durante varios minutos, horas e inclusive días.

Los investigadores han propuesto las diferentes clasificaciones en función de unos datos, o los aspectos más relevantes en sus investigaciones, pero no hay un consenso general al respecto.

Además, la falta de abstracción en la generalización en las propuestas actuales, en términos de número de individuos en la escena (es decir, de grupo de dos o más personas hasta multitudes), dificulta el establecimiento de una arquitectura de referencia para definir cómo abordar diferentes los casos posibles utilizando propuestas similares.

2.3. Evolución del aprendizaje profundo

El Análisis del Comportamiento Humano (HBA), también conocido como Reconocimiento del Comportamiento Humano, consiste en detectar la acción/actividad/comportamiento de las personas utilizando técnicas de Inteligencia Artificial (IA). Diferentes enfoques abordan este problema desde diferentes perspectivas, ya sea en un bajo nivel de comprensión como una sola acción (por ejemplo, mover una mano), o en un comportamiento más complicado (por ejemplo, ir de compras) [39, 50]. Por otro lado, el objetivo puede centrarse en analizar comportamientos de una sola persona o en grupos y multitudes [100].

El análisis de comportamiento humano se abordó inicialmente con las técnicas tradicionales de Aprendizaje Automático, y en los últimos años, las soluciones de Aprendizaje Profundo (DL) han mostrado una mejora en la precisión de los resultados. Las propuestas clásicas de HBA utilizaron varios de los métodos conocidos de IA, como los modelos de Markov (HMMs) [288], SVM [304] y los clasificadores como AdaBoost[249], Sistema de detección de intrusos (IDS)[378], métodos de clasificación probabilísticos[245, 273] y muestreo estocástico[266, 345], análisis del modelo de formas de datos a partir de datos 2D y 3D[110], interacciones humano-computadora[6, 269, 269], clasificación semántica del comportamiento[6, 362, 402]. También, métodos como el Mapa Auto-Organizado (SOM)[187], Mapas Auto-Organizados Supervisados (SSOM)[270], el Mapa de Papadimitriou 2002 supervisado, Neural GAS (NGAS)[242], Análisis Lineal del Discriminante (LDA)[26], Cadenas de Markov-Monte Carlo (MCMC)[225], Model Gausiano Mixto (GMM)[225],Histograma of Flujo Optico (HOF)[137] han sido usados.

Sin embargo, estos métodos antes citados necesitaban un pre-procesamiento de las imágenes para segmentar y rastrear a la persona o grupo de personas en escena, y después ingresar estos datos de entrada al clasificador. Este proceso hacía que la clasificación dependiera en gran medida de la calidad de cada uno de los pasos (segmentación, seguimiento, clasificación). Se ha demostrado que la utilización de un descriptor de trayectoria mejora los resultados, ya que reduce al mínimo el ruido de rastreo y proporciona una

representación homogénea del movimiento en la escena. El Activity Descriptor Vector (ADV) [18], por ejemplo, se evaluó con varios métodos de ML que superan los enfoques anteriores. Además, se extendió al Grupo HBA (GADV [19]) con una mejora similar, y se probó en problemas de predicción [23].

En campo de investigación de Visión por Computador, las Redes Neuronales Profundas han evolucionado para ser usadas consistentemente debido a sus buenos resultados. Los métodos de Aprendizaje Profundo han ganado superioridad sobre otros en el campo del reconocimiento y la clasificación en imágenes y video, como el reconocimiento de acciones basado en LSTM [101, 229, 347, 347], arquitecturas basadas en múltiples flujos para el reconocimiento del comportamiento [138, 175, 375], análisis basados en esqueletos para el reconocimiento del comportamiento humano [90, 235, 280].

Como describimos en el Capítulo 1, desde la década de los años 1950, el Aprendizaje Máquina (ML), un subconjunto de la Inteligencia Artificial (IA), ha revolucionado varios campos de investigación en las últimos años. Las redes neuronales son un subconjunto del Aprendizaje Máquina (ML), de esto surge el Aprendizaje Profundo (DL). Desde un inicio, el DL ha conseguido resultados más eficientes, en muchos dominios de aplicación. En estos términos, podemos definir al ML como un proceso que estima los parámetros de un modelo, para que pueda resolver un problema específico. Algunos autores describen al DL como un enfoque de aprendizaje universal que es capaz de resolver diversos problemas en diferentes dominios de aplicación. En otras palabras, el DL no fue pensado para resolver un problema específico.

En la tabla 2.3 se muestra, los eventos más relevantes sobre los avances de las investigaciones en DL desde el año 1962 hasta 2013, y como complemento hasta los estudios actuales se muestra en 2.3. En 2.3 se observa la evolución del ML según el punto de vista del autor Schmidhuber en su trabajo [320].

El ML data originalmente de los años 60, con propuestas de neuronas inspiradas en la neurobiología. Más adelante en la década de los años 70 se proponen conceptos básicos como Backpropagation, Convoluciones, etc.

Año	Título	Referencia
1962	Neurobiological Inspiration Through Simple Cells and Complex Cells	[158]
1970	Backpropagation	[11, 43, 181]
1979	Deep Neocognitron, Weight Sharing, Convolution	[107, 107, 108]
1987	Autoencoder Hierarchie	[145, 146, 282]
1989	Backpropagation for CNN	[107, 107, 206, 207]
1991	Fundamental Deep Learning Problem	[147, 148]
1997	Supervised Deep Learner (LSTM)	[114, 128, 129, 130]
2006	Deep Belief Networks / CNN Results	[25, 146, 291, 336]
2009	First Competitions Won by Deep Learning	[131, 322]
2010	Plain Backpropagation on GPUs Yields Excellent Results	[68, 70]
2011	MPCNN on GPU / First Superhuman Visual Pattern Recognition	[67, 68, 71]
2012	First Contests Won on Object Detection and Image Segmentation	[66, 69, 191, 305]
2013	More Contests and Benchmark Records	[118, 369, 401]

Durante los años 80 y 90 se produce una gran avance con el desarrollo de los Autoencoders, planteamiento del problema del DL, entre otros. A finales de los 2000 y desde 2010 en adelante es cuando la potencia de cálculo usando GPU permite el gran avance del Deep Learning.

La mayor parte de Redes Neuronales Convolucionales Profundas están conformadas por capas llamadas básicas, que incluyen capas de convolución, muestreo, densas, completamente conectadas, entre otras. Entre los ejemplos de este tipo de redes podemos citar a LeNet[208], VGGNet[410], NiN[221]. Así mismo, con el avance de las investigaciones en este campo, se han propuesto otras arquitecturas más avanzadas como: DenseNet[156], FractalNet[200], GoogleNet[358], Inception[357] y Redes Residuales. Otras redes son DCNN, AlexNet[191], VGG, GoogleNet, Dense CNN y FractalNet. Para analizar datos a gran escala se conoce a GoogleNet y Resnet, como alternativa FractalNet Network, mientras que VGG es para uso general.

En la tabla 2.3 se muestra una organización en orden cronológico de las principales redes para Aprendizaje Profundo como complemento a las innovaciones mostradas en la tabla 2.3. Como se aprecia, es a partir del año 2012 cuando se proponen la mayor parte de implementaciones.

Año	Arquitectura	Referencia
2012	AlexNet	[191]
2013	ZFNet / Clarifai	[401]
2014	VGGNET, GoogLeNet	[338, 357, 358]
2015	Residual Network(ResNet)	[143, 163]
2016	FractalNet	[200]
2017	Densely Connected Network(DenseNet)	[156]

2.4. Redes Neuronales Artificiales (NN) y arquitecturas

Las Redes Neuronales Artificiales (NN) han sido motivo de estudio de los investigadores con el objetivo de solucionar problemas grandes y complejos de manera eficiente. Los modelos de Redes Neuronales son capaces de encontrar patrones de forma inductiva utilizando algoritmos de aprendizaje basado en los datos existentes sin requerir ayuda de un agente externo para programar su forma de funcionamiento. Los diferentes algoritmos pueden funcionar en forma independiente o combinada para mejorar su eficiencia. Las Redes Neuronales ajustan en forma dinámica los valores de los pesos de las interconexiones entre neuronas, además su funcionamiento está fundamentado en el del cerebro humano conformado por neuronas biológicas que por si solas pueden almacenar poca información, sin embargo, cuando se estableces conexiones entre estas su rendimiento mejora notablemente. Las neuronas artificiales están agrupadas en capas que tienen una gran cantidad de conectividad entre ellas, esta conectividad es ponderada por el valor de los pesos. Mediante un algoritmo de aprendizaje que puede ser supervisado, no supervisado y por refuerzo, estas redes realizan ajustes en su arquitectura y valores de parámetros con el objetivo de encontrar un mínimo valor en la función de error.

Las Redes Neuronales Artificiales o arquitecturas de estas pueden ser de varios tipos o usos, y características según la necesidad de los investigadores para resolver un tipo de problema, a continuación se muestra una clasificación de redes y arquitecturas con sus principales ventajas y características:

Universitat d'Alacant
Universidad de Alicante

2.4.1. Redes Neuronales Convolucionales (CNNs):

[216] Este tipo de redes son parecidas a redes simples como el perceptrón multi-capa, se conforman de neuronas que tienen pesos y sesgos que pueden aprender. Además están inspiradas en la estructura de la corteza visual que está conformada por dos tipos de células, unas simples y otras complejas. Adopta cuatro ideas principales del sistema biológico: las conexiones locales y uso compartido de parámetros, pooling y multicapas. En las Redes Neuronales Convolucionales (CNNs), la operación de convolución reemplaza la multiplicación general de la matriz en red neural general. De esta manera, la complejidad de la red se reduce debido a la disminución de número de pesos. Cabe señalar que las CNNs es la primera arquitectura de Aprendizaje Profundo (DL) con la arquitectura jerárquica de capas. Para utilizarlas en problemas de comportamiento humano, se tiene dos ventajas principales. La primera ventaja permite a CNN extraer características localizadas de las posiciones que están relacionadas con el espacio, en lugar de hacerlo desde una sola posición. La segunda ventaja permite que la información sobre el ritmo o la frecuencia se mantenga en la información extraída.

Arquitecturas CNN En los últimos años desarrollo tecnológico en el campo de la Inteligencia Artificial ha sido muy evidente ya que ha permitido implementar muchas cosas que antes se les consideraban simplemente como teóricas, entre los avances en investigación que se ha logrado implementar con mucho éxito es las arquitecturas de redes neuronales convolucionales, entre estas existen algunas que se les puede considerar como las pioneras ya que fueron las primeras en ser publicadas, en base a estas redes muchos investigadores han diseñado sus propias arquitecturas o han hecho uso de estas en sus trabajos, las arquitecturas más comunes en orden cronológico son las siguientes:

LeNet: [208] Es pionera de las Redes Neuronales Convolucionales, implementada con muchas restricciones debido a la potencia de procesamiento requerida para este tipo de tareas. La variante más conocida desarrollada por Yann LeCun[208], la variante más conocida es la Arquitectura LeNet-5 que se utilizó para el reconocimiento de dígitos.

AlexNet:[191] fue desarrollada por Alex Krizhevsky, Ilya Sutskever y

Geoff Hinton en el año 2012, fue el primer trabajo que popularizó Convolutional Neural Networks (CNN) en Visión por Computador. La red AlexNet fue probada en el reto ImageNet ILSVRC en el año 2012 y la red neuronal, con 60 millones de parámetros y 650.000 neuronas, tiene cinco capas convolucionales, con capas de máxima concentración, y tres capas completamente conectadas con un softmax final de 1000 vías.

ZFNet: en 2013 Matthew D. Zeiler y Rob Fergus desarrollaron ZF Net[401]. (Zeiler & Fergus Net), fue ganador en ILSVRC 2013. Esta red fue una adaptación específica a la arquitectura AlexNet[191], pero aún así desarrolló algunas ideas originales para mejorar el rendimiento. Los investigadores explican ConvNets y muestran cómo visualizar los filtros y pesos correctamente.

GoogleNet: Ch. Szegedy et al.[358] fue el mejor trabajo de Convolutional Neural Networks presentado en ILSVRC GoogLeNet tiene 22 capas de red profunda, cuya calidad se evalúa en el contexto de la clasificación y detección. La principal contribución fue el desarrollo de un Módulo de Inicio (Inception Module)[357] que redujo drásticamente el número de parámetros en la red (4M, comparado con AlexNet con 60M).

VGGNet: Karen Simonyan and Andrew Zisserman fue desarrollado en 2014 en una CNN llamada VGGNet [338]. La contribución principal es una evaluación completa de las redes con profundidad creciente utilizando una arquitectura con filtros de convolución muy pequeños (3x3), lo que muestra que se puede lograr una mejora significativa en las configuraciones empujando la profundidad a 16-19 capas de peso. Este trabajo fue ganador en la presentación de ImageNet Challenge 2014 a las pistas de localización y clasificación respectivamente.

ResNet: Las Deep Residual Networks (e.g. ResNet[142], [163], [143]) han evolucionado como una familia de arquitecturas extremadamente profundas que muestran una alta precisión y un comportamiento de convergencia apropiado. Residual Network fue el ganador del ILSVRC 2015. Este informe de trabajo mejoró los resultados utilizando un ResNet de 1001 capas en CIFAR-10 (error del 4,62 %) y CIFAR-100, y una red ResNet de 200 capas en ImageNet.

FractalNet: Esta arquitectura es una arquitectura avanzada y alter-

Universitat d'Alacant
Universidad de Alicante

nativa del modelo ResNet, que es eficiente para diseñar modelos grandes con profundidad nominal, pero con trayectos más cortos para la propagación de gradiente durante el entrenamiento [200]. Este concepto se basa en el "drop-path", que es otro enfoque de regularización para la creación de redes grandes. Como resultado, este concepto ayuda a cumplir los objetivos de velocidad frente a los de precisión.

Network in Network: En este tipo de red se consideran dos conceptos importantes [221], el primero es utilizar la convolución de percepción multicapa, donde se realizan convoluciones con filtros de dimensión 1 x 1, que ayudan a añadir más no linealidad a los modelos. Este concepto se utiliza a menudo en la capa cuello de botella en un modelo de aprendizaje profundo. El segundo concepto es utilizar el Global Average Pooling(GAP) como una alternativa a las capas totalmente conectadas. Esto ayuda a reducir significativamente el número de parámetros de la red. Entonces, GAP cambia significativamente la estructura de la red. Aplicando GAP sobre un mapa de características de gran tamaño, podemos generar un vector final de características de baja dimensión sin reducir la dimensión de los mapas de características.

Densely Connected Network(DenseNet): Esta arquitectura fue propuesta en 2017 por los autores en[156], está conformada por capas de CNN densamente conectadas, las salidas de cada capa están conectadas con todas las capas sucesoras en un bloque denso. Por lo tanto, se forma con una conectividad densa entre las capas, esta característica le da el nombre de DenseNet. Este concepto es eficiente para la reutilización de funciones, lo que reduce en forma significativa los parámetros de la red. DenseNet está formada por varios bloques densos y bloques de transición, que se colocan entre dos bloques densos adyacentes. Cada capa toma todos los mapas de características anteriores como entrada, la capa n-ésima recibe todos los mapas de características de las capas anteriores.

Universitat d'Alacant
Universidad de Alicante

2.4.2. Redes Neuronales Recurrentes(RNNs)

En términos de informática, los pensamientos humanos guardados en el cerebro y tienen persistencia, la acción de preservar la información de un objeto de forma permanente, es decir queda guardado, se le conoce como

persistencia, pero a su vez también se refiere a la capacidad de recuperar o leer la información para que pueda ser nuevamente utilizada. Esto quiere decir, que los humanos no tiran por la borda toda la información que reciben a través de los pensamientos. Los enfoques tradicionales de la red neuronal, incluyendo las DNN y CNN no pueden hacer frente a este tipo de problemas por las siguientes razones: Primero, estos enfoques sólo manejan un vector de tamaño fijo como entrada (por ejemplo, una imagen o fotograma de vídeo) y producir un vector de tamaño fijo como salida (por ejemplo, probabilidades de las clases). En segundo lugar, estos modelos funcionan con un número fijo de pasos computacionales (por ejemplo, el número de capas en el modelo). Los RNNs son únicos ya que permiten la operación sobre una secuencia de vectores con el tiempo. El concepto de este tipo de red fue pensado desde el año 1974[317]. Las Redes Neuronales Recurrentes [216] tienen la capacidad de extraer información temporal y semántica, es decir recuerda la información anterior y la utiliza para influir en la salida de los siguientes nodos. Sin embargo, la RNN convencional tiene una limitación: las dependencias a largo plazo. Para superar esta desventaja de funcionamiento, se creó la memoria a corto plazo (LSTM). Si se compara con CNN, que solamente puede procesar datos con tamaño único, la predicción de una red RNN y sus variantes incrementan su precisión cuando se dispone de mayor cantidad de datos. El resultado de la predicción es cambiando con el tiempo. Por consiguiente, RNN es más sensible al cambio de los datos de entrada que una CNN. Para Análisis de Comportamiento Humano, RNN y sus variantes tienen la superioridad de explotar las correlaciones temporales en una actividad, que es una cuestión crucial para el reconocimiento de las actividades humanas. A continuación describiremos varias redes que han sido creadas fundamentadas en el concepto antes descrito:

Recurrent Neural Networks (RNNs): Son un tipo de red neuronal artificial [246],[315] han sido un tema relevante de la investigación de las redes neuronales durante la década de 1990. Esta red añade pesos adicionales a la red para crear ciclos en el gráfico de red en un esfuerzo por mantener un estado interno, las redes neuronales recurrentes son una familia de redes neuronales para el procesamiento de datos secuenciales,

Universitat d'Alacant
Universidad de Alicante

donde las conexiones entre unidades forman un ciclo dirigido, esto le permite mostrar el comportamiento temporal dinámico, también utilizado para analizar secuencias de acciones.

Long Short-Term Memory Networks (LSTMs): [113], es una novedosa formación en arquitectura de red recurrente, supera a la RNN tradicional con un algoritmo de aprendizaje basado en el grado adecuado, puede aprender a guardar los retrasos mínimos que superan los 1.000 pasos de tiempo discreto mediante la imposición de un flujo de errores constante a través de **carruseles de errores constantes** dentro de unidades especiales, sin pérdida de capacidades de retardo de corto tiempo, puede aproximarse a los dominios problemáticos ruidosos, representaciones distribuidas y valores continuos.

Gated Recurrent Unit Neural Networks (GRUs): [89], Cho et al. en el año 2014 propusieron hacer que cada unidad recurrente capturara de forma adaptativa dependencias de diferentes escalas de tiempo, han demostrado ser exitosas en varias aplicaciones que involucran datos secuenciales o temporales. Al igual que la unidad LSTM, la GRU dispone de unidades de inyección que modulan el flujo de información dentro de la unidad, pero sin disponer de células de memoria separadas.

Hybrid Deep Model: [216] Como describimos anteriormente cada uno de los modelos tiene sus ventajas y desventajas cuando realizan una tarea, una alternativa para aprovechar estas ventajas y mitigar las desventajas es la utilización de Modelos Híbridos Profundos, es decir se puede integrar varias redes juntas y aprovechar las bondades de todas estas redes. El objetivo principal de proponer este tipo de modelos es mejorar la precisión. En el Análisis de Comportamiento Humanos, CNNs y RNNs se pueden combinar comúnmente, porque son buenos para abstraer diferentes características del dominio: CNN captura las relaciones espaciales mientras que RNN captura las relaciones temporales[373, 412].

2.4.3. Auto-Encoders(AEs)

Las arquitecturas pueden ser utilizadas para la detección de anomalías en imágenes y vídeos es un reto muy importante y difícil para los investigadores, en los últimos tiempos los trabajos sobre este reto se fundamentan

en métodos basados el aprendizaje profundo como: los autoencoders (AE). Se trata de una red neuronal entrenada por backpropagation, es usada para representar las diferentes transformaciones lineales y no lineales de las imágenes o movimientos, definidos como comportamientos normales en los vídeos de vigilancia. Las anomalías representan las desviaciones mal reconstruidas. Esta red tiene algunas variaciones, Auto-Encoders Convolucionales (CAEs), Auto-Encoders Contractivos, Autoencoders Variacionales (VAEs), Auto-Encoders Adversariales (AAEs), Auto-Encoders Desactivadores (DAE) y DAEs Apilados (SDAEs), estos métodos han sido agrupados y descritos en [186], y su objetivo es la detección de comportamientos anormales en videos o imágenes. Algunos de estos métodos se describen a continuación:

Auto-Encoder:[216] Es una red neuronal de retro-alimentación que tiene como objetivo reconstruir los datos de entrada de la red con ciertas restricciones. Tiene la capacidad de aprender representaciones de características profundas de entradas no etiquetadas, es decir de aprendizaje no supervisado, a través de varias repeticiones de procedimientos de codificación y decodificación. Cuando los datos de entrada son muy parecidos, el Auto-Codificador (AE) es capaz de descubrir matices en los datos en sí mismos por pre-entrenamiento no supervisado. Además, la formación previa no supervisada tiende a funcionar como un que impide potencialmente que la red se sobreentrene. En Análisis de Comportamiento Humano tiene algunas variantes como: Auto-Codificador de Pila (SAE) que apila varios AEs dispersos para adquirir una codificación de características más compacta. Auto-Codificador Convolucional (CAE) que combina esencialmente CNNs y AEs, y los procedimientos de codificación y descodificación se realizan por convolución y deconvolución.

Convolutional Auto-Encoders (CAEs): [243], se utilizan para el aprendizaje de características sin supervisión. Una pila de CAEs clasifica una Red Neuronal Convolucional (CNN). Cada CAE es entrenado usando un gradiente en descenso en línea, sin términos de regularización adicionales. Una CNN puede ser inicializada por una pila CAE. Mientras que la representación oculta demasiado completa del CAE hace que el aprendizaje sea aún más difícil que en el caso de los Auto-Codificador (CAEs)

estándar, surgen buenos filtros si utilizamos una capa de max-pooling.

Contractive Autoencoders (COAes): [299], La función principal de un Contractive Autoencoder, es crear explícitamente la invariancia añadiendo el jacobiano del espacio latente. La mejor manera de hacerlo es añadir un término de penalización a la función de coste que estamos intentando minimizar, lo que penaliza la sensibilidad de la representación a la aportación de formación. Los investigadores probaron su enfoque en un punto de referencia de problemas de clasificación de imágenes, a saber: CIFAR-bw: una versión en escala de grises de la tarea de clasificación de imágenes CIFAR10 (Krizhevsky y Hinton, 2009) y MNIST.

Variational Autoencoders (VAEs): [14], Este método es una detección de anomalías que utiliza la probabilidad de reconstrucción del Autocodificador Variacional. El codificador y el decodificador son una sola capa oculta con 400 dimensiones, la dimensión latente es de 200 dimensiones. Los investigadores utilizan un Auto Codificador denotativo de dos capas ocultas con 400, 200 dimensiones para la primera y segunda capa oculta, respectivamente, la segunda capa fue entrenada apilando la salida de la capa anterior.

Universitat d'Alacant
Universidad de Alicante

Adversarial Autoencoders (AAEs): [238], es un autoencoder probabilístico que utiliza las Redes Generativas Adversarias (GANs) recientemente propuestas, para realizar la inferencia variacional haciendo coincidir la parte posterior agregada del vector de código oculto del Auto Codificador con una distribución previa arbitraria, el Adversarial Autoencoder puede ser utilizado en aplicaciones tales como la clasificación semi-supervisada.

De-noising Auto-Encoders (DAE) and Stacked DAEs (SDAEs): son métodos de extracción de características para entornos de aprendizaje no supervisado, en estos métodos, los criterios de minimización de errores de reconstrucción se complementan con los de reconstrucción a partir de entradas dañadas. Los SDAE se utilizan para aprender representaciones de un vídeo utilizando tanto la apariencia, por ejemplo, valores originales o videos en bruto, como por ejemplo el movimiento detectado (flujo óptico) en secuencias de imágenes[186].

2.4.4. Redes Generativas Adversarias(GANs)

El área de visión por computador tiene diferentes sub-áreas, segmentación, clasificación, detección, entre otras. Cuando utilizamos aprendizaje profundo para un mejor funcionamiento de este, se requieren de grandes cantidades de datos etiquetados. Este problema se ha intentado resolver generando muestras similares o datos sintéticos utilizando un modelo generativo. La Rede Generativa Adversaria (GAN)[122] es un enfoque de Aprendizaje Profundo recientemente inventado por Goodfellow en 2014. Las GAN ofrecen un enfoque alternativo a las técnicas de estimación de máxima verosimilitud. La GAN es un enfoque de Aprendizaje Profundo no supervisado en el que dos redes neuronales compiten entre sí en un juego de suma-cero[77], describe una situación en la que la ganancia o pérdida de un jugador se equilibra con exactitud, con las pérdidas o ganancias de los otros jugadores. En el caso del problema de generación de imágenes, el generador comienza con el ruido Gaussiano para generar imágenes y el discriminador determina la calidad de las imágenes generadas. son. Este proceso continúa hasta que las salidas del generador se acercan a las muestras de entrada reales.

El uso de Aprendizaje Máquina debido a los avances tecnológicos actuales especialmente de hardware nos permite la posibilidad de experimentar el funcionamiento en tiempos reducidos de arquitecturas cada vez más complejas con redes neuronales de diferente tipo como se describió en líneas anteriores, este tipo de soluciones han sido planteadas por los investigadores y se han vuelto populares debido a que los resultados obtenidos superan a los métodos tradicionales, mejorando de este modo los resultados. La posibilidad de experimentar en equipos con mucha potencia de almacenamiento y cálculo se pueden realizar pruebas con muchas combinaciones de redes organizadas en arquitecturas, en este trabajo se tiene como objetivo combinar algunas redes conocidas como VGG[410], ResNet50[142], AlexNet[191, 308], Yolo [294, 295, 326], RetinaNet[223], y otras para incrementar la precisión de la predicción. Esta arquitectura propuesta será explicada y analizada más adelante.

Con la aparición de técnicas y algoritmos de Aprendizaje Profundo, el avance de muchos campos de la investigación como el reconocimiento de

voz o reconocimiento de objetos. Como lo habíamos descrito antes el modelo de aprendizaje tiene múltiples capas de procesamiento para aprender representaciones de alto nivel automáticamente. Como resultado de estos avances, el Aprendizaje Profundo ha impulsado notablemente el desarrollo de muchos campos, incluyendo Análisis de Comportamiento Humano. En esta sección, se expuso varios modelos de Aprendizaje Profundo y se analizó sus ventajas de funcionamiento.

2.5. Deep Learning para HBA

En el presente capítulo mostraremos cómo puede usarse el DL para solucionar problemas relacionados con el HBA, para obtener resultados favorables para predicción de comportamiento humano con mayor o menor semántica, y también con su tiempo de duración. En los trabajos actuales se ha tratado de resumir en términos que agrupen conocimientos sobre el tema, entre los principales temas que los investigadores usan para analizar el comportamiento humano están: Análisis de Comportamiento Humano (HBA)[48, 49, 51, 84, 194, 268], aplicaciones tipo Activities of Daily Living (ADL)[373] que generalmente utilizan datos extraídos con diferentes tipos de sensores. La aceleración y la velocidad angular se modifican en función de los movimientos del cuerpo humano, por lo que pueden inferir actividades humanas. Estos sensores se pueden encontrar usualmente en teléfonos inteligentes, relojes, bandas, gafas y cascos , Reconocimiento de Actividades Humanas (HAR)[169, 170, 250], es algo común en todos los aspectos de la vida diaria. y por lo tanto, este tema se ha vuelto interesante para los investigadores. se estudia las actividades diarias básicas como sentarse, caminar, correr, y pueden ser monitoreadas y con un alto nivel de precisión si el usuario lleva una gran cantidad de nodos de sensores principalmente en teléfonos inteligentes. En nuestro estudio la fuente de datos para detectar actividades humanas es un video, aplicaciones tipo Ambient Assisted Living (AAL) tienen como objetivo mejorar la calidad de vida y mantener la independencia especialmente de las personas mayores y vulnerables usando tecnología. Las tecnologías AAL de última generación utilizan muchos sensores integrados en el entorno o en la vivienda de las personas.

Universitat d'Alacant
Universidad de Alicante

Estos sensores suelen ser invasivos y muchas veces se activan presentando falsas alarmas, ante esto una opción alternativa es el uso de cámaras que puedan detectar principalmente caídas de personas, ya que es este tipo de actividad de personas mayores que es más frecuente, por esta razón existe una creciente tendencia en soluciones basadas en vídeo y visión artificial para aplicaciones AAL[27, 41, 80, 120, 254].

En [367] los autores presentan un modelo de aprendizaje basado en relaciones contextuales que utiliza una red neuronal profunda para reconocer las actividades realizadas por un grupo de personas en una secuencia de vídeo. El modelo propuesto comprende el aprendizaje contextual utilizando un enfoque ascendente, aprendiendo desde las acciones humanas individuales hasta las actividades a nivel de grupo, así como aprendiendo de la información de la escena. Tomando en cuenta que puede identificar actividades de grupo e individuales se consideraría un avance, ya que es uno de los primeros trabajos que pude identificar comportamientos de acuerdo al número de personas en escena, aclarando que todavía no se llega al nivel de identificación de comportamientos en multitudes con una misma aplicación [136, 176, 366, 367, 370, 380].

En [88] los autores proponen una solución basada en la visión para identificar las Actividades de Vida Diaria (ADL), a través de datos esqueléticos capturados con una cámara RGB-D. Tras la descomposición de una secuencia esquelética en segmentos temporales cortos, las actividades se clasifican a través de una red llamada Long-Short Term Memory Network (LSTM) de dos capas, que permite analizar la secuencia a diferentes niveles de granularidad temporal. La propuesta se evalúa en el dataset Watch-n-Patch[383], en el que se encuentran ejemplos 11 diferentes actividades diarias de personas como: traer cosas desde la nevera, volver cosas hacia la nevera, derramamiento de liquido, beber un líquido, salir de la cocina, traer cosas del horno, meter cosas al microondas, preparar la comida, llenar una tetera, conectar una tetera la toma eléctrico, mover la tetera. La principal contribución de los autores es un modelo de actividades de múltiples escalas y dependencia temporal, basado en la comparación de las características del contexto que caracterizan los resultados de reconocimientos anteriores, y una representación jerárquica con una capa de

Universitat d'Alacant
Universidad de Alicante

reconocimiento de unidades de comportamiento de bajo nivel y otra unidad de alto nivel. Es decir, es una solución que maneja dos diferentes niveles de semántica.

En este trabajo se han analizado varios trabajos que tratan sobre Redes Neuronales Convolucionales (CNN), Auto Codificadores (AE) y Redes Neuronales Recurrentes (RNN), cada una con algunas de sus variantes, para proponer una clasificación de los trabajos de análisis de comportamiento de grupos y multitudes de acuerdo con el número de individuos y el nivel de comprensión en cuanto a la duración de la conducta detectada.

Desde el inicio del reto de ImageNet [192], las técnicas de Aprendizaje Profundo han avanzado vertiginosamente los trabajos relacionados con este tema, se ha mejorando la precisión y utilidad de esta poderosa herramienta en el área de la Visión por Computador.

Anteriormente se había descrito los términos movimiento, acción, actividad y comportamiento, que son definiciones de acuerdo al nivel de semántica, la descripción taxonómica de la interacción de seres humanos con objetos en un determinado entorno. Esto comúnmente se conoce como comportamiento humano, y los investigadores han clasificado las investigaciones de este tipo en campos como el Análisis de Comportamiento Humano o HBA con sus siglas en idioma inglés. Para abordar este problema inicialmente se ha propuesto varios métodos para fines de explicación los llamaremos tradicionales, posterior a esto tenemos otro grupo de técnicas para reconocer comportamiento que son las Redes Neuronales Convolucionales (CNNs), y anteriormente explicamos, finalmente en la actualidad se han están proponiendo redes especializadas y conjuntos de redes agrupadas llamadas arquitecturas. Estas herramientas nos ayudan a detectar diferentes tipos de comportamientos, actividades y otro tipos de términos relacionados con la interacción humana del día a día con el mundo real.

En [354] se definen las acciones humanas en las secuencias de vídeo son señales espacio-temporales tridimensionales (3D) que caracterizan tanto la apariencia visual como dinámica del movimiento de los seres humanos y objetos involucrados. Tomando en cuenta el éxito generado por los resultados positivos de las Redes Neuronales Convolucionales (CNN) para

Universitat d'Alacant
Universidad de Alicante

la clasificación de imágenes, se han hecho intentos recientes de aprender CNN 3D para reconocer las acciones humanas en los videos, pero debido a la alta complejidad del entrenamiento de los núcleos de convolución 3D y la necesidad de grandes cantidades de videos de entrenamiento que este tipo de redes requiere para su entrenamiento, sólo existen pocos casos de éxito, siendo un amplio tema todavía que requiere madurez en la investigación.

A continuación describiremos los diferentes tipos de arquitecturas que actualmente se han vuelto populares para estudiar HBA:

RetinaNet: [223] Es una red convolucional unificada, responsable de calcular un mapa de características convolucional sobre una imagen de entrada completa, conformada por dos sub-redes, cada una destinada a tareas específicas, la primera sub-red realiza la clasificación de objetos convolucionales en la salida de la red troncal; la segunda sub-red realiza una regresión convolucional de la caja delimitadora. Las dos sub-redes tienen un diseño sencillo que permite la detección densa de una sola etapa.

YOLOv3: [94, 124, 295]: Esta propuesta aplica una sola red neuronal a la imagen completa en el momento de la prueba, de modo que sus predicciones se basan en el contexto global, divide la imagen en regiones y predice cuadros delimitadores y probabilidades para cada región. Estas cajas delimitadoras son ponderadas por las probabilidades pronosticadas. tiene varias ventajas sobre los sistemas basados en clasificadores. Además realiza predicciones con una sola evaluación de red a diferencia de sistemas como R-CNN que requieren miles para una sola imagen. Esto lo hace extremadamente rápido, más de 1000 veces más rápido que R-CNN y 100 veces más rápido que Fast R-CNN.

FasterR-CNN:[117, 172, 297] Se basa en redes convolucionales profundas para clasificar eficientemente objetos, emplea varias innovaciones para mejorar la velocidad de entrenamiento y pruebas, también aumenta la precisión de la detección. Fast R-CNN es 9 veces más rápida que R-CNN[118], 213 veces más rápida en el momento de la prueba, y logra un mayor mAP en PASCAL VOC 2012[97]. En comparación con SPPnet[141, 284], comparado con VGG16[338] es 3 veces más rápido, en pruebas 10 veces más rápido y es más preciso.

R-FCN: [82] Es una red regional totalmente convolucional para una

Universitat d'Alacant
Universidad de Alicante

detección de objetos, con casi todo el cálculo computacional compartido en la imagen completa. Para lograr este objetivo, los autores proponen mapas de puntuación sensibles a la posición para abordar un dilema entre la variación de traducción en la clasificación de imágenes y la variación de traducción en la detección de objetos. De este modo, este método puede adoptar naturalmente backbones de clasificadores de imágenes totalmente convolucionales, como las últimas Redes Residuales (ResNets)[142], para la detección de objetos. los resultados en el dataset PASCAL VOC [97] (83.6 % mAP en la versión 2007) con una ResNet de 101 capas. Además el resultado de cálculo se obtiene a una velocidad de prueba de 170 ms por imagen, 2,5-20 veces más rápido que el de la versión R-CNNN [118] más rápida.

SSD: [230] Es un método para la detección de objetos en imágenes utilizando un único método red neuronal profunda, discretiza el espacio de salida de las cajas delimitadoras en un conjunto de cajas predeterminadas sobre diferentes relaciones de aspecto y escalas por ubicación de mapa de características. En la predicción, la red genera puntuaciones por la presencia de cada categoría de objeto en cada caja por defecto, y produce ajustes en la caja para que coincida mejor con la forma del objeto. Además, la red combina predicciones de múltiples mapas de características con diferentes resoluciones para manejar naturalmente objetos de varios tamaños. SSD es simple en relación con los métodos que requieren propuestas de objetos porque elimina completamente la generación de propuestas y las etapas subsiguientes de remuestreo de pixeles o características, además encapsula todos los cálculos en una sola red. Esto hace que las unidades SSD sean fáciles de entrenar y fáciles de integrar en sistemas que requieren un componente de detección. Los resultados experimentales en los conjuntos de datos PASCAL VOC[97], COCO [60, 224] e ILSVRC [191] confirman que la SSD tiene una precisión competitiva con respecto a los métodos que utilizan un paso adicional de propuesta de objeto y es mucho más rápida, a la vez que proporciona un marco unificado tanto para la formación como para la inferencia.

FPN: [222] es una arquitectura de tipo top-down con conexiones laterales para construir mapas de características semánticas de alto nivel a

todas las escalas. Esta arquitectura, llamada Feature Pyramid Network (FPN), muestra una mejora significativa como extractor de características genéricas en varias aplicaciones. Utilizando FPN en un sistema básico R-CNN [118] más rápido, este método logra resultados de modelo único de última generación en el punto de referencia de detección COCO[60, 224]. Además, este método puede funcionar a 5 fps en una GPU y, por lo tanto, es una solución práctica y precisa a las múltiples escalas detección de objetos.

CornerNet: [203] Detecta un objeto en una imagen como un par de puntos, en la esquina superior izquierda y la esquina inferior derecha del cuadro delimitador. Es una red convolucional que predice dos conjuntos de mapas de calor para representar la ubicación de las esquinas de diferentes objetos, un conjunto para las esquinas superior izquierda, y el otro para la esquina inferior derecha. Además predice un vector de incrustación para cada esquina detectada de tal manera que la distancia entre las incrustaciones de dos esquinas desde el el mismo objeto es pequeño. Para producir cajas delimitadoras más ajustadas, la red también predice compensaciones para ajustar ligeramente las posiciones de las esquinas. Con la predicción mapas de calor, incrustaciones y compensaciones, aplica un simple algoritmo de post-procesamiento para obtener las cajas delimitadoras finales de los objetos.

Universitat d'Alacant
Universidad de Alicante

2.6. Datasets

Un conjunto dataset es un conjunto de datos, que en el caso de los utilizados en esta tesis, contienen miles de imágenes destinadas a las pruebas, entrenamiento y validación. También se incluyen archivos que detallan la información de las diferentes imágenes, por ejemplo, la descripción de lo que representa la imagen, detalles de coordenadas de algún elemento dentro de ella, tipos de acciones que una imagen representa, y otros.

Existen datasets con características especiales según el tipo de investigación, e.g. gran cantidad de datos para DL, secuencias de imágenes para HBA, profundidad para estudios 3D, datos de grupos de personas, etc. A continuación describiremos los diferentes tipos de datasets que actualmen-

te son usados HBA para métodos de ML tradicionales en la sección 2.6.1, y los usados para aprendizaje profundo en la sección 2.6.2.

2.6.1. Datasets HBA tradicionales

En esta sección se revisan los principales datasets utilizados en ML para HBA. Puesto que esta tesis se centra en comportamientos de grupo y multitudes, los datasets que se muestran son muy centrados en estos datos. Las tablas 2.1 y 2.2 muestran una clasificación de el tamaño del grupo y las actividades que realizan. Además, se especifican los métodos, algoritmos y formas de reconocimiento HBA que se han utilizado para su estudio. Podemos ver los siguientes campos analizados: Ref=Referencia al artículo, CL=Clasificación y (número de personas si existe), TE = Técnica, D = Nombre del conjunto de datos, LA = Abstracción de nivel. En la columna LA = Abstracción de Niveles mostramos tres niveles de abstracción: Mov = Movimiento, Act = Acción, Actv = Actividad, también dos tareas automáticas, CP = Contar-Personas y Tra=Rastreo.

La clasificación de acuerdo al número de personas es en dos grupos principales GROUP y CROWD. El grupo se define como el conjunto de dos o más personas en un sitio determinado y que realizan una acción o actividad. La multitud es una composición de personas más grande que un grupo que realiza actividades simultáneas.

Los tipos de comportamientos que se analizan con la video-vigilancia son limitados y específicos. Los comportamientos más frecuentemente estudiados son los siguientes: rastreo, trayectorias, ciclistas, peatones, patinadores, cuenta personas en un grupo o multitud, peleas callejeras, interacción objetos-personas, movimientos o acciones relacionadas con los deportes, acciones humanas frecuentes del día a día como (caminar, trotar, correr, boxear, saludar con la mano y aplaudir).

El dataset o conjunto de datos llamado para fines de diferenciación como “tradicional” que se utiliza frecuentemente para la experimentación con el objetivo de Análisis del Comportamiento Humano (HBA) de grupos y multitudes es el siguiente: BEHAVE[37], BIWI[274], CAVIAR[102], VS-PETS, ETH[96], DGPI[62], UHD[351], HMDB[171], SportVU[244], PETS[95], UNM[103], ViF[140], y otros. También en algunos casos los investigadores

utilizan su propio conjunto de datos o videos obtenidos en YouTube. La cantidad de datos, puntualmente imágenes o videos que contienen este tipo de datasets por lo general es menor a los utilizados en entrenamientos de redes neuronales o arquitecturas similares, que en la siguiente sección se explicará ya que ese tipo de datasets es el que se usará en esta tesis.

A partir de la información analizada en los trabajos que usan estos datasets, es posible proponer una clasificación según el nivel de abstracción del HBA de grupos y multitudes según el caso, en orden de menor a mayor duración del comportamiento proponemos tres niveles de abstracción: Movimiento, Acción, Actividad, también dos tareas automáticas, Contar Personas y Seguimiento.

Las técnicas o métodos llamados tradicionales en este trabajo que se utilizan con frecuencia para HBA de grupos y multitudes a través de la video-vigilancia son los siguientes: Bolsa de Palabras (Bag of Words BoW), Modelos Markov Ocultos (Hidden Markov Models HMD), Monte Carlo, Modelo de Mezcla Gaussiana (Gauss Mixture Model), Seguimiento Múltiple Humano (Multiple Human Tracking), Máquinas Vectoriales de Soporte (Support Vector Machine SVM). Muchos autores utilizan estos métodos incluyendo ciertas variaciones o modificaciones de estos para obtener un mejor rendimiento.

Universitat d'Alacant
Universidad de Alicante

2.6.2. Datasets HBA Aprendizaje Profundo

La tabla 2.3 es un resumen de datasets creados en los últimos veinte años, desde 2004 hasta 2015 para aprendizaje profundo y relacionado con HBA, entre los principales dominios de aplicación que definen los autores en [343] están acciones e interacciones humano-humano, humano-objetos, deportes, ADLs(actividades de vida diarias), cocina, compras.

En la tabla 2.4 se incluyen datasets para arquitecturas CNN desde los últimos ocho años, entre 2010 y 2018 que contienen datos de escenas HBA relacionado con el comportamiento de multitudes, esto incluye conteo de personas en una multitud, comportamiento general de una multitud que puede clasificarse en normal o anormal. Entre los conjuntos de datos más destacados están: UCSD[237], ShanghaiTech[411], DynTex[279], Dyntex++[115], UCLA[92], NUS-HGA[421], UMN[103], Mall Dataset[58],

Tabla 2.1: Datasets HBA tradicionales

AR	CL	TÉCNICA	D	LA
[22]	G	Self-Organizing Map (SOM) Supervised Self-Organizing Map (SSOM) Neural GAS (NGAS) Linear Discriminant Analysis (LDA) k-Nearest Neighbour (kNN) Multiclassifier (MC)	CAVIAR[102]	Actv
[331]	C	Collective Transition priors (CT) Mixture of dynamic texture (DTM) Hierarchical clustering (HC) Coherent filtering (CF)	CUHK[418]	Mov
[392]	C	Pedestrian Simulation(PS) Person re-identificatio(PT) Pedestrian tracking(MPF)	NY Central Station[393] Shanghai World Expo[403]	Mov
[394]	C	Motion Pattern Features(MDA)	N	Mov
[7]	G	Stability Features(HDP)	BEHAVE[37]	Actv
[334]	G	Hidden Markov Models(HMM) Dynamic Probabilistic Networks(DPN)	Shanghai World Expo[403] BEHAVE[37]	Mov
[368]	G(50)	Inter-Relation Pattern Matrix(IRPM) Game-Theoric Conversational Groups(GTCG) Spectral Clustering (R-GTCG SC)	DGPI[62]	Actv
[214]	C	Model Dynamic Textures Temporal(MDT-temp) Local Motion Histogram(LMH) Spatail(MDT-spat)	UNM[103] UCSD[237]	Mov
[228]	G(25)	Markov Chain Monte Carlo (MCMC) Gaussian Mixture Model (GMM)	FIFA WC 2006	Tra
[225]	G	Category Feature Vectors (CFVs) Gaussian Mixture Models(GMM) Recognizing algorithm (CFR)	N	Actv
[47]	G	Multiple Human Tracking (MHT) Correct Detected Tracks (CDT) False Alarm Tracks (FAT) Track Detection Failure (TDF)	ETH[96] UHD[351]	Tra
[137]	C	Histogram of Oriented Gradients (HOG) Histogram of Optical Flow (HOF) Motion Boundary Histogram (MBH)	UMN[103] UCSD[237] CUHK[418] PETS2009[95] ViF[140] Rodriguez's[302] UCF[346] Own Dataset	Act
[104]	C	Hidden Markov Models (HMM) Support Vector Machine (SVM) Robust Local Optical Flow (RLOF)	PETS[95] UMN[103]	Act

Tabla 2.2: Datasets HBA tradicionales

AR	CL	TÉCNICA	D	LA
[121]	G	Cumulative Match Characteristic (CMC) Synthetic Disambiguation Rate (SDR) Center Rectangular Ring Ratio-Occurrence(CRRO) Block based Ratio-Occurrence (BRO)	VIPeR[132]	Act
[219]	C	Support Vector Machine (SVM) Library for Support Vector Machines(LIBSVM) Basis Radial Function(BRF) Block Matching Algorithm (BMA)	UMN[103]	Act
[57]	C	Fast Corner Detect(FAST) Support Vector Machine (SVM)	BEHAVE[37]	Act
[119]	G	Evolving Networks(EN) Monte Carlo(MC)	N	Mov
[275]	G	Linear Trajectory Avoidance (LTA)	N	Mov
[72]	G(20)	Bag of words modelling (BoW)	Novel Dataset	Mov
[278]	G	Gaussian Mixture Model(GMM) EM algorithm	N	Actv
[395]	G	Minimum Description Length (MDL)	COLLECTIVE ACTIVITY[64] BEHAVE[37]	Actv
[112]	G	Hidden Markov Models(HMM) Dynamic Bayes Networks(DBN)	BIWI[274]	Tra
[202]	G(20)	Multi-model MHT	Own	Tra
[167]	G	Voronoi Diagrams Model(VDM)	N	Mov
[309]	G	Dynamic Probabilistic Networks (DPNs) Dynamically Multi-Linked (DML) Hidden Markov Model(HMM)	PETS 2004[95] YouTube	Mov
[323]	G(25)	Support Vector Machines (SVM)		Act
[15]	C	Hidden Markov Model (HMM)	N	Mov
[112]	G	Sampling Importance Resampling (SIR) Discrete Choice Model (DCM) Multi Hypothesis Tracking (MHT) Statistical Shape Modeling(SSM)	BIWI[274]	Tra
[184]	G(90)	Heuristic learned(HL)	N	CP
[405]	C	Bag of Words (BoW) Locality-constrained Linear Coding (LLC) Vector Quantization (VQ)	BEHAVE[37]	Mov
[42]	C	Unsupervised Bayesian Clustering Framework(UBCF)	N	Mov
[111]	C	Bayesian Marked Point Process (MPP)	CAVIAR[102] VSPETS SOCCER	CP
[248]	C	Social Force Model(SFM) Pure Optical Flow(POF)	UNM[103]	
[79]	C	Detection of moving regions	METRO	Tra

Universidad de Alicante

Tabla 2.3: Datasets HBA Aprendizaje Profundo

Dataset	Año	Tipo grupo	Clases	Ref.	Dominio de aplicación
KTH	2004	Gp(>2)	6	[198]	Reconocimiento de acciones humanas en ambientes abiertos
Weizmann	2005	Gp(>2)	9	[125]	Reconocimiento de acciones humanas
IXMAS	2006	Gp(>2)	11	[381]	Reconocimiento de acciones humanas
CASIA Action	2007	Gp(>2)	-	[379]	Comportamiento humano e interacciones humano-humano
UIUC Sport	2008	Gp(>2)	-	[361]	Reconocimiento de acciones en deportes
UCF Sports	2008	Gp(>2)	9	[197]	Reconocimiento de acciones en deportes
Olympic Games	2008	Gp(>2)	16	[264]	Reconocimiento de acciones en deportes
Hollywood	2008	Gp(>2)	8	[199]	Reconocimiento de acciones en películas
Hollywood2	2009	Gp(>2)	12	[241]	Reconocimiento de acciones en películas
UT-Interaction	2009	Gp(>2)	10	[310]	Reconocimiento de actividades humano-humano
BEHAVE	2009	Gp(>5)	6	[37]	Reconocimiento de actividades de grupo
HMDB51	2011	Gp(>2)	51	[195]	Reconocimiento de interacciones humano-humano y humano-objetos
UCF50	2011	Gp(>2)	50	[293]	Reconocimiento de actividades humanas en deportes
UT-Interaction	2012	Gp(>2)	6	[310]	Reconocimiento de interacciones humano-humano
MPII Cooking	2012	Gp(>2)	65	[303]	Reconocimiento de actividades diarias en la cocina
UCF101	2013	Gp(>2)	101	[349]	Reconocimiento de acciones en deportes
YouTube Sports 1M	2013	Gp(>5)	487	[178]	Reconocimiento de acciones en deportes
ActivityNet	2015	Gp(>2)	203	[45]	Reconocimiento de actividades humanas
THUMOS'15	2015	Gp(>2)	20	[123]	Reconocimiento de acciones en videos
ChaLearn	2015	Gp(>5)	235	[30]	Aprendizaje de acciones humanas e interacciones
FCVID	2015	Gp(>5)	239	[174]	Reconocimiento de actividades humanas
MOBISERV-AIIA	2015	Gp(>2)	13	[165]	Reconocimiento de actividades de vida diaria (ADLs)
MERL Shopping Dataset	2016	Gp(>2)	5	[342]	Reconocimiento de actividades humanas, interacción humano-objetos
YouTube 8M	2016	Gp(>5	4800	[2]	Reconocimiento de actividades humanas
Okutama Action	2017	Gp(>2)	12	[29]	Reconocimiento de acciones desde vistas aéreas
Something–Something	2018	Gp(>2)	174	[127]	Interacciones humano-objetos

Universitat d'Alacant
Universidad de Alicante

UCF_CC_50[162], BEHAVE[37], Pedestrian dataset[55], Who do What at someWhere (WWW)[330], Shanghai World Expo Dataset[411], PETS2009[95].

La mejor estrategia para obtener los mejores resultados en el campo del Aprendizaje Profundo o en la vida real es la práctica, mientras más veces se realiza una experiencia de aprendizaje más se aprende. Específicamente en el área de Visión por Computador existen muchos problemas, generalmente todos están relacionados con el procesamiento de imágenes o videos, estos elementos son los datos con los que trabajan los computadores para resolver los problemas a través de métodos, técnicas y algoritmos relacionados con el Aprendizaje Profundo. Los datos para generar los documentos de investigación pueden ser encontrados en forma organizada y tabulada en conjuntos de datos o datasets, que pueden ser públicos o privados y además tienen un tamaño enorme que hace difícil por la gran cantidad de tiempo que se requiere para el etiquetado de datos en caso de usar Aprendizaje Profundo supervisado y semi-supervisado. Además, si no fuera necesario etiquetarlos, los conjuntos de datos con buena calidad para el Aprendizaje Profundo no supervisado, también pueden ser complicados y costosos de generar. Luego de etiquetar un conjunto de datos, los modelos de Aprendizaje Profundo se pueden aplicar a los datos no etiquetados de tal manera que se tendría mejores resultados sobre ese fragmento de datos.

A veces puede ser complicado encontrar un conjunto de datos específico para usar en un problema de la vida real, problemas de Aprendizaje Máquina o incluso para experimentar, debido a que un problema puede ser muy específico con soluciones puntuales, por lo tanto distinto a otro que requiere una solución diferente.

A continuación se describen algunos de los datasets o conjunto de datos utilizados para problemas relacionados con HBA:

- Watch-n-Patch (RGB-D): Es un conjunto de datos con tipo RGB-D de actividad de personas, grabado por una cámara Kinect v2. Los videos del conjunto de datos contienen de 2 a 7 acciones de personas que interactúan con diferentes objetos (apagar monitor, llenar una tetera, tomar algo de la nevera, tomar un objeto de la mesa). La cámara utilizada tiene una resolución RGB-D (RGB: 1920x1080,

Tabla 2.4: Datasets HBA Aprendizaje Profundo

Dataset / Año / Referencia	Año	Tipo	Personas	Referencia	Areas
UCLA	2003	Crowd	–	[92]	Análisis de escena de multitudes
UCSD	2008	Crowd	>30	[55]	Conteo de multitudes
Subway benchmarks	2008	Crowd	>30	[4]	Anomalías en multitudes
QMUL	2009	Crowd	>20	[149]	Análisis de escena de multitudes
UMN Social Force	2009	Crowd	>20	[247]	Anomalías en multitudes
BEHAVE	2009	Crowd	>15	[37]	Análisis de escena de multitudes
PETS / 2009	2009	Crowd	>30	[95]	Densidad, análisis de multitudes
UCSD Ped 1	2010	Crowd	>20	[237]	Anomalías en multitudes
PASCAL VOC	2010	Crowd	>10	[97]	Conteo de multitudes
UIUC Sports Event Dataset	2010	Crowd	–	[211]	Análisis de escena de multitudes
Social Event Image Dataset (SocEID)	2011	Crowd	>100	[271]	Análisis de escena de multitudes
Caltech 10X / 2011	2011	Crowd	>100	[91]	Conteo de multitudes
Mall Dataset / 2012	2012	Crowd	>100	[58]	Conteo de multitudes
UCF / 2012 /Crowd	2012	Crowd	>100	[346]	Análisis de escena de multitudes
ImageNet	2012	Crowd	>100	[192]	Análisis de escena de multitudes
Avenue1	2013	Crowd	>100	[233]	Anomalías en multitudes
UCF_C_0	2013	Crowd	>100	[162]	Conteo de multitudes
UCF-CROWD	2013	Crowd	>100	[162]	Conteo de multitudes
CUHK	2013	Crowd	>100	[418]	Análisis de escena de multitudes
S-Hock	2015	Crowd	>100	[73]	Densidad de multitudes
AHU-CROWD	2016	Crowd	>100	[154]	Conteo de multitudes
ShanghaiTech	2016	Crowd	>100	[411]	Conteo de multitudes
UCFCC	2016	Crowd	100	[328]	Conteo de multitudes
CI	2016	Crowd	>100	[397]	Análisis de escena de multitudes
Train Station	2016	Crowd	>100	[391]	Análisis de escena de multitudes
Who do What at someWhere (WWW)	2016	Crowd	>100	[330]	Análisis de escena de multitudes
NUS-HGA	2017	Crowd	>50	[421]	Análisis de escena de multitudes
Subway	2018	Crowd	>50	[312]	Densidad de multitudes

profundidad: 512x424) y un seguimiento corporal de los esqueletos humanos basado en 25 articulaciones del cuerpo. El dataset contiene 458 vídeos con una duración total de unos 230 minutos. Actúan 7 personas que realizan actividades humanas cotidianas en 8 oficinas y 5 cocinas con un fondo complejo. En cada entorno las actividades se registran en diferentes vistas. Con el objetivo de obtener una variación en las actividades, los autores utilizan diferentes combinaciones de acciones y orden de forma natural. Algunas acciones ocurren juntas, a menudo como el tomar un objeto de la nevera y el regreso del objeto a la nevera, mientras que otras no siempre están en el mismo video[383].

- CAD-60 y CAD-120 [355]: Los videos son filmados con una cámara Kinect, que produce una imagen RGB junto con profundidades alineadas en cada pixel a una velocidad de fotogramas de 30 Hz. Produce una imagen de profundidad de 640x480 con un alcance de 1,2 m a 3,5 m. En las filmaciones se identificaron de tres a cuatro actividades comunes para cada lugar, lo que da un total de doce actividades únicas realizadas por cuatro personas diferentes en un tiempo de 45 segundos: dos hombres y dos mujeres en cinco ambientes diferentes de una casa: cuarto de baño (lavarse las manos, cepillarse los dientes, colocarse lentes de contacto), dormitorio (hablar por teléfono, beber agua, abrir el recipiente de las pastillas), cocina (cocinar-picar, cocinar-batir, beber agua, abrir el recipiente de las pastillas), sala de estar (hablar por teléfono, beber agua, hablar en el sofá, relajarse en el sofá), oficina (hablar por teléfono, escribir en la pizarra, beber agua, trabajar en la computadora). El objetivo de los autores es realizar la detección de actividad humana, es decir, el algoritmo propuesto debe ser capaz de distinguir las actividades deseadas de otras actividades aleatorias que realizan las personas. Con este objetivo se toman actividades aleatorias pidiendo a la persona que actuara de una manera diferente a cualquiera de las actividades realizadas anteriormente. La actividad aleatoria contiene una secuencia de movimientos aleatorios que van desde una persona parada, hasta una que camina y estira su cuerpo.

Universitat d'Alacant
Universidad de Alicante

- MSRDailyActivity3D [374]: Es un conjunto de datos de acción de secuencias de profundidad capturadas por una cámara de profundidad. Este conjunto de datos contiene veinte acciones: ola de brazo alto, ola de brazo horizontal, martillo, captura de mano, puñetazo hacia adelante, tiro alto, dibujar x, dibujar tictac, dibujar círculo, aplauso, ola de dos manos, boxeo lateral, doblar, patada hacia adelante, patada lateral, trotar, swing de tenis, servicio de tenis, swing de golf, recoger y lanzar una pelota. Cada acción fue repetida por diez personas durante tres veces. La frecuencia de imagen es de 15 imágenes por segundo y la resolución es de 640x480. En total, el conjunto de datos tiene 23797 cuadros de mapa de profundidad para 402 muestras de acción.

  Las posiciones de las uniones 3D se extraen de la secuencia de profundidad utilizando el algoritmo de seguimiento del esqueleto en tiempo real. Dado que no hay interacción persona-objeto en este conjunto de datos, sólo se extrae la posición 3D de la articulación.

- NTURGB+D [327]: Para recopilar los datos, se usaron sensores Kinect v2. Se recopilaron en cuatro modalidades principales de datos proporcionados por el sensor: mapas de profundidad, información de uniones en 3D, cuadros RGB y secuencias de infrarrojos.

  Los mapas de profundidad son secuencias de valores de profundidad bidimensionales en milímetros. Para mantener toda la información, los autores han aplicado una compresión sin pérdidas para cada cuadro individual. La resolución de cada fotograma de profundidad es de 512x424.

  La información de las articulaciones consiste en ubicaciones tridimensionales de 25 articulaciones principales del cuerpo para detectar y rastrear los cuerpos humanos en la escena. Los pixeles correspondientes en los marcos RGB y los mapas de profundidad también se proporcionan para cada unión y cada marco.

  El dataset contiene 60 clases de acciones en total, que se dividen en tres grupos: 40 acciones diarias (beber, comer, leer, etc.), 9 acciones relacionadas con la salud (estornudar, tambalearse, caerse, etc.) y 11

acciones mutuas (puñetazos, patadas, abrazos, etc.). Estas acciones se realizaron con 40 personas distintas, las edades entre los 10 y los 35 años.

Se utilizaron tres cámaras simultáneamente para capturar tres vistas horizontales diferentes de la misma acción. Para cada configuración, las tres cámaras se ubicaron en la misma altura pero desde tres ángulos horizontales diferentes: -45^o , 0^o , $+45^o$. Cada persona realizó una acción dos veces, una hacia la cámara de la izquierda y otra hacia la cámara derecha. De esta manera, se capturaron dos vistas frontales, una vista del lado izquierdo, una vista del lado derecho, una vista de 45^o del lado izquierdo y una vista de 45^o del lado derecho. A las tres cámaras se les asignan números. La cámara 1 siempre observa las vistas de 45^o, mientras que la cámara 2 y 3 captura las vistas frontal y lateral.

- Breakfast [194]: En este dataset se identifican 52 personas, cada una realizó 10 actividades culinarias distintas, grabadas en 18 cocinas diferentes. Una de las principales motivaciones para la configuración de grabación propuesta **en la naturaleza**, el objetivo principal es que el conjunto de datos debe reflejar lo más cercano a las condiciones del mundo real, en lo que se refiere a el seguimiento y análisis de las actividades cotidianas de personas.

  Contiene aproximadamente 77 horas de vídeo con 4 millones de fotogramas. Las cámaras utilizadas fueron cámaras web, cámaras industriales estándar (Prosilica GE680C) y una cámara estéreo (BumbleBeeR , Pointgrey, Inc). Para equilibrar los puntos de vista se grabaron videos con cámaras colocadas lateralmente. Para reducir la cantidad total de datos, todos los vídeos fueron sometidos a un muestreo a una resolución de 320x240 pixeles con una velocidad de fotogramas de 15 fps.

  Las actividades culinarias incluyeron la preparación de café, jugo de naranja, leche chocolatada, té, un tazón de cereales, huevos fritos, panqueques, ensalada de frutas, sándwiches y huevos revueltos. Este conjunto de actividades dirigidas por objetivos que las personas rea-

Universitat d'Alacant
Universidad de Alicante

lizan comúnmente en las cocinas como las antes descritas, así como actividades muy similares (por ejemplo, huevo frito comparado con preparación de huevos revueltos) para permitir una evaluación completa del sistema de reconocimiento. A diferencia de la mayoría de conjuntos de datos existentes, la actuación del actor es realizada sin guión, sin ensayo ni dirección.

- Epic-Kitchens [83]: Los datos de este dataset se capturaron utilizando un GoPro montado en la cabeza con un sistema de montaje ajustable. para controlar el punto de vista para diferentes ambientes y alturas de los 32 participantes. La cámara GoPro estaba ajustada a un campo lineal de vista, 59.94fps y resolución Full HD de 1920x1080, sin embargo algunos sujetos hizo cambios menores como FOV o resolución amplia o ultra amplia, ya que grabaron múltiples secuencias en sus hogares, y por lo tanto se apagaba el dispositivo y durante varios días. Específicamente, el 1 % de los videos fueron grabados a 1280x720. y 0.5 % a 1920x1440. También, 1 % a 30fps, 1 % a 48fps y 0.2 % a 90fps.

  La duración de las grabaciones en promedio tienen una duración de 1,7 horas, con un máximo de 4,6 horas. Cocinar una sola comida puede abarcar múltiples secuencias, dependiendo de si una se queda en la cocina, o se va y vuelve más tarde. En promedio, cada persona grabó 13,6 secuencias.

2.7. Estado del Arte, Conclusiones y Retos

Actualmente, los métodos de Aprendizaje Profundo (DL Deep Learning) están logrando grandes resultados que están revolucionando la forma de abordar los problemas de Visión Artificial, estas técnicas pueden solucionar problemas que antes no se podían resolver, incluso en algunos casos superando los resultados presentados por un ser humano, especialmente en el reconocimiento de imágenes. El objetivo de este trabajo es revisar el estado del arte del Aprendizaje Profundo (DL Deep Learning) para analizar y reconocer el comportamiento humano de grupos o multitudes (por

ejemplo, conteo de peatones [210], detección de peatones [360], personas que miran un partido de fútbol en un estadio [73], personas que juegan voleibol, cálculo de la densidad de multitudes [403], [177],[341], [38, 89], etc.). Además, analizamos el Aprendizaje Profundo para identificar el tipo de grupo o grupo de comportamiento, el nivel de abstracción (desde acciones o movimientos simples hasta comportamientos muy complejos) y las arquitecturas utilizadas para lograr este propósito. Adicionalmente se ha planteado para esta tarea una arquitectura de redes neuronales basada en CNNs principalmente y además usando la ayuda del descriptor ADV, también se ha pensado incluir otro tipo de redes que reconozcan el contexto para diferenciar cierto tipo de comportamiento humanos de acuerdo al sitio en donde se desarrolla.

De acuerdo con el objetivo anterior, las contribuciones de este trabajo son: analizar las principales redes de las Redes Neuronales Convolucionales (CNN Convolutional Neural Networks), Auto Codificadores (AE Auto Encoders) y Redes Neuronales Recurrentes (RNN Recurrent Neural Networks), cada una con algunas de sus variantes, para proponer una clasificación de los trabajos de análisis de comportamiento de grupos y multitudes de acuerdo con el número de individuos y el nivel de comprensión con respecto a la duración de la conducta detectada, así como los conjuntos de datos con información de visión para la conducta humana. Además se propone una arquitectura que detecte en forma automática el tipo de comportamiento de un grupo o multitud en función del número de personas y el contexto en el cual estas desarrollen alguna actividad.

Universitat d'Alacant
Universidad de Alicante

2.8. Propuesta de taxonomía

Tras hacer la revisión del estado del arte acerca del Análisis de Comportamiento y Actividades Humanas, proponemos una taxonomía de clasificación de las propuestas basada en dos dimensiones. En la primera dimensión se define la componente semántica o nivel de complejidad de la acción. Como se explica con mayor profundidad en la sección 2.2, los niveles de semántica (bajo, medio y alto) se estiman en función del tiempo. La segunda dimensión define el tamaño del grupo de personas (individual,

grupos pequeños, grupos grandes y multitudes). Esta segunda dimensión se explica en la sección 2.1. Visualmente se presenta la taxonomía propuesta en la Figura 2.2.

Con esta propuesta de taxonomía se pueden definir mejor zonas que actualmente se han estudiado en el estado del arte y aquellas que aun no han sido trabajadas o presentan retos no resueltos. Tener una taxonomía bidimensional permite, además, describir mejor la problemática del análisis del comportamiento humano. La madurez que ha alcanzado el área hace que describir una propuesta solo con el nivel de semántica que aborda o el tamaño del grupo de personas es insuficiente. Esta propuesta de taxonomía resuelve ese problema.

2.8.1. Taxonomía de número de personas

Varios autores han tratado de clasificar el número de personas que intervienen en una escena en donde se desarrolla algún tipo de proceder, ya sea individual o colectivo. No existe un criterio unificado debido a que solo existen descripciones, como por ejemplo, "individual o individuo" si se trata de una persona, "grupo" si existen dos o más personas juntas en una escena con un actuar coordinado, y finalmente, si existen muchas personas, por lo general más de veinte, sin especificar una cantidad exacta se puede considerar como una "multitud", especificando que también tienen una conducta coordinada en la misma escena. En esta propuesta de taxonomía no se definen umbrales claros de división entre grupo pequeño, grupo grande y multitud. Sí que lo hay, obviamente, entre un único individuo y más de uno. Aquí tampoco se define un umbral, puesto que depende del contexto (un grupo de 15 personas reunidas en una casa se considera grande, mientras que 15 personas reunidas en un parque para hacer una manifestación puede considerarse un grupo pequeño). Sin embargo, sí podemos decir que, en términos generales para los investigadores de esta área, más de 50 individuos se considera multitud.

Universitat d'Alacant
Universidad de Alicante

2.8.2. Taxonomía de comportamiento

En referencia a la taxonomía de la semántica del comportamiento en Análisis de Actividades o Comportamiento Humano no existe una clasificación universal para la complejidad que tiene un comportamiento de un individuo, grupo o multitud. En este existe similitudes entre clasificaciones que han propuesto diferentes investigadores en cuanto según los autores de [371], se definen cuatro niveles (gestos, acciones, interacción, actividad de de grupo). De forma similar, Chaaraoui et al. [49], definen tres niveles (movimiento, actividad, comportamiento), con descripciones muy parecidas a las encontradas en [371]. Existen definiciones que algunos autores comparten como las acciones atómicas, cuando se refieren a acciones que no se pueden dividir en otras más sencillas y que forman parte inicial de otras más complejas. El concepto de comportamiento, entendiéndose como el nivel mayor de complejidad en términos de actividades humanas, puede depender de muchos parámetros. Sin embargo, siguiente la clasificaciones propuestas en la literatura, es lógico utilizar el tiempo como parámetro para definir la complejidad de la actividad o comportamiento. Es por ello que en esta taxonomía que se propone utilizamos el tiempo para definir el nivel de semántica llevada a cabo por las personas en una escena. No se pueden definir estratos o niveles de semántica claros separados por valores exactos de tiempo. Sin embargo, se definen aquí niveles tres niveles, bajo, medio y alto, donde bajo serían acciones en términos de pocos segundo (andar, agacharse), nivel medio podría corresponder a actividades en la franja de segundo a varios minutos (recorrer un pasillo, interactuar brevemente entre personas, entrar en un comercio), y por último el nivel más alto de comportamiento que, en términos de tiempo, podría ir desde varios minutos hasta horas y días (hacer la comida, comportamiento de un equipo durante todo el partido de algún deporte, etc.).

2.8.3. Conclusiones sobre taxonomías estudiadas

La taxonomía propuesta basada en dos dimensiones describe un espacio de problemas que tiene varias regiones ya estudiadas y otras que aun necesitas ser analizadas por la comunidad científica para darles solución.

Universidad de Alicante

La figura 2.2 muestra una representación visual de este espacio, donde el eje x representa el número de personas desde una persona, grupos pequeños y grupos grandes, por último multitudes. En el eje y se muestran tres niveles diferentes de semántica (bajo, medio, alto). Como ya se ha dicho, aquí definimos el nivel de semántica en base al tiempo de la acción. Se asume que esto va relacionado directamente con el tiempo que dura un comportamiento. Adicionalmente, se han marcado tres zonas en el plano cartesiano (A,B,C,D), estás zonas son la representación de diferentes tipos comportamientos según la complejidad de estos y el número de personas en escena. A continuación se describe el significado de las zonas A, B, C y D: 1) Zona A describe comportamientos con bajo nivel semántico y corto tiempo de duración desde una sola persona hasta un grupo pequeño, 2) Zona B incluye comportamientos más sencillos que en A de grupos pequeños, 3) Zona C incluye comportamientos de muchas personas o también llamadas multitudes con un tiempo y grado bajo de semántica, es decir comportamientos muy sencillos. Finalmente, la Zona D incluye los problemas no resueltos del tema, y datasets que no se han construido todavía, es decir, comportamientos muy complejos, con un alto grado de semántica, y tiempos de duración de horas o días.

Para resumir acerca de los trabajos que los investigadores han publicado sobre cada una de lo que para fines de esta tesis hemos llamado zonas antes descritas, podemos concluir que pueden existir más zonas, y por lo tanto existirán más problemas para los cuales se pueda proponer soluciones. Sin embargo, se citan algunos trabajos que encajan por sus características en nuestra definición.

Zona A: en este espacio se ubican pocas personas que ejecutan acciones de corta duración o baja componente semántica, como caminar, saltar, aplaudir y otros similares [364, 364]. También se incluyen Ambientes de Vida Asistidos (AAL), ya que la investigación aborda los comportamientos individuales que tengan mayor tiempo de duración o componente semántico alto pero con pocas personas, como los trabajos de [159, 413], que proponen aplicaciones que apoyan la calidad de vida monitoreando el comportamiento de personas adultas.

Zona B: esta zona corresponde a un mayor número de personas podemos

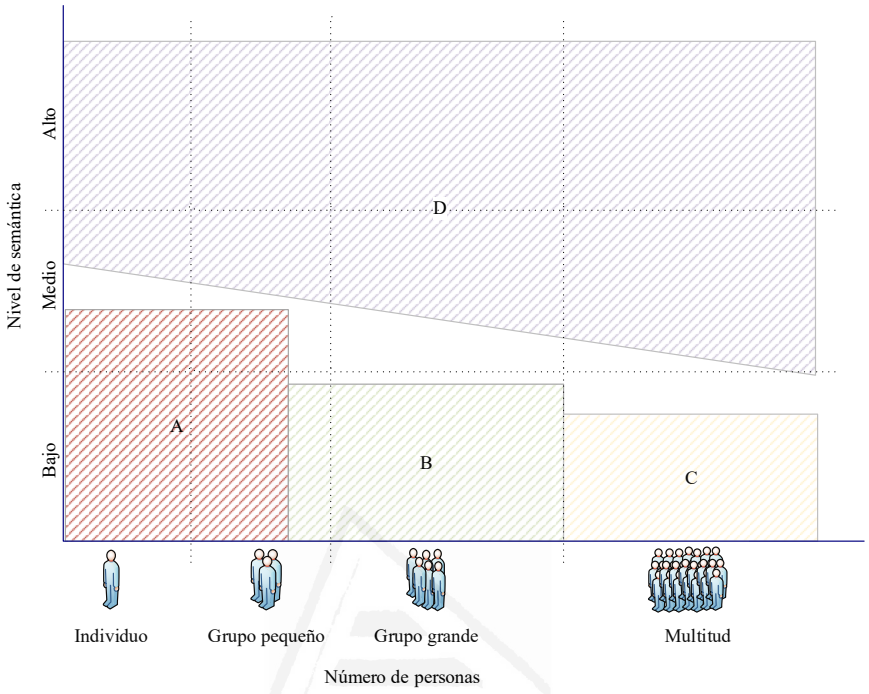


Figura 2.2: La figura gráficamente la taxonomía bidimensional propuesta para clasificar el espacio de problemas del análisis de comportamiento humano. En el eje x se representa el número de personas. En el eje y se incluye los niveles de semántica gráficamente la taxonomía bidimensional propuesta para clasificar el espacio de problemas del análisis de comportamiento humano.

Universidad de Alicante

citar ejemplos de comportamiento que se limitan a actividades o acciones muy concretas y sencillas, usualmente, de corta duración o baja componente semántica como una jugada de fútbol [166, 283], jugadas de volleyball [24, 335], grupos de personas caminando en un solo sentido o grupos de personas conversando[63, 86], detección de peatones en una calle [112, 275], violencia entre grupos [133, 213], entre otras.

Zona C: aquí se ubican los trabajos relacionados con multitudes se limitan específicamente a tareas como contar personas y calcular densidad de una multitud [341, 403], o detectar movimientos de una masa de personas o colisiones de multitudes [260, 329].

Zona D: en esta zona según el análisis del estado del arte realizado en

apartados anteriores no se ha encontrado trabajos o datasets que tengan características como un alto componente semántico que incluya formas de actuar de muchas personas, como por ejemplo el desempeño de un equipo de fútbol durante toda una temporada. Tampoco existen datasets que incluyan comportamientos de larga duración, de horas o días por ejemplo.

Universitat d'Alacant
Universidad de Alicante

Capítulo 3

Propuesta de arquitectura

En este capítulo se propone una arquitectura de aprendizaje para el reconocimiento de actividades de grupo basada en la propuesta general del capítulo 1. La arquitectura está compuesta por un primer bloque capas de describir el movimiento presente en una escena, y un segundo bloque encargado de la clasificación del mismo en actividades grupales. El modelo arquitectural propuesta se instancia en esta tesis para resolver el problema de la clasificación de múltipes actividades (*multi-class*) y de comportamientos anómalos (*one-class*).

El descriptor de movimiento se fundamenta en el *Activity Description Vector* (ADV), proponiendo una variante de este, el D-ADV, con formato de imágenes que permite ser utilizado como entrada para arquitecturas de aprendizaje profundo. El bloque de clasificación propuesto considera múltiples flujos de información: los flujos correspondientes al D-ADV y al contexto donde se desarrolla la escena.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

3.1. Introducción

Actualmente existen diversas aplicaciones relacionadas con el problema del análisis del comportamiento humano (HBA), destacando la videovigilancia [126] y la vida asistida por el entorno (AAL) [50]. En general, el comportamiento humano está relacionado con cómo las personas desarrollan sus actividades en una escena, pero además esa actividad puede estar relacionada con la interacción de estas con elementos adicionales como objetos o con otras personas. De esta forma, el comportamiento se puede analizar a través de movimientos simples o complejos que impliquen interacciones persona-persona o persona-objetos [39]. Como ya hemos analizado en capítulos anteriores, se suelen utilizar técnicas de visión por computador y métodos de aprendizaje automático para hacer frente a los diferentes aspectos del problema.

Independientemente de la aplicación, el proceso tradicional para abordar el HBA se divide en dos etapas principales: la extracción de características de la escena y el análisis de las mismas. La etapa de extracción de características suele ser clave ya que está estrechamente relacionada con la capacidad que tiene el método posterior de clasificar correctamente el comportamiento. Es por ello que el proceso de diseño de una solución requiere, en la mayoría de ocasiones, de un conocimiento del dominio del problema a resolver. Esta etapa es heterogénea y diferente para cada uno de los métodos propuestos en la literatura. La extracción de las características de una secuencia de imágenes incluye detectar y seguir la Región de Interés (ROI) con técnicas de segmentación y tracking. La segmentación de imágenes se ha abordado ampliamente mediante técnicas tradicionales como [258, 318] y, recientemente, se han propuesto enfoques utilizando aprendizaje profundo (Deep Learning, DL) [109]. En cuanto al tracking, se han propuesto varios trabajos para estudiar el movimiento en escena. Cabe destacar el trabajo de Ojha et al. donde se realiza una extensa revisión de diferente técnicas de tracking [265], y más recientemente, el trabajo de Yazdi y Bouwmans que realizar una revisión de métodos de incluyendo enfoques de aprendizaje profundo [389]. Para el análisis de las características de la escena, tradicionalmente se ha abordado a través de técnicas de aprendiza-

je máquina, alimentando a los clasificadores con las mismas (velocidades, trayectorias, ROIs, etc.) para estimar el comportamiento, la acción, la actividad, etc. [253]. En la actualidad, tal como se ha analizado en el capítulo 2, las propuestas de clasificación utilizan técnicas de aprendizaje profundo.

Como hemos comentado anteriormente, entre las características de las escena, resulta un aspecto fundamental el movimiento de los sujetos en la misma. Esto es, desde un punto de vista de la Física, el cambio de posición que sufre un cuerpo en el espacio, tomando en cuenta el tiempo y un punto de referencia donde está ubicado el observador del fenómeno. Esto significa, que las propiedades de todo movimiento dependen del sistema del punto de vista desde donde se lo observe. En el presente trabajo uno de los aspectos que se considera es el análisis cambios de posición del cuerpo humano o sus partes. Adicionalmente, se debe tener en cuenta que de todos estos movimientos básicos pueden ser ejecutados por varias personas simultáneamente o en combinaciones. Entender el significado de ese conjunto de actos realizados por las personas, es una tarea difícil que aun supone un reto a nivel científico.

En la literatura se ha estudiado ampliamente el movimiento de individuos, proponiendo descriptores del mismo a diferentes niveles semánticos dependiendo de su duración: acción, actividad y comportamiento. Respecto a descriptores del movimiento, destacamos el Activity Description Vector (ADV) presentado en el trabajo [20] y sus diferentes versiones para clasificar comportamiento [18, 23]. En esta tesis, se propone una nueva variante que es independiente del número de individuos o grupos y que permite describir con dos imágenes el movimiento existente en la escena a partir del cálculo del flujo óptico [31, 99, 116].

Por último, en la propuesta desarrollada en esta tesis, se propone la utilización del contexto como mecanismo que permite discriminar entre comportamientos que pueden ser normales o no dependiendo de cuándo o dónde se realicen, de los sujetos que aparecen en la escena, etc. De la misma forma, podrían ayudar a concretar con mayor precisión las actividades que se llevan a cabo en la escena. En otras palabras, para fines de entender la aplicación del término contexto en la presente investigación debemos describir que se refiere a todo lo que rodea, sea en forma física o simbólica,

a un acontecimiento. En esta propuesta de solución se plantea también utilizar aprendizaje máquina para determinar el entorno.

3.1.1. Clasificación one-class

A los efectos dela presente propuesta, partimos del supuesto que la mayoría de los elementos del dataset de entrenamiento son considerados como datos "normales" (para este análisis, el término "normal" se define como no anómalo y no relacionado con una distribución Gaussiana). El objetivo es aprender de los datos que se describen a detalle y con precisión el significado de "normal". Las desviaciones o los datos que difieran de esta definición se considerarán como anomalías o anormales. Esta descripción se denomina clasificación de una clase (OCC)[255]. Se basa en el Problema del Círculo Mínimo de Recubrimiento [356], que trata de encontrar la esfera con el radio más pequeño que contiene un conjunto de puntos, para un conjunto finito de puntos. En este caso, consiste en aprender el límite que define la única clase normal conocida, y asumir que todo lo demás es otra clase anormal desconocida.

La combinación de OCC y redes neuronales se ha llevado a cabo anteriormente en varios trabajos, como el caso de Chalapathy et al. [52] que propuso un modelo de red neuronal de una clase (OC-NN) para detectar anomalías en conjuntos de datos complejos. El OCC se encuentra en muchas aplicaciones de Visión por Computador del mundo real como: detección de anomalías [52], clasificación profunda [277, 306], detección de novedades [1, 313], y otras.

En situaciones reales el porcentaje de eventos anormales es significativamente menor comparado con eventos que son considerados normales. Por ejemplo en muchos de los casos el análisis de comportamiento humano se comporta como un problema de clasificación una sola clase (OCC del inglés, One Class Classification) que consiste en aprender la frontera que define una sola clase conocida, y asumiendo que todo lo demás algo anormal. Más concretamente, se trata de definir el hiper-espacio de parámetros que corresponden a una de las clases y asumir que el resto del espacio pertenece a la segunda clase. Para ello, se define una hiper-esfera con radio R que agrupará el comportamiento normal. La base matemática utilizada

es la del problema del círculo de recubrimiento mínimo [356], que se basa en encontrar la esfera de menor radio que contiene un conjunto de puntos, para un conjunto finito de puntos.

En la figura 3.1 se muestra una descripción gráfica del problema de la Clasificación Una-Clase, que toma como base teórica **El problema del círculo de cobertura mínima**, donde el conjunto de entrenamiento contiene una sola clase, representada con triángulos verdes. Estos elementos están en la hiper-esfera definida por su radio R y centro C. La segunda clase, que se dispone de ella en la fase de clasificación, se representa con cuadrados rojos y está fuera de la hiper-esfera. En nuestro caso, la clase interna es el comportamiento normal y la clase externa corresponde al comportamientos anormales.

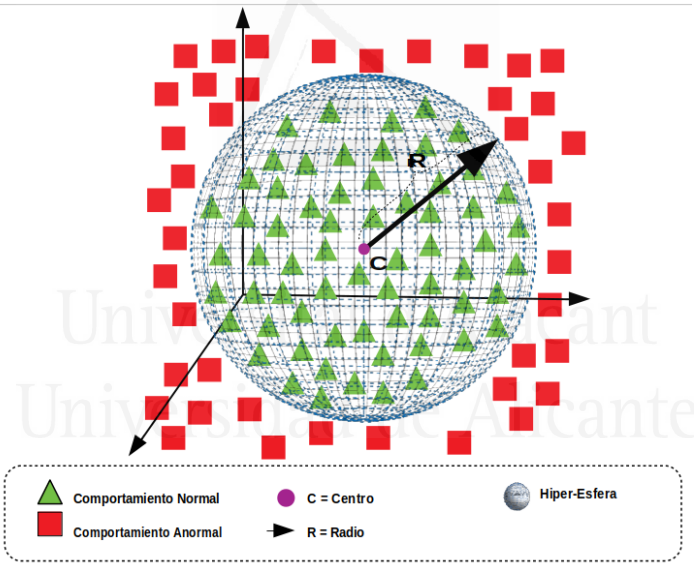


Figura 3.1: Los componentes que representan la Clasificación Una-Clase son: Ejes, que representan n dimensiones, cuadrados rojos (comportamientos anormales), triángulos verdes (comportamiento normal), hiper-esfera (envuelve los comportamientos normales), radio R de la hiper-esfera, centro C de la hiper-esfera.

El problema del círculo de cobertura mínima, base de la OCC, postula que el problema de agrupar un conjunto de datos en un círculo puede ser representado en un espacio de dimensión n, con una n-esfera o hiper-esfera que contenga a todos los puntos agrupados en un conjunto específico. Este

problema fue propuesto por primera vez en 1857 por el matemático inglés James Joseph Sylvester [356]. Estos puntos representan los valores normales con características definidas, mientras que los puntos que no cumplan las características antes definidas serán considerados como anomalías.

Adicionalmente un enfoque relacionado con esta forma de solucionar este problema existe una propuesta alternativa que consiste en calcular las elipsoides más pequeñas de los conjuntos de puntos en dimensiones extendidas, con el objetivo de ampliar la cobertura de elementos que posiblemente queden excluidos de una esfera [382].

En el problema de clasificación OCC solo se puede definir una clase que tiene una frontera de decisión debe rodear alrededor de los datos considerados como normales. También es necesario decidir cuáles son las características debe tener la frontera entre los objetos de clase normales y anormales.

Para solucionar este tipo de situaciones, en la mayoría de casos, los métodos detectan e identifican las principales características de los valores normales. Conceptualmente, la solución más simple para la detección de valores anormales es generar datos anómalos en torno al objetivo establecido, es decir, los valores normales. Esto sería como crear valores que son representados por los cuadrados rojos mostrados en la Figura 3.1.

Universitat d'Alacant
Universidad de Alicante

Los autores de [183] presentan un enfoque unificado de OCC proponiendo tres aspectos importantes sobre este tipo de método de clasificación. 1) Disponibilidad de datos de aprendizaje: Aprender sólo con datos positivos o datos no etiquetados y/o alguna cantidad de ejemplos atípicos. 2) Metodología utilizada: Algoritmos basados en máquinas vectoriales de soporte de una clase (OC-SVM) o metodologías distintas a estas. 3) Dominio de aplicación: OCC aplicado a la clasificación de textos o en otros campos o dominios de aplicación.

A continuación se describen las características recomendadas que debería tener una método OCC para mejorar el rendimiento y resultados, las propiedades se refieren a la robustez de los valores anormales, la incorporación de valores anómalos conocidos, los parámetros mágicos de ajuste, y los requerimientos de cálculo.

- **Robustez a valores anómalos:** En este método se asume que el

conjunto de entrenamiento contiene elementos que poseen características de una distribución objetivo, es decir, se entrena con valores considerados como normales. Este es uno de los métodos en que se optimiza los elementos más parecidos entre si o también se considera la cercanía de los mismos hacia el umbral, estando cerca al borde o límite son los mejores candidatos para ser considerados como atípicos.

- **Inclusión de valores anómalos conocidos:** En este caso se considera la existencia de valores anómalos para realizar ajustes en la descripción del conjunto objetivo, para incluir esta información en un método el conjunto de entrenamiento debe ser lo suficientemente flexible como para incluir o rechazar un valor anormal.

- **Parámetros mágicos:** Un aspecto importante en la operación de un método es la configuración de parámetros elegidos por el usuario, lo que incluye la elección de un buen conjunto de entrenamiento. Este factor influye de forma significativa en los resultados, ya que si no el resultado puede ser diferente al esperado. La elección adecuada de parámetros iniciales se conoce como "parámetros mágicos". Para describir detalles adicionales sobre estos parámetros podemos pensar varios ejemplos como el número de capas, las neuronas a utilizar. Muchas veces estos parámetros en se obtienen tras experimentar mediante ensayo y error.

- **Requerimientos de procesamiento y almacenamiento:** Una consideración final recomendada es la capacidad de procesamiento y almacenamiento que debe tener el hardware donde se procesan los datos con los métodos. Se conoce que cada vez existen ordenadores con más capacidad de procesamiento, y de forma unísona, la demanda de recursos por parte de los procesos y métodos es cada vez mayor para experimentar con grandes volúmenes de datos.

Universitat d'Alacant
Universidad de Alicante

3.1.2. Clasificación multi-class

Es un proceso encargado de clasificar dos o más clases; por ejemplo clasificar de entre varias imágenes cuáles de estas corresponden a comportamientos como caminar, saltar, unirse a un grupo, separarse de un grupo (asignar múltiples categorías a las observaciones). La clasificación multi-clase toma como punto de partida que cada muestra tiene asignada sola una etiqueta, es decir, no se pueden etiquetar dos comportamientos con na misma etiqueta a la vez. Un ejemplo de este tipo de clasificación es el reconocimiento de caracteres de números escritos a mano (aquí las clases son de 0 a 9).

El problema de la clasificación multi-clase puede descomponerse en varios problemas de tipo binario, de tal modo que una solución eficiente luego de la descomposición es usar clasificadores binarios. Existen métodos que nos permiten utilizar en clasificación multi-clase, para realizar este procedimiento existen dos estrategias conocidas: Uno-contra-todos (OVA de inglés one-versus-all) y uno-contra-uno (OVO del inglés one-versus-one) [10].

One-versus-all (OVA) [324]: esta es una de las más usadas, consiste en tomar una clase específica y discriminar esa clase del resto de clases. Transforma un problema de c clases en c problemas binarios. Estos problemas representados en dos clases se construyen usando los ejemplos de una clase como ejemplos verdaderos y los ejemplos del resto de clases como los ejemplos falsos.

One-versus-one [105]: esta técnica consiste en entrenar un clasificador separado para cada par de clases del problema. El método transforma un problema de c clases en $c(c-1)/2$ problemas binarios $< i, j >$, uno por cada conjunto de clases i, j, donde $i,\ j = 1, ..., c$ e $i < j$. El clasificador binario para un caso $< i, j >$ es entrenado con clases i y j, mientras que las muestras de las clases $k \neq i,\ j$ no son tomadas en cuenta. Para resolver el problema de clasificación binaria se han propuesto diferentes métodos. Sin embargo, la clasificación multi-clase se trata de un caso especial ya que inicialmente se introdujeron muchos algoritmos para resolver problemas de clasificación binaria. El problema de la clasificación multi-clase puede resolverse extendiendo la técnica de clasificación binaria en el caso

de algunos algoritmos como: neural networks (NN) [35, 93], decision trees (DT) [231, 252, 352], k-Nearest Neighbor (kNN) [28, 33], Naive Bayes (NB) [185, 300, 407], y Support Vector Machines (SVM) [44, 74, 151]:

- **Neural Networks (NN) [35, 93]:** Las redes neuronales con múltiples capas entradas múltiples proporcionan una extensión al problema de la multi-clase. La diferencia está en que en lugar de tener una neurona en la capa de salida, con salida binaria, se utiliza n neuronas, una para cada clase.

- **Decision Trees (DT) [231, 252, 352]:** El árbol trata de inferir una división de los datos de entrenamiento basado en valores de las características disponibles para producir una posible generalización. Cada nodo de la hoja corresponde a una etiqueta de clase. Una nueva muestra se clasifica siguiendo un camino desde el nodo raíz hasta el nodo de la hoja, donde en cada nodo es una prueba de alguna característica de esa muestra. El nodo de la hoja alcanzado es considerado la etiqueta de clase. El algoritmo puede manejar problemas de clasificación binaria o multi-clase.

- **k-Nearest Neighbors (k-NN) [28, 33]:** Se considera uno de los algoritmos más antiguos de clasificación no paramétrica. Para clasificar una muestra desconocida, se mide la distancia desde una muestra específica a todas las demás muestras de entrenamiento. Se identifican las distancias k más pequeñas, y la clase más representada en estas k clases se considera la etiqueta de clase de salida. El valor de k se determina normalmente utilizando un conjunto de validación o utilizando una validación cruzada.

- **Naive Bayes (NB) [185, 300, 407]:** Este algoritmo asume la presencia o ausencia de alguna característica particular de los datos analizados. Una ventaja es que solo se necesita una muestra pequeña de datos de entrenamiento para estimar los parámetros (medias y varianzas) necesarias para la clasificación. Como las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza. Es-

Universidad de Alicante

te enfoque es naturalmente extensible al caso de tener más de dos clases.

- **Support Vector Machines (SVM) [44, 74, 151]:** Su fundamento parte de la idea de maximizar el margen, es decir, maximizar la distancia mínima desde el hiper-plano de separación hasta el ejemplo más cercano. Este algoritmo soporta la clasificación binaria, sin embargo puede adaptarse a clasificación multi-clase. Se conocen como vectores, en lugar de puntos, porque dichos puntos tienen tantos elementos como dimensiones pueda tener el espacio de entrada de datos. Es decir, estos puntos multi-dimensionales son representados con vector de dimensión n.

3.2. Modelo arquitectural

A pesar del gran esfuerzo realizado por la comunidad científica para mejorar el análisis automático del comportamiento humano, todavía existen muchos retos abiertos [189]. Como hemos visto en el capítulo 2, las propuestas actuales para el análisis del comportamiento humano se basan en soluciones íntegras de aprendizaje profundo. Es decir, una misma red neuronal alimentada directamente con la secuencia de imágenes de la escena es capaz de clasificarla. Esto es debido a que la red es capaz de aprender, por sí sola, tanto las características de la imagen, o la secuencia, que mejor definen la solución al problema, como los parámetros de la red que mejor separan esas características para determinar la correcta clasificación. De esta forma, el aprendizaje se realiza de todo el proceso proporcionando unos rendimientos muy altos pero careciendo, en ciertos casos, de generalidad. Por otro lado, los resultados suelen estar condicionados por el volumen de datos disponibles: a mayor volumen de datos, mejores resultados. Esto es, un conjunto de datos pequeño no asegura la bondad de las soluciones propuestas. Por último, entre los retos actuales, la falta de generalidad en las propuestas actuales, en términos de número de individuos en la escena (es decir, de grupo de dos o más personas hasta multitudes), dificulta el establecimiento de una arquitectura de referencia para definir cómo abordar diferentes casos utilizando propuestas similares.

Universitat d'Alacant
Universidad de Alicante

Con el objetivo de avanzar en la solución de estos problemas, en esta tesis se propone un modelo arquitectural de visión por computador capaz de aprender y reconocer las actividades de grupos de personas utilizando los movimientos de los mismos en la escena. El modelo consta de dos grandes bloques (ver Figura 3.2): la representación local del movimiento y la clasificación. De esta forma, se intenta aunar las mejores capacidades del aprendizaje (profundo) junto con descriptores robustos que permitan generalizar los datos de entrada. Para ello, se parte de la hipótesis de que la utilización de la apariencia de movimiento en la escena y el contexto podría utilizarse para establecer un modelo arquitectural para resolver la clasificación de actividades de grupos y la detección de anomalias independientemente del número de personas que intervienen.

En el estado del arte se ha demostrado que el uso de descriptores de trayectoria mejora la calidad de la estimación del comportamiento real, ya que proporciona un simple y elevado nivel de comprensión de las actividades de grupos complejos. Por tanto, el primer bloque del modelo (Fig. 3.2) es capaz de describir el movimiento en la escena a partir de representaciones locales del movimiento en una región de la misma. Está basado en el Vector Descriptor de Actividad (Activity Descriptor Vector, ADV), un descriptor capaz de representar la información de la trayectoria de una secuencia de imágenes como una colección de los movimientos locales que ocurren en regiones específicas de la escena. El ADV [18, 21] mostró un muy buen desempeño, independientemente del uso de diferentes clasificadores, en la descripción de las actividades relacionadas con individuos. También, demostró sus capacidades de predicción no sólo para predecir el comportamiento a partir de nuevas entradas, sino también de detectar el comportamiento utilizando una porción de la entrada, para detectar en forma anticipada el comportamiento realizado por una persona en una escena [23]. Por último, una variante de ADV también se especificó para analizar el comportamiento del grupo (G-ADV) en [19] mostrando también excelentes resultados. El G-ADV se calcula a partir de la trayectoria descrita por el grupo y por los individuos que lo forman. En concreto, utiliza tres componentes diferentes: la trayectoria seguida por el grupo, la coherencia de las trayectorias individuales en el grupo y, finalmente, las relaciones de

Universitat d'Alacant
Universidad de Alicante

movimiento entre los diferentes grupos de la escena.

El segundo bloque del modelo propuesto (Fig. 3.2) permite en sí realizar una clasificación de las actividades del grupo. En las investigaciones de visión por computador, las redes neuronales profundas han evolucionado para ser utilizadas de manera consistente debido a sus buenos resultados, como hemos comentado anteriormente. Los métodos de aprendizaje profundo han ganado superioridad sobre otros en el campo del reconocimiento de imágenes y la clasificación en imágenes individuales y secuencias, como el reconocimiento de acción basado en LSTM [101, 229, 347, 347], arquitecturas basadas en múltiples flujos para el reconocimiento del comportamiento [138, 175, 375], basadas en esqueletos para el reconocimiento del comportamiento humano [90, 235, 280], etc. En esta tesis, el modelo arquitectural se instancia para resolver el problema de aprendizaje con una sola clase (OCC) con el objetivo de detectar comportamientos anormales así como para la clasificación de múltiples actividades (MCC).

Por lo tanto, el principal objetivo al proponer esta tipo de arquitectura es combinar las ventajas del ADV para representar actividades basadas en las trayectorias de los individuos, y el enfoque de aprendizaje profundo, cuando sea necesario, introduciendo la descripción del movimiento como una variante del ADV. A partir de este objetivo, la contribución de esta arquitectura es la mejora de la generalidad y el rendimiento en la clasificación de las actividades de grupo. Además, el uso de la variante ADV permite entrenar el modelo utilizando pequeños datasets etiquetados en lugar de utilizar grandes volúmenes de datos. En lugar de aprender de las secuencias de imágenes en bruto, el uso de la variante ADV permite a la red aprender las características de ese descriptor, reduciendo el espacio de la solución.

3.2.1. Descriptor local de movimiento

El primer bloque de la arquitectura tiene como objetivo extraer una representación de los movimientos que se producen en la escena. La propuesta que se realiza en esta tesis es una variante del descriptor ADV que permite ser utilizada en sistemas de aprendizaje profundo basados en imagen. Concretamente, se propone el D-ADV que permite describir una

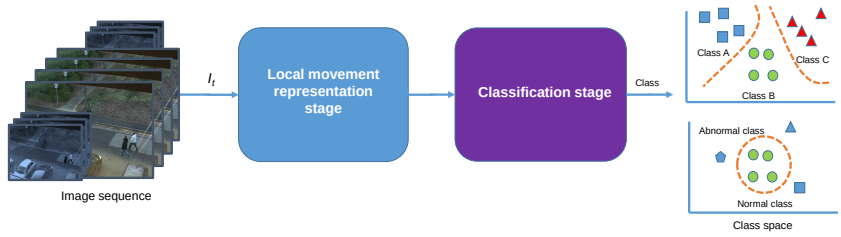


Figura 3.2: Modelo arquitectural propuesto en dos etapas: etapa de representación de movimientos locales y etapa de clasificación.

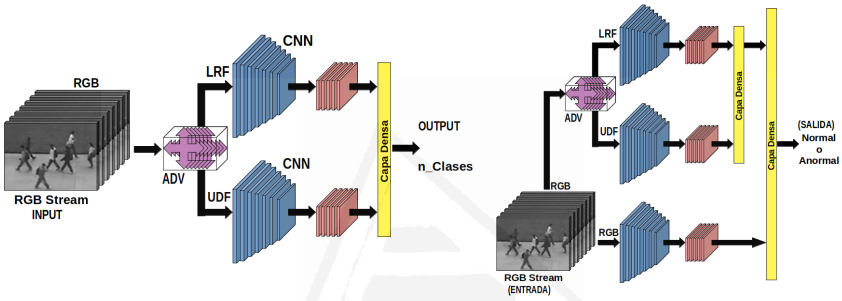


Figura 3.3: En esta figura se se resume la arquitectura propuesta en donde las líneas y flechas de color negro representan el flujo de datos en dos y tres streams, en la parte izquierda para MCC, y la derecha para OCC.

Universitat d'Alacant
Universidad de Alicante

escena con imágenes de movimientos locales en regiones de la misma. En esta sección se detalla este descriptor y se resume, con el objetivo de completitud, el descriptor ADV.

3.2.1.1. Activity Description Vector (ADV)

La propuesta de ADV [18, 21] consiste en un método de representación que toma como referencia la escena o terreno en donde una persona se mueve como un modelo geométrico básico para describir su trayectoria considerando que los datos del escenario no tienen perspectiva. Por lo tanto, el espacio de valores debe ser perpendicular al enfoque de la cámara. En el supuesto caso que la cámara no esté ubicada en el techo, cualquier información contenida en el plano de la imagen capturada desde la cámara estática debe ser transformada en el plano correspondiente que se ajus-

ta al suelo a través de una homografía, H (Eq. 3.1). La transformación proyectiva permite considerar todo el espacio de movimientos de las personas en un espacio Euclideo. Entonces, cualquier punto p_i en la imagen es transformado en un punto p_g en el plano del suelo G.

$$p_g = H \cdot p_i \tag{3.1}$$

Como sólo nos interesa la información espacial de la trayectoria para obtener una simple representación para analizar el comportamiento, la información necesaria para seguir los objetos en la escena es la posición de un individuo en la escena. Se establecen, así, una lista de sucesivos puntos LTP en G.

$$LTP = \{p_1, p_2, p_3, \cdots, p_n\} \tag{3.2}$$

ADV utiliza el número de ocurrencias de una persona en un punto específico del escenario y sus movimientos locales en él. Este método describe el "suelo"del escenario, G, en celdas como una cuadrícula, C, para discretizar el entorno. Cabe mencionar que para minimizar errores producidos durante fases de segmentación y seguimiento, produciendo ruido en las posiciones, se muestrea tomando datos cada t tiempo y modelando la trayectoria con una curva *spline*. De ese modelo se calculan puntos intermedios que forman una trayectoria más suave y continua. Con estos nuevos datos, no se calcula la trayectoria completa, sino que se tiene en cuenta el movimiento entre puntos consecutivos en cada uno de los cuatro ejes considerando desplazamiento y dirección. Por lo tanto, calculamos las cuatro direcciones: arriba (U), abajo (D), izquierda (L), derecha (R), como el producto vectorial del vector de desplazamiento entre puntos en LTP, p_i y p_{i-1}, y el vector normal para cada eje (ver figura 3.4).

Los distintos desplazamientos se calcula de la siguiente forma:

$$U(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otros... \end{cases} \tag{3.3}$$

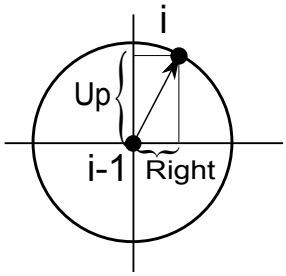


Figura 3.4: Representación en el eje x e y de los movimientos Arriba (U) y Derecha (R) en un desplazamiento entre los puntos p_{i-1} y p_i. Figura tomada de [18]

$$D(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otros... \end{cases} \tag{3.4}$$

$$L(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otros... \end{cases} \tag{3.5}$$

$$R(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otros... \end{cases} \tag{3.6}$$

Por otro lado, el ADV considera la Frecuencia, F, como el número de ocurrencias de una persona ubicada en un punto específico de G. Esto es, el número de frames en los que la persona ha estado en un lugar específico. F contiene información sobre la trayectoria espacial de una persona pero sin considerar los movimientos en sí. Finalmente, el plano del suelo G se muestrea espacialmente en una matriz C de celdas $m \times n$ de modo que los puntos transformados p_g y las funciones de frecuencia y movimientos de la misma, están en una de las celdas de la matriz C. Cada celda describirá

la actividad ocurrida en esa región de la escena considerando el vector de valores más relevantes, el *Activity Description Vector* (ADV_C). Este vector estará compuesto por la frecuencia y los movimientos U, D, L y R de todos los puntos del plano dentro de una celda:

$$ADV_C = \langle F, U, D, L, R \rangle \tag{3.7}$$

Por lo tanto, dentro de una celda particular, se calculan los histogramas acumulativos de los movimientos U, D, L, R y F para los puntos sobre G de la celda $c_{i,j}$ de C. Sea $u \times v$ el tamaño real del escenario, dividido en $m \times n$ celdas y, $p_{k,l}$ el punto situado en la posición k y l del espacio G, cada ADV de una celda es:

$$\begin{gathered} \forall c_{i,j} \in C \wedge \forall p_{k,l} \in G / i = \left\lfloor \frac{k \times m}{u} \right\rfloor \wedge j = \left\lfloor \frac{k \times n}{v} \right\rfloor \\ ADV_{i,j} = \Big(\sum F(p_{k,l}), \sum U(p_{k,l}), \sum D(p_{k,l}), \sum L(p_{k,l}), \sum R(p_{k,l}) \Big) \end{gathered} \tag{3.8}$$

Así, para un escenario de $u \times v$, dividido en $m \times n$ celdas, cada dato en el ADV será de $5 \times m \times n$. La Figura 3.5, extraída del artículo [19], muestra un ejemplo de representación del ADV.

Universitat d'Alacant
Universidad de Alicante

3.2.1.2. D-ADV

El D-ADV utiliza una secuencia de imágenes como conjunto de entrada. A diferencia del ADV, el D-ADV no se basa en los movimientos específicos e individuales de una persona en la escena y las ocurrencias en ella (es decir, la frecuencia). Considera el movimiento aparente de los individuos en la escena visual y la apariencia de los mismos asumiendo un fondo específico. Para los primeros, el cálculo del flujo óptico es la etapa inicial del proceso. Calcula el flujo óptico entre dos frames consecutivos $(t, t + \delta t)$ de la secuencia utilizando el método diferencial como el método más utilizado [180]. Se basa en el supuesto de la constancia del brillo de la imagen: dada una secuencia de video, la intensidad del pixel (x, y) del frame t, $I_t(x, y)$, permanece igual a pesar de pequeños cambios de posición y de período de tiempo. Si (δx,δy,δt) se expresa como un pequeño cambio del movimiento, y asumiendo la constancia del brillo y la expansión como serie de Taylor,

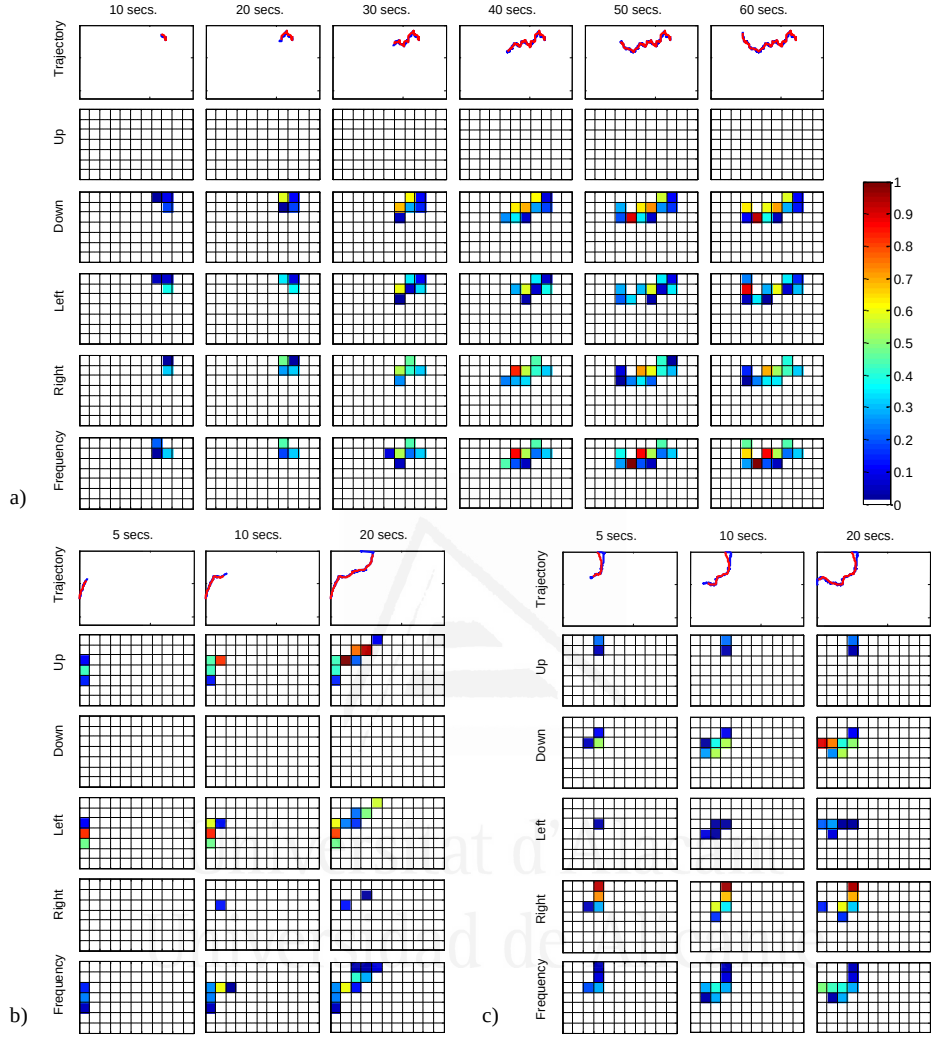


Figura 3.5: Ejemplo de representación del ADV para secuencias de las clases *Window Shopping* (a), *Shop enter* (b) y *Shop exit* (c) del dataset CAVIAR para distintas observaciones de tiempo. La primera fila muestra el original (azul) y los datos suavizados con el *spline* (rojo). El resto de filas muestran los movimientos U, D, L, R y F

puede ser expresado y aproximado como se lo describe en [40, 180]):

$$I_{t+\delta t}(x+\delta x, y+\delta y) \approx I_t(x,y) + dato\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t$$

, resolviendo y dividiendo el segundo término a lo largo del mismo por δt, es posible obtener:

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}U + \frac{\partial I}{\partial y}V + \frac{\partial I}{\partial t} \approx 0$$

donde $U = \frac{\delta x}{\delta t}$ y $V = \frac{\delta y}{\delta t}$ son los dos componentes del flujo óptico en t.

En este caso, los puntos p_i utilizados para calcular los componentes de ADV como en las ecuaciones 3.3, 3.4, 3.5 y 3.6 fueron extraídos de puntos consecutivos de una trayectoria en un plano. Si asumimos la imagen como un plano del suelo y una cámara estática (es decir, el movimiento aparente sólo es generado por los individuos en la escena, no por el observador que este caso sería la cámara), la diferencia en la trayectoria $(p_i - p_{i-1})$ podría aproximarse como los derivadas de los pixeles en x y y para el frame t como $(p_i - p_{i-1}) \approx (\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}) = (U, V)$. Además, como los movimientos se consideran en cada eje, los movimientos U y D están estrechamente relacionados con la componente V del flujo óptico, y las componentes L y R relacionadas con la componente U. En consecuencia, las componentes podrían calcularse como:

$$\begin{aligned} U(I_t) &= \begin{cases} -V_t & if \;\; V_t < 0 \\ 0 & otros... \end{cases} \\ D(I_t) &= \begin{cases} V_t & if \;\; V_t > 0 \\ 0 & otros... \end{cases} \end{aligned} \tag{3.9}$$

Universitat d'Alacant
Universidad de Alicante

Con respecto al componente F, se estima que:

$$F = |I - B| > 2 \cdot std(I - B), \tag{3.10}$$

donde B es el fondo calculado a partir de una secuencia de imágenes, y std es la desviación estándar de la diferencia entre un frame y el fondo. El primer plano se extrae para obtener los individuos que aparecen en la escena independientemente de si se están o no moviéndose.

Esta etapa acumulativa es la responsable de calcular el ADV en una determinada celda, tal como se presenta en la ecuación 3.7. Por un lado, el

desplazamiento acumulativo es responsable de los parámetros L, R, U y D, y el primer plano acumulativo es para la componente F. La acumulación se considera para un conjunto de frames consecutivos. Para describir este proceso en forma gráfica se explica el flujo de datos en la figura 3.6.

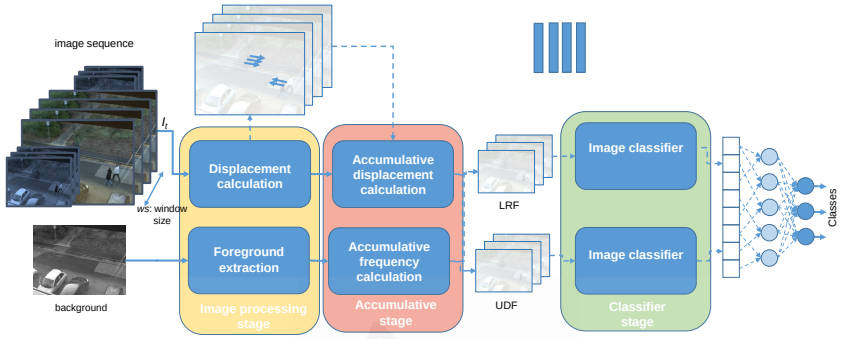


Figura 3.6: Flujo de datos en el la etapa de procesamiento de imágenes, donde se calcula el desplazamiento y se extrae el foreground de la secuencia de imágenes de entrada. En la etapa de acumulación se calcula el desplazamiento y la frecuencia acumulada. Continuando con al etapa de clasificación de imágenes donde recibe como dato de entrada LRF y UDF, para proporcionar la salida las diferentes clases de imágenes

Universitat d'Alacant
Universidad de Alicante

Actualmente se ha hecho popular el uso de arquitecturas de Redes Neuronales para el reconocimiento de comportamiento humano. En la mayoría de casos se utiliza una sola hilera de capas para extraer características. Sin embargo muchos autores consideran que se puede utilizar varias redes de forma paralela, cada una especializada en extraer ciertas características, y una vez obtenidas unirlas para mejorar el porcentaje de acierto. Sobre este tema existen algunos trabajos que muestran la utilización varas hileras de redes neuronales para detectar comportamientos [152, 188, 218, 218, 227, 296, 311]. En este trabajo se propone una arquitectura de dos y tres hileras: la primera y segunda extraen las características de comportamiento según el movimiento de las personas utilizando el flujo óptico, la tercera analiza la escena o contexto. En la figura 3.3 se describe en forma general el flujo de datos para los casos de OCC y MCC de tamaño ws como se muestra en la figura 3.7. El parámetro ws es ajustado en función de la necesidad de precisión resultante, este dato indica la cantidad

de imágenes que se toma del total de la secuencia como dato de entrada para calcular el desplazamiento, y luego el desplazamiento acumulado. En este caso, los componentes no se concatenan todos juntos, se separan conformando dos imágenes LRF compuestas por los componentes L, R y F, y , de forma similar, UDF combina los componentes U, D y F. En la figura 3.8 se muestra por separados las imágenes de LRF, UDF, y la imagen de una escena de comportamiento anormal.

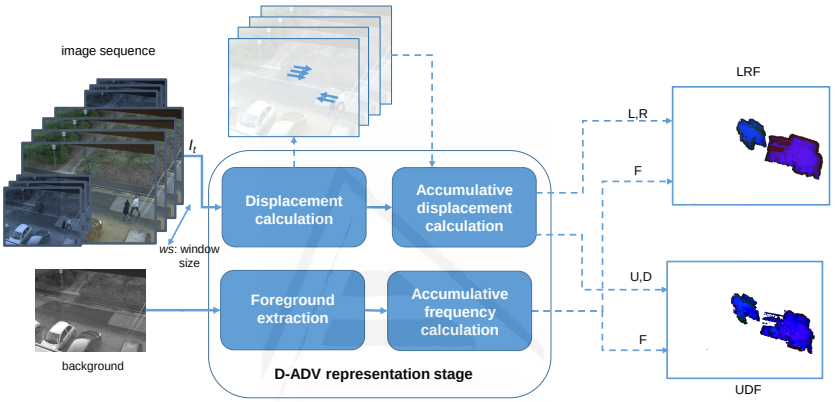


Figura 3.7: Flujo de datos en el la etapa de representación de D-ADV-OC donde se calcula el desplazamiento y se extrae el primer plano. Para el desplazamiento la entrada es un grupo de imágenes dado por el parámetro windowsize, y la salida el desplazamiento acumulado, para la extracción de foreground, la entrada es el background y la salida es la frecuencia acumulada.

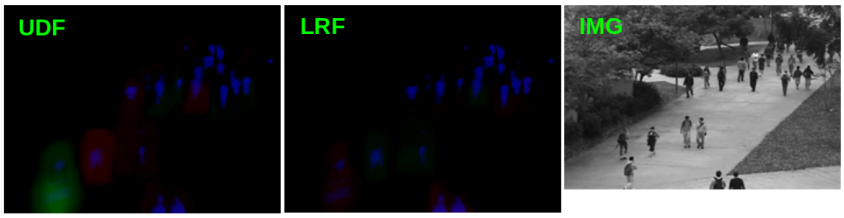


Figura 3.8: Imágenes LRF, UDF, IMG del D-ADV de una escena de comportamiento tipo anormal.

3.2.2. Clasificación de múltiples flujos

El uso de arquitecturas de redes neuronales para análisis de secuencias de video se ha hecho muy popular en los últimos tiempos. La idea principal es asignar una única tarea específica de clasificación a la arquitectura de red, encontrando múltiples trabajos que reconocen objetos en la escena [223, 230, 295, 297, 326], lugares [415, 416, 417], tipos de plantas [285, 314, 316], detección de cáncer[239, 259, 404], y otros muchos problemas heterogéneos en el ámbito de la clasificación de imágenes.

Una reciente tendencia sugiere la combinación de algunos tipos de arquitecturas en dos o más flujos de datos por donde ingresan imágenes o video para que sean analizados por las mismas con el objetivo de extraer características específicas, clasificar, etc. y combinar sus resultados en una sola clasificación final. De esta forma se utilizan varias redes de forma paralela, cada una especializada en extraer ciertas características, y una vez obtenidas unirlas para mejorar el porcentaje de acierto. Por ejemplo, existen arquitecturas de múltiples flujos que tienen como fin la emulación de la visión humana asumiendo que el sistema visual humano procesa lo que vemos a través de dos tipos de flujos de datos: la vía ventral y la vía dorsal respectivamente. La vía ventral es responsable de procesar la información espacial, como la forma y el color, mientras que la vía dorsal es responsable de procesar la información de movimiento [193]. Para procesar secuencias de video y extraer características existen propuestas separan al el video en componentes espaciales y temporales. La parte espacial como un frame individual, contiene información sobre las escenas y objetos encontrados en el video. La parte temporal, en forma de movimiento a través frames, transmite el movimiento del observador, que en este caso es la cámara, y los objetos en escena. Las partes espacial y temporal de la arquitectura propuesta en [337] son separadas en dos flujos de datos diferentes. A esta manera de analizar las secuencias de video se el conoce como arquitecturas two-streams, y existen trabajos adicionales con la misma lógica de funcionamiento de dos o más flujos de datos [155, 218, 339, 399, 420]. En cuanto al ámbito concreto de detectar comportamientos existen varias trabajos actuales. [152, 188, 218, 218, 227, 296, 311].

En esta tesis doctoral, se propone una arquitectura de clasificación de

actividades (ver Fig. 3.7 de dos flujos relacionados con el descriptor D-ADV. Se establece un flujo de datos para cada conjunto de imágenes del D-ADV: UDF y LRF. La combinación de ambos flujos se realiza mediante una fusión final de los clasificadores concatenándolos en un sólo vector que está conectado directamente a una capa de neuronas completamente conectada.

Además, el modelo arquitectural contempla el contexto en la escena como un flujo más de la clasificación. En este caso, la fusión de los flujos se realiza de forma ponderada a cada uno de los mismos y dependerá de la aplicación concreta en la que se instanciará la arquitectura de visión.

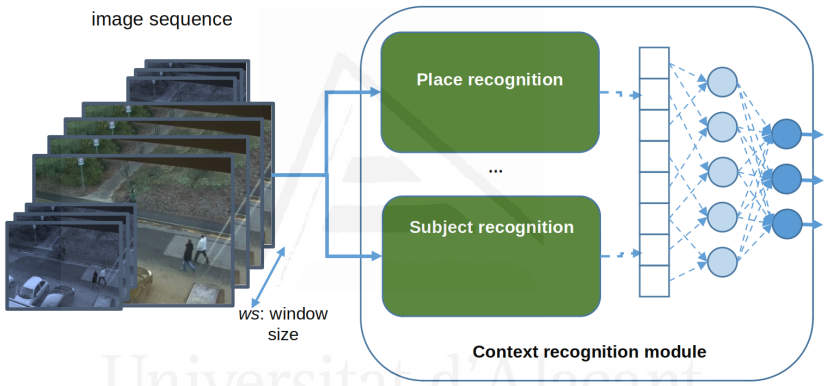


Figura 3.9: Como datos de entrada al bloque de reconocimiento de contexto se tiene un grupo de imágenes dadas por el parámetro windowsize que son un segmento representativo del total de frames del video analizado, la salida de este bloque nos muestra si el comportamiento es normal o anormal.

3.3. Arquitecturas propuestas

Para el desarrollo de esta tesis se han propuesto dos arquitecturas de red neuronales instanciando el modelo arquitectural definido en la sección anterior: instancia con clasificación One-Class (OCC) e instancia con clasificación Multi-Class (MCC). Estas arquitecturas son capaces de detectar comportamientos específicos que forman parte del conjunto de datos en entrenamiento y comportamientos anormales, a partir de entrenamiento

de datos normales, respectivamente. En las sub-secciones 3.3.1 y 3.3.2 se explican los detalles acerca de las etapas de clasificación y entrenamiento, así como algunos parámetros importantes de configuración de cada una de las instancias de la arquitectura.

3.3.1. Clasificación de múltiples actividades (D-ADV-MC)

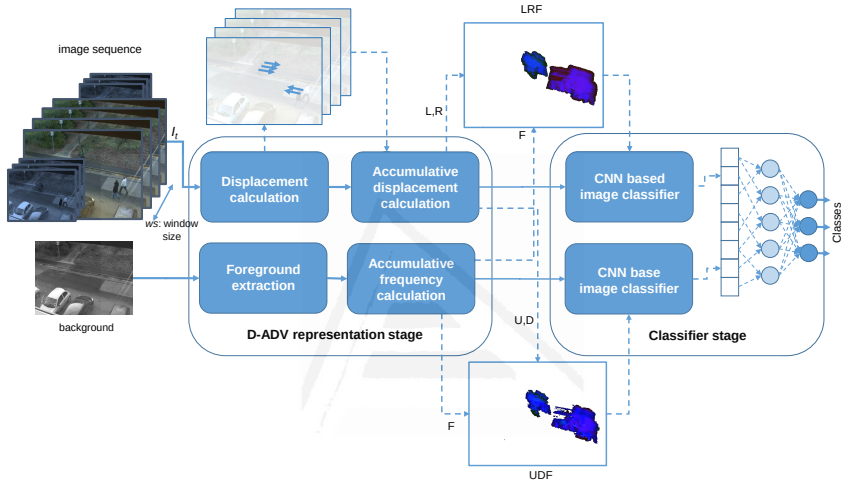


Figura 3.10: Flujo de datos de la propuesta D-ADV. La arquitectura D-ADV se divide principalmente en dos partes, la etapa de representación D-ADV-MC donde el desplazamiento se calcula usando el descriptor ADV a partir de una secuencia de imágenes y su movimiento de flujo óptico. La segunda etapa define el clasificador usando clasificadores CNN y una capa totalmente conectada para predecir la clase.

La arquitectura instanciada a partir del modelo arquitectural propuesto en esta tesis para la clasificación de múltiples actividades de grupo se puede observar en la Fig. X. Esta arquitectura utiliza la clasificación de actividades de dos flujos y realiza un fusión tardía tal como se ha analizado en la sección anterior capaz de clasificar las imágenes individuales previamente calculadas: LRF y UDF. La propuesta para el clasificador es abierta y se puede utilizar cualquier arquitectura CNN (VGG, Resnet, Alexnet, LeNet, etc.). Este tipo de redes generalmente utilizan una capa totalmente conectada en la salida con una función de activación softmax para decidir la clase a la que corresponde la imagen (por ejemplo, obje-

tos, lugares, poses, etc.). La arquitectura D-ADV no tiene en cuenta las capas densas individuales. Sin embargo, las capas anteriores en el modelo se concatenan en una fusión tardía con una capa de concatenación. Finalmente, una capa totalmente conectada con función de activación sigmoidea se utiliza para conectar la capa de concatenación para predecir múltiples clases.

Para evitar los problemas de los datasets con gran cantidad de imágenes para entrenar nuestro modelo y con el objetivo de que el modelo pueda ser usado para pequeños datasets (como por ejemplo BEHAVE y CAVIAR [37, 102] utilizado en la experimentación), se propone realizar una transferencia de aprendizaje (transfer learning) de los modelos entrenados con el dataset ImageNet [191]. En consecuencia, la red basada en CNN se refina tres veces. En primer lugar, la capa fully-connected de la arquitectura de ImageNet es reemplazada por una nueva que se ajusta con las nuevas clases. Después, un subconjunto de las capas inferiores se entrenan debido a que la entrada de LRF y UDF son diferentes a las imágenes RGB de ImageNet. Finalmente, un subconjunto de las capas superiores es reentrenado nuevamente para realizar un ajuste fino.

En la fase de entrenamiento se utiliza la entropía cruzada binaria (binary cross entropy) como función objetivo, a fin de considerar cada clase de salida como una distribución independiente de Bernoulli. En cuanto a la clasificación, y teniendo en cuenta que se podría presentar más de una clase en un frame de la secuencia, se consideran diferentes umbrales ϵ_i para cada neurona i de salida. En este caso los umbrales ϵ_i Se calculan como el valor que maximiza la tasa de verdaderos positivos (TPR) y minimiza la tasa de falsos positivos (FPR), para cada clase, C_i.

3.3.2. Clasificación de actividades anómalas (OCC)

En este apartado se especifica la instancia de la arquitectura propuesta para la detección de actividades anómalas en una escena. Esta arquitectura utiliza tres flujos de datos: los dos flujos relacionados con el movimiento y el asociado al contexto de la escena. Es decir, para el caso de OCC se tiene un esquema similar de dos flujos para la clasificación de imágenes LRF y UDF, a diferencia de MCC, esta clasificación se incluye un flujo adicional

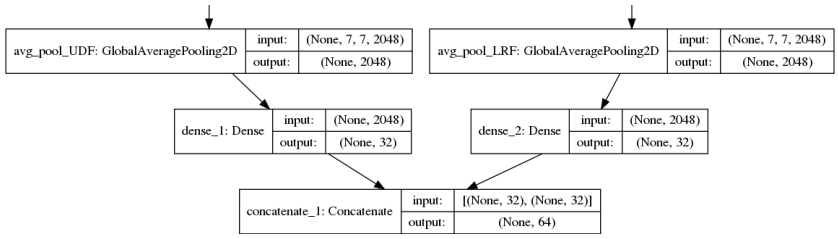


Figura 3.11: El flujo de datos de UDF (rama izquierda) y LRF (rama derecha) se conectan a una capa densa *dense-1* y *dense-1* respectivamente se concatenan con una capa totalmente conectada llamada *concatenate-1* que tiene como salida las las clases con las que se ha entrenado.

para extraer características del contexto, dando como resultado una arquitectura conformada por tres flujos. A esta arquitectura se le ha acuñado el término D-ADV-OC, y no utiliza las capas densas individuales ya que se conecta con capas previas a estas. Por esto, las capas anteriores en el convnet se concatenan de manera de fusión tardía utilizando en una capa de concatenación de las dos corrientes. Finalmente, una capa totalmente conectada con activación lineal se utiliza para conectar la capa de concatenación para predecir la actividad anormal en el grupo. Este diseño de arquitectura se basa en el reciente trabajo propuesto por Ruff et al. [307] que proporciona un modelo profundo para entrenar una red neuronal minimizando el volumen de una hiper-esfera que encierra las representaciones de la red de datos. Esta propuesta difiere del trabajo de Chalapathy et al. [53] al combinar la capacidad de las redes basadas en CNN para aprender progresivamente de un subconjunto de imágenes que son la representación de los datos de entrada junto con el objetivo one-class. A diferencia de este último trabajo, que utiliza auto-encoders para establecer la representación de los datos de entrada, definiendo el centro de la hiper-esfera, en esta tesis algunas capas de la red basadas en CNN son entrenables, lo que permite a la arquitectura continuar entrenando tanto el centro, como ajustando el radio de la hiper-esfera. Para evitar los problemas de los datasets grandes para entrenar nuestro modelo y con el objetivo de que pueda ser usado para datasets pequeños, se usa la transferencia de aprendizaje de los modelos entrenados con ImageNet.

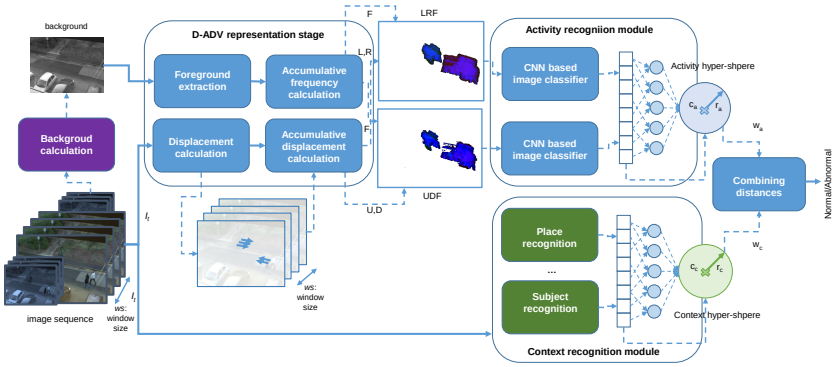


Figura 3.12: El flujo de datos del método D-ADV-OC. La arquitectura D-ADV-OC se divide principalmente en dos partes, la etapa de representación D-ADV-OC donde el desplazamiento se calcula utilizando el descriptor ADV a partir de una secuencia de datos y su movimiento de flujo óptico. La segunda etapa define el clasificador usando clasificadores CNN y una capa totalmente conectada para identificar la clase si es normal o anormal.

El tercer flujo de esta arquitectura está relacionado con la información de contexto en la escena, y en la fase de entrenamiento, se calculan los valores máximos de los patrones de entrada para normalizar los datos de salida que pueden ser objetos, lugares, etc. El valor medio de la normalización realizada establece el centro de la hiper-esfera, optimizando la longitud del radio de la misma a través de una capa totalmente conectada al final de la red.

El módulo de combinación de distancias utiliza los pesos w_a y w_c para la actividad y las funciones de pérdida de contexto con el fin de entrenar la red y calcula la distancia de un patrón de entrada a la clase normal según la etapa de predicción utilizando la siguiente función:

$$dist = \frac{1}{n} w_a \sum_i ||i_a - c_a||^2 + w_c ||i_c - c_c|^2$$

, siendo i_a la representación calculada para las actividades que utilizan el movimiento; i_a, la representación calculada del contexto en la escena; y, finalmente, c_a y c_c los centros de las hiper-esferas. Adicionalmente en este modelo se toma en cuenta la combinación del valor de pesos α para comportamiento y β.

Universitat d'Alacant
Universidad de Alicante

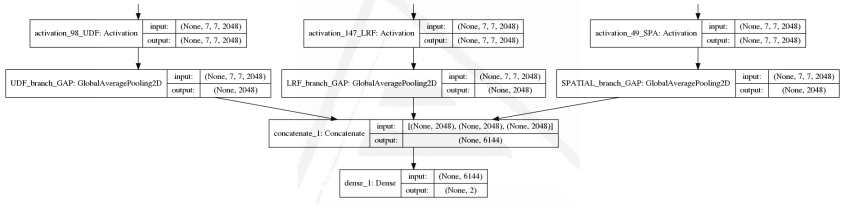


Figura 3.13: El flujo de datos de UDF (rama izquierda), LRF (rama media) e IMG (rama derecha) se conectan en una capa densa *dense-1* y *dense-1* respectivamente se concatenan con una capa totalmente conectada llamada *concatenate-1* que tiene como salida las las clases con las que se ha entrenado.

Universidad de Alicante

Capítulo 4

Experimentación

La experimentación para esta tesis se ha realizado en dos instancias, la primera para Multi-Class Classification (MCC) con los datasets INRIA, CAVIAR y BEHAVE para clasificar diferentes tipos de comportamientos, mientras que para la segunda etapa con One-Class Classification (OCC) con los datasets PED 1, PED 2 y Avenue para clasificar comportamientos de tipo normal o anormal. En los dos casos se ha utilizado secuencias de video de donde se extraen las diferentes características según el caso. Adicionalmente en las pruebas realizadas se ha probado con diferente valor para el parámetro windowsize (10 y 40) para el caso de MCC y variando la arquitectura (base-model y resnet-model) tanto en OCC como en MCC.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

4.1. Introducción

Con el objetivo de evaluar el desempeño de la propuesta y su generalidad, se han realizados experimentos para validar la arquitectura con las dos instancias D-ADV-MC y D-ADV-OC con los datasets públicos ampliamente utilizados, como son BEHAVE, INRIA y CAVIAR [37, 102] para multi-class, y para one-class los datasets (Ped 1, Ped 2) [237] y Avenue [233].

4.2. Experimentación D-ADV-MC

Para la instancia de la arquitectura basada en clasificación multi-clase la efectividad de la propuesta es evaluada en los datasets (BEHAVE, INRIA y CAVIAR) [37, 102], un conjunto de datos con varias secuencias de vídeo etiquetadas de dos vistas en varios escenarios con grupos de personas que realizan siete tipo de actividades.

4.2.1. Datasets

Universitat d'Alacant
Universidad de Alicante

En la experimentación relativa al D-ADV-MC se han utilizado los datasets públicos BEHAVE, CAVIAR e INRIA [37]. Los datasets contienen videos filmados desde dos puntos de vista de varios escenarios con personas que actúan en diversas interacciones. Los videos han sido capturados a 25 fps con una resolución es de 640x480, y están disponibles como AVI o como un conjunto numerado de archivos JPEG imágenes independientes. Los datos etiquetados incluyen varios tipos de comportamientos, para el dataset BEHAVE se describen a continuación:

- InGroup (EnGrupo): Las personas están en un grupo y tienen poco movimiento.
- Approach (Acercarse): Dos personas o grupos con uno (o ambos) acercándose el uno al otro.
- WalkTogether (CaminarJunto): Grupo de personas que caminan juntas.

- Meet (Reunirse): Dos o más personas que se reunidas.
- Split (Dividirse): Dos o más personas divididas o alejadas entre sí.
- Ignore (Ignorarse): Personas que se ignoran mutuamente.
- Chase (Perseguir): Un grupo atrás de otro que lo persigue.
- Fight (Pelear): Dos o más grupos peleando.
- RunTogether (CorrerJunto): Grupo de personas que corren juntos.
- Following (Seguir): Grupo de personas que es seguido por otro grupo.

El la figura 4.1 se muestra los diferentes tipos de comportamientos anteriormente descritos.

Adicionalmente el dataset INRIA, parte del proyecto CAVIAR [102] tiene imágenes de Personas/grupos reuniéndose, caminando juntos y separándose; y dos personas luchando en diferentes escenarios. Las clases específicas que se utilizan en este trabajo son *Fighting (peleando), Leaving (saliendo), Meeting (reuniéndose)*. Para el dataset CAVIAR tenemos las etiquetas de las clases *Meeting (reuniéndose), ShopEnter(EntrarTienda), ShopExit(SalirTienda), Walking(caminando).*

Universitat d'Alacant
Universidad de Alicante

4.2.2. Parámetros de configuración

Las pruebas realizadas con los datasets BEHAVE, CAVIAR e INRIA utilizan los mismos parámetros en todos los casos, salvo el valor del parámetro windowsize que representa al número de frames consecutivos considerados en el proceso de acumulación y son los datos de entrada de la arquitectura. El tamaño de las celdas que conforman las imágenes LRF y UDF es de 224 x 224. El clasificador de imágenes basado en la CNN es una red ResNet50 sin la última capa. Finalmente, los ajustes de la arquitectura se han realizado en las 139 capas inferiores en el primer paso y, finalmente, de la capa superior a la capa 249.

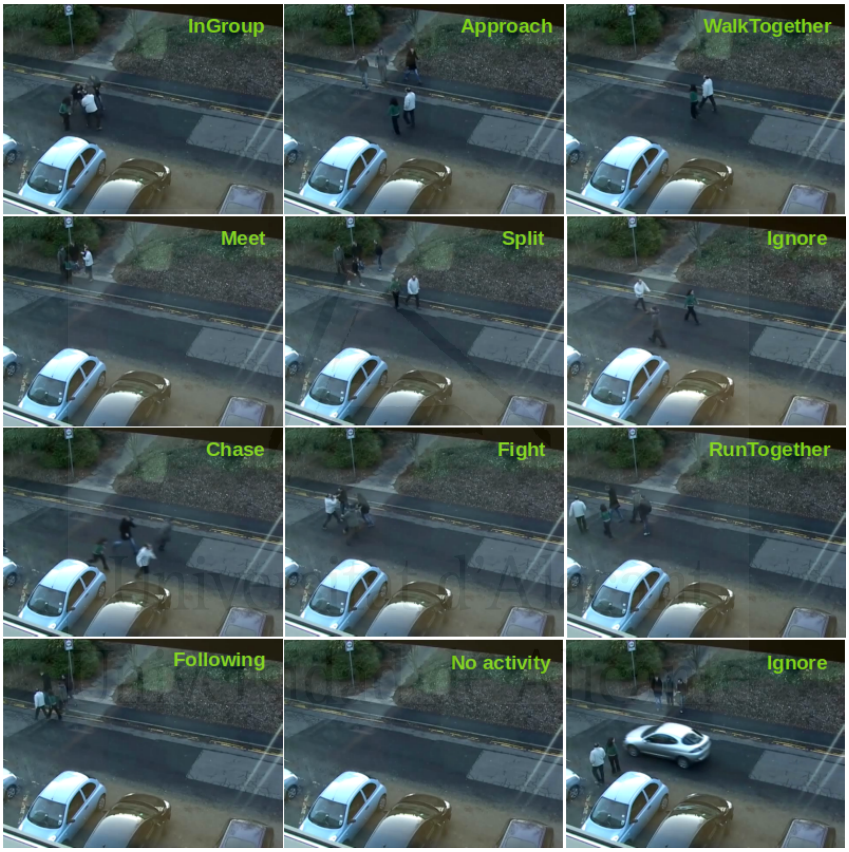


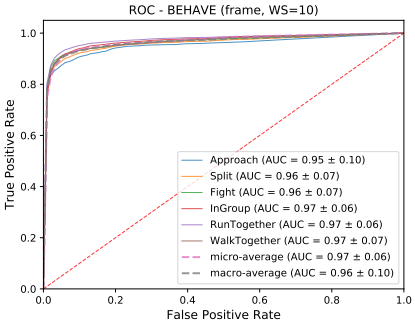
Figura 4.1: Comportamientos InGroup, Approach, WalkTogether, Meet, Split, Ignore, Chase, Fight, RunTogether, Following y No Activity del dataset BEHAVE.

4.2.3. Resultados

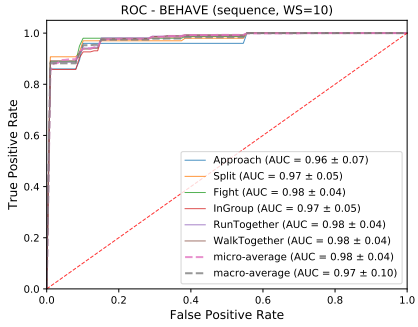
La Tabla 4.1 muestra los resultados de sensibilidad y especificidad, y las curvas AUC y ROC se muestra en la figuras 4.3, 4.2 y 4.4, a fin de analizar el rendimiento de la instancia D-ADV-MC para los frames y las secuencias, con valores variables del parámetro windowsize(ws) de 10 y 40. Los resultados obtenidos para frames, y con un valor para el parámetro windowsize (ws) de 10 alcanzan, para el dataset INRIA son 71,70 % de sensibilidad y 84,85 % en especificidad como promedio, para el dataset BEHAVE, un total de 91,47 % de sensibilidad y 94,51 % en especificidad, para CAVIAR (Corridor) 78,18 % de sensibilidad y 87,12 % en especificidad. Usando un valor mayor del parámetro windowsize, de 40 en este caso, los resultados mejoran en los tres conjuntos de datos. Obtenemos en total 89,93 % de sensibilidad y 95,65 % de especificidad para INRIA, 92,55 % de sensibilidad y 94,79 % de especificidad para BEHAVE, y 80,00 % de sensibilidad y 93,06 % de especificidad para CAVIAR (Corridor). En cuanto al rendimiento por secuencia indicado en la parte derecha de la tabla 4.1, el D-ADV obtiene altos resultados. Considerando un valor de 10 del parámetro windowsize, para el dataset INRIA, se alcanza un total de 91,67 % de sensibilidad y 95,83 % de especificidad. Además, el D-ADV obtiene un total de 95,07 % de sensibilidad y 95,52 % de especificidad para el BEHAVE. Se repite el proceso con los resultados considerando un valor de 40 para el parámetro windowsize logrando una mejora en los resultados. En promedio, el 95,83 % de sensibilidad y el 97,92 % de especificidad para el dataset INRIA y el 95,52 % de sensibilidad y el 95,70 % de especificidad para el BEHAVE. Para Caviar se tiene el 80.00 % de sensibilidad y el 93,00 % de especificidad. Como conclusión final se puede afirmar que tanto para frame como para sequence cuando se usa el parámetro windowsize = 40 todos los resultados de sensibilidad y especificidad son más altos que cuando windowsize = 10.

4.2.4. Comparativa con otros trabajos

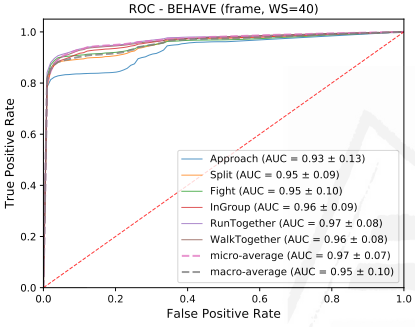
Para los experimentos de D-ADV-MC en esta tesis se han tomado seis clases diferentes del dataset INRIA (Approach, Split, WalkTogeter, In-



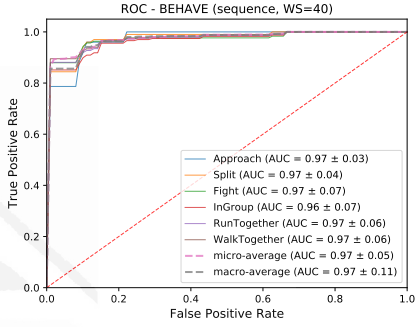
(a) ROC en frame para windowsize = 10.



(b) ROC en secuencia con windowsize = 10.



(c) ROC en frame con windowsize = 40.



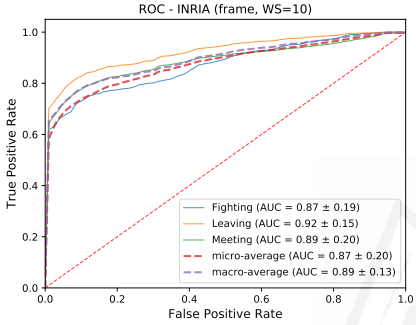
(d) ROC en secuencia con windowsize = 40.

Figura 4.2: Curvas ROC para el dataset BEHAVE para frame y sequence con el valor del parámetro windosize de 10 y 40.

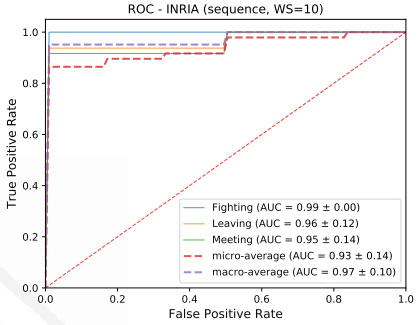
Universitat d'Alacant
Universidad de Alicante

Tabla 4.1: Comparación de resultados con los datasets (INRIA, BEHAVE, CAVIAR) calculados para frame y sequence con valores de windowsize (ws) 10 y 40.

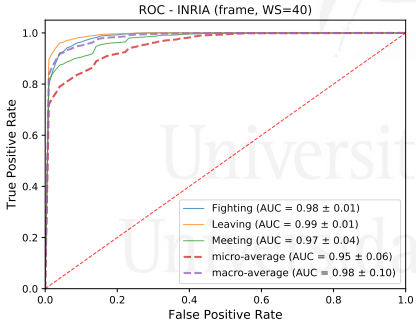
		Frame				**Sequence**			
		10		**40**		**10**		**40**	
Dataset	**Class**	Sensibilidad	Especificidad	Sensibilidad	Especificidad	Sensibilidad	Especificidad	Sensibilidad	Especificidad
Inria	Fighting	82.84 %	76.46 %	95.48 %	94.00 %	100.00 %	100.00 %	100.00 %	100.00 %
	Leaving	95.87 %	94.79 %	99.68 %	99.74 %	87.50 %	93.75 %	100.00 %	100.00 %
	Meeting	65.11 %	75.10 %	86.85 %	87.58 %	87.50 %	93.75 %	87.50 %	93.75 %
	Overall	**71.70 %**	**84.85 %**	**89.93 %**	**95.65 %**	**91.67 %**	**95.83 %**	**95.83 %**	**97.92 %**
Behave	Approach	90.88 %	92.45 %	92.02 %	92.68 %	93.94 %	92.08 %	93.94 %	95.05 %
	Split	92.58 %	93.23 %	95.18 %	93.92 %	97.14 %	95.96 %	97.14 %	96.97 %
	Fight	95.52 %	96.35 %	93.40 %	95.27 %	100.00 %	98.28 %	94.44 %	93.10 %
	InGroup	94.41 %	94.32 %	94.15 %	93.75 %	94.83 %	94.74 %	93.10 %	93.42 %
	RunTogheter	99.87 %	99.95 %	100.00 %	99.99 %	100.00 %	100.00 %	100.00 %	100.00 %
	WalkTogheter	84.12 %	88.30 %	87.92 %	90.82 %	92.31 %	88.41 %	96.92 %	94.20 %
	Overall	**91.47 %**	**94.51 %**	**92.55 %**	**94.79 %**	**95.07 %**	**95.52 %**	**95.52 %**	**95.70 %**
Caviar	Meeting	87.59 %	86.25 %	90.58 %	90.47 %	88.24 %	93.18 %	94.12 %	100.00 %
	Shop-Enter	91.28 %	91.14 %	85.80 %	93.86 %	100.00 %	95.56 %	87.50 %	93.33 %
	Shop-Exit	81.21 %	87.10 %	83.63 %	85.82 %	90.00 %	95.12 %	90.00 %	90.24 %
	Walking	72.16 %	76.49 %	73.57 %	75.17 %	65.96 %	78.57 %	68.09 %	78.57 %
	Overall	**78.18 %**	**87.12 %**	**79.00 %**	**88.88 %**	**80.00 %**	**93.06 %**	**80.00 %**	**93.06 %**



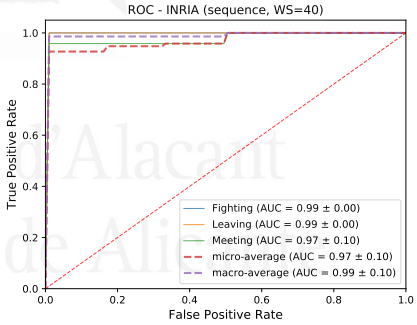
(a) ROC en frame con windowsize = 10.



(b) ROC en secuencia con windowsize = 10.



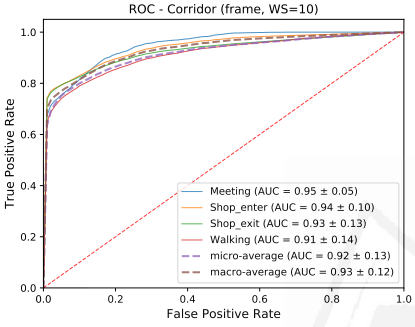
(c) ROC en frame con windowsize = 40.



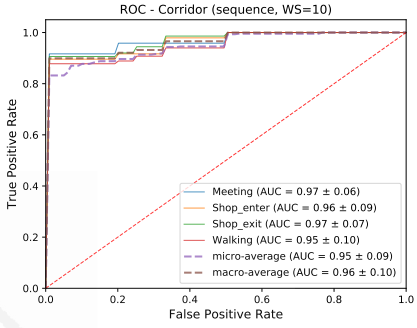
(d) ROC en secuencia con windowsize = 40.

Figura 4.3: Curvas ROC para el dataset INRIA para frame y sequence con el valor del parámetro windosize de 10 y 40.

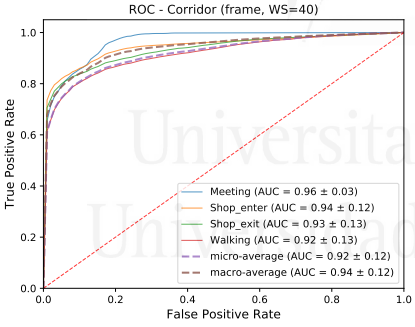
Universitat d'Alacant
Universidad de Alicante



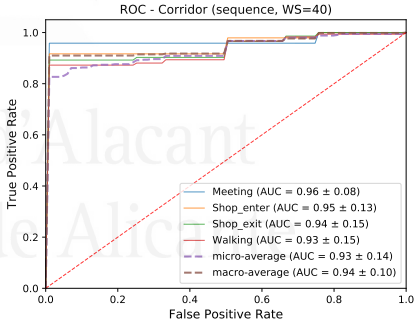
(a) ROC en frame con windowsize = 10.



(b) Curvas ROC en secuencia para el dataset CORRIDOR con windowsize = 10.



(c) ROC en frame con windowsize = 40.



(d) ROC en secuencia con windowsize = 40.

Figura 4.4: Curvas ROC para el dataset CORRIDOR para frame y sequence con el valor del parámetro windosize de 10 y 40.

Universitat d'Alacant
Universidad de Alicante

Group, Fight y RunTogether), y se han comparado los resultados con cinco trabajos diferentes ([19], [61], [406], [257], [396]), en donde [19] y [61] tienen resultados de las seis clases para comparar, los trabajos [406] y [257] hacen sus experimentos con las clases (Approach, Split, WalkTogeter y InGroup), y [396] con las clases (Split, WalkTogeter, InGroup y Fight). Los valores mostrados en la tabla 4.2 en los que el porcentaje de predicción por cada clase es más bajo que en los trabajos comparado es para la clase Approach que alcanza una valor de $93,94\,\%$, que es $6\,\%$ más bajo que en [19]. Para la clase Split se tiene $97,14\,\%$, que es $3\,\%$ más bajo que los valores alcanzados en [19] y [61]. Para la clase WalkTogether el valor de $96,92\,\%$ es más alto que en todos los trabajos con los que se compara. En la clase InGroup $93,1\,\%$ es ligeramente $1\,\%$ más bajo que en [396] y $7\,\%$ menor al valor de [61]. En la clase Fight solamente es $1\,\%$ más bajo que el valor de [396]. El valor alcanzado en la clase RunTogether es el más alto con $100\,\%$, que es igual al de [19]. Si se compara el descriptor D-ADV de este trabajo de tesis con los métodos propuestos en [19], [61], [406], [257], [396] considerando las siete clases del conjunto de datos BEHAVE. Sólo en el trabajo [61] y nuestro trabajo previo (G-ADV) consideran las siete clases en sus experimentos. El resto de los trabajos utilizan un subconjunto de cuatro clases. La tabla 4.2 muestra la comparación de los resultados de Sensibilidad. Como podemos ver, nuestra propuesta, D-ADV, logra en promedio los mejores resultados superando todos los métodos comparados. A continuación se detallan los valores de los porcentajes y diferencias específicas para cada clase y caso de comparación. Finalmente, comparamos nuestro descriptor D-ADV-MC con los métodos propuestos en [19], [61], [406], [257], [396] considerando las siete clases del conjunto de datos BEHAVE. Sólo en el trabajo [61] y un trabajo previo (G-ADV) consideran las siete clases en sus experimentos. El resto de los trabajos utilizan un subconjunto de cuatro clases. La tabla 4.2 muestra la comparación de los resultados de sensibilidad. Como se puede ver, la propuesta, D-ADV-MC, logra en promedio los mejores resultados superando todos los métodos comparados. Como cálculo final se hace un promedio de todos los valores resultantes de las clases obteniendo $95,93\,\%$, siendo este un valor superior al de todos los trabajos comparados.

Universitat d'Alacant
Universidad de Alicante

Tabla 4.2: Comparación de los resultados de D-ADV-MC con otras propuestas. Las celdas con el contenido de las letras ND significa que no existen valores disponibles

	D-ADV-MC	**G-ADV[19]**	**[61]**	**[406]**	**[257]**	**[396]**
Approach	93,94 %	100,00 %	83,33 %	71,00 %	60,00 %	ND
Split	97,14 %	100,00 %	100,00 %	79,00 %	70,00 %	93,10 %
WalkTogether	96,92 %	86,67 %	91,66 %	88,00 %	45,00 %	92,10 %
InGroup	93,10 %	86,67 %	100,00 %	88,00 %	90,00 %	94,30 %
Fight	94,44 %	90,00 %	83,33 %	ND	ND	95,10 %
RunTogether	100,00 %	100,00 %	83,33 %	ND	ND	ND
Promedio	**95,93 %**	**93,89 %**	**90,28 %**	**81,50 %**	**66,25 %**	**93,65 %**

4.3. Experimentación D-ADV-OC

En el caso de clasificación una-clase, evaluamos la efectividad de nuestra arquitectura propuesta en lo conjuntos de datos de referencia, Avenue [233], UCSD (Ped 1 y Ped 2)[237] de escenas que contienen multitudes, el dataset UCSD está conformado de 100 secuencias de video y cinco categorías de comportamientos anómalos bien definidas. Específicamente, sólo hay una clase definida como *normal*, y cualquier cosa diferente a ésta se considera *anormal*. Para cada criterio, se utilizan las métricas comúnmente usadas incluyendo una curva de la Característica Operativa del Receptor (ROC, del inglés Receiver Operating Characteristic) es dibujar el Área Bajo Curva (AUC) y la Tasa de Error Igual (EER), las métricas como se explica en [237], son usualmente reportadas para comparar los métodos. Además, llevamos a cabo una evaluación a nivel de frame en Ped 1, Ped 2 y Avenue.

Universidad de Alicante

4.3.1. Datasets

Las imágenes del dataset UCSD (PED 1 y PED 2) fueron filmadas con una cámara estacionaria montada sobre una elevación, con vistas a los pasillos peatonales de la Universidad de California en San Diego. Todos los videos son en escala de grises de 8 bits, con dimensiones de 238x158 a 10 fps. El video original es 740x480 a 30 fps, y está disponible bajo pedido. El directorio de video contiene los videos de las dos escenas, las densidades de multitud son variables entre escasa y muy concurrida. Para eventos considerados como normales los videos sólo contienen peatones. Los eventos

identificados como anormales se deben a: 1) La circulación de elementos no peatonales en los pasillos, o 2) peatones con patrones de movimiento anómalos. Los eventos anormales más comunes incluyen ciclistas, patinadores, carros de golf y personas que caminan a través de una pasarela o en la hierba que rodea a la zona peatonal. También se registraron algunos casos de personas en silla de ruedas. Todas las anomalías ocurren naturalmente, es decir, no fueron puestas en escena con el propósito de ensamblar el dataset. Los datos se dividieron en 2 subconjuntos, cada uno correspondiente a una escena diferente. La primera escena contiene grupos de personas que caminan hacia y alejándose de la cámara, y cierta cantidad de distorsión de la perspectiva. La segunda contiene escenas con movimiento peatonal paralelo al plano de la cámara. El video grabado de cada escena se dividió en varios clips de unos 200 frames. En cada clip, una anotación indica si existe o no una anomalía en ese frame. Además, se proporciona un subconjunto de 10 clips con máscaras binarias a nivel de pixel generadas manualmente, que identifican las regiones que contienen anomalías. Esto ayuda a la evaluación del rendimiento con respecto a la capacidad de localizar anomalías. Para el dataset Avenue se tiene 15 secuencias, cada una dura unos 2 minutos. El número total de frames es de 35.240. Existen 14 eventos considerados anormales, incluyendo correr, lanzar objetos y vagabundear. Se utilizan 4 vídeos como datos de entrenamiento con 8.478 frames en total. El la figura 4.5 se puede apreciar diferentes tipos de comportamientos considerados anormales, entre la multitud aparece un carro de golf y una personas en bicicleta.

4.3.2. Parámetros de configuración

Se han realizado experimentos para dos tamaños de ventana diferentes (10 y 40) con el objetivo de evaluar la capacidad de la representación para sintetizar la información extraída de la escena. Además, las imágenes LRF y UDF se han normalizado en un rango de (0,1) dividiendo cada pixel (celda) por el valor máximo de cada componente. Para obtener resultados que puedan ser generalizados a un dataset independiente, se ha realizado una validación cruzada de 10 veces. Para el proceso de entrenamiento, se ha utilizado el 25 % de los datos para el conjunto de validación. Para la

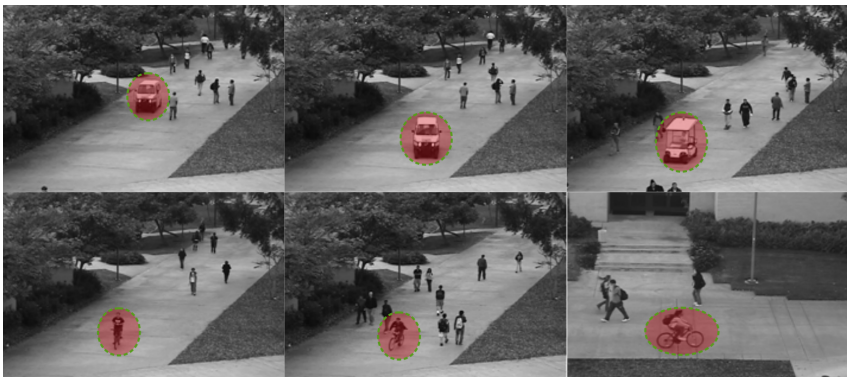


Figura 4.5: Comportamientos anormales de los datasets PED 1 y PED 2.

función de loss se ha usado los parámetros de two-stream y three-stream, con variaciones en la arquitectura de Resnet y base-model.

4.3.3. Resultados

Los resultados de la experimentación de las pruebas realizadas para D-ADV-OC se han hecho con los datastets Ped 1, Ped 2 y Avenue. Se ha calculado los valores de AUC y EER. En el caso de los experimentos hecho con el dataset Ped 1 tenemos un valor de 84.4 % para AUC con una función de loss three-stream y una arquitectura tipo base-model. Para el dataset Ped 2 el valor alcanzado en AUC es de 84.8 % con una función de loss two-stream y una arquitectura tipo Resnet. Finalmente para el dataset Avenue se tiene un valor de 78.7 % en AUC con una función de loss three-stream y una arquitectura tipo Resnet.

4.3.4. Comparativa con otros trabajos

En la tabla 4.3 se muestran los resultados de la comparación de nuestra propuesta con otros enfoques para reconocimiento de comportamientos anormales utilizando los datasets PED 1, PED 2 y Avenue. Los valores obtenidos en la arquitectura propuesta para PED 1 son 84.41 % para AUC y 24.36 % para EER, en PED 2 se tiene 95.04 % para AUC y 11.49 % para EER, y para Avenue 82.29 % en AUC y 22.70 % en EER. Se ha hecho una comparativa con los trabajos [153, 209, 217, 234, 262, 263, 298, 372,

377, 390, 419], en los que los valores promedio de AUC y EER son PED 1 (AUC 78.78 % y EER 29.08 %), PED 2 (AUC 94.07 % y EER 11.51 %) y Avenue (AUC 83.85 % y EER 26.60 %), mientras que los valores obtenidos con la arquitectura propuesta son PED 1 (AUC 84.41 % y EER 24.36 %), PED 2 (AUC 95.04 % y EER 11.49 %) y Avenue (AUC 82.29 % y EER 22.70 %). Concluyendo que el valor de AUC es superior para PED 1 y PED 2, mientras que para Avenue es menor con 1.56 %. El valor de EER es menor en todos los datasets.

Si analizamos los valores obtenidos para los tres datasets (PED 1, PED2 y Avenue) se tiene que se ha usado dos funciones de loss diferentes (OC-SVDD y OCC), en el método se ha combinado el D-ADV con contexto (D-ADV+Context), D-ADV con una red neuronal convolucional (D-ADV+CNN), D-ADV con una red neuronal convolucional y con el contexto (D-ADV+CNN+Context), finalmente también se ha probado solamente el D-ADV. También se ha calculado el valor promedio de los métodos usados con la funciones de loss OC-SVDD y OC-NN, además el promedio global.

En el caso de los resultados mostrados en la tabla 4.4 con el dataset PED 1 se tiene con la función de loss OC-SVDD el valor más altos es cuando se usa un clasificador con el contexto (D-ADV+Context) obteniendo para sensibilidad 73.8 %, especificidad 73.9 %, AUC 81.0 % y EER 26.1 %. Para la función de loss OCNN de igual modo que en el anterior, el uso de contexto (D-ADV+Context) influye para que se tenga valores más altos en sensibilidad 74.5 %, especificidad 76.0 %, AUC 84.4 % y EER 24.4 %. Los valores más bajos en sensibilidad, especificidad, AUC y EER se obtienen cuando se usa cualquiera de las dos funciones de loss (OC-SVVD y OC-NN) combinadas con el clasificador (D-ADV+CNN). El promedio de los valores obtenidos con las dos funciones de loss (OC-SVVD y OC-NN) es menor que el máximo obtenido cuando se usa el contexto en el clasificador (D-ADV+Context).

Para el dataset PED 2 los resultados mostrados en la tabla 4.5 con la función de loss OC-SVDD el valor más altos es cuando se usa un clasificador con el contexto (D-ADV+Context) obteniendo para sensibilidad 88.5 %, especificidad 88.4 %, AUC 95.0 % y EER 11.49 %. Para la función

de loss OCNN de igual modo que en el anterior, el uso de contexto (D-ADV+Context) influye para que se tenga valores más altos en sensibilidad 88.53 %, especificidad 88.40 %, AUC 95.04 % y EER 11.49 %. Los valores más bajos en sensibilidad, especificidad, AUC y EER se obtienen cuando se usa cualquiera de las dos funciones de loss (OC-SVVD y OC-NN) combinadas con el clasificador (D-ADV+CNN). El promedio de los valores obtenidos con las dos funciones de loss (OC-SVVD y OC-NN) es menor que el máximo obtenido cuando se usa el contexto en el clasificador (D-ADV+Context). Los resultados con el dataset Avenue mostrados en la tabla 4.6 se tiene con la función de loss OC-SVDD el valor más altos es cuando se usa un clasificador con el contexto (D-ADV+Context) obteniendo para sensibilidad 77.3 %, especificidad 77.3 %, AUC 82.3 % y EER 22.7 %.

Para la función de loss OCNN de igual modo que en el anterior, el uso de contexto (D-ADV+Context) influye para que se tenga valores más altos en sensibilidad 77.29 %, especificidad 77.30 %, AUC 82.29 % y EER 22.70 %. Los valores más bajos en sensibilidad, especificidad, AUC y EER se obtienen cuando se usa cualquiera de las dos funciones de loss (OC-SVVD y OC-NN) combinadas con el clasificador (D-ADV+CNN). El promedio de los valores obtenidos con las dos funciones de loss (OC-SVVD y OC-NN) es menor que el máximo obtenido cuando se usa el contexto en el clasificador (D-ADV+Context).

En la tabla 4.7 se ha calculado los valores promedios de los datasets PED 1, PED 2 y Avenue. Con la función de loss OC-SVDD, si comparamos los valores promedio de sensibilidad con 79.89 %, especificidad con 79.89 %, AUC con 86.12 % y EER con 20.11 % de los tres datasets (PED 1, PED2 y Avenue), tenemos que el más alto que el promedio es para el dataset PED 2 para sensibilidad con 88.5 %, seguido de especificidad con 88.4 %, AUC con 95.5 % y EER con 11.49 % para el valor más bajo. Con la función OC-NN y clasificador D-ADV+Context se tiene un valor promedio de sensibilidad de 78.22 %, especificidad de 78.09 %, para AUC 86.23 % y para ERR 21.84 %.

Universitat d'Alacant
Universidad de Alicante

Tabla 4.3: Comparación de resultados de D-ADV-OC a nivel de frame con otros métodos con los datasets PED 1, PED 2 y Avenue. Las celdas que contienen las letras ND indican que no existe datos disponibles.

	Frame level					
	Ped 1		**Ped 2**		**Avenue**	
Referencia	**AUC %**	**EER %**	**AUC %**	**EER %**	**AUC %**	**EER %**
[372]	82.34	23.50	97.52	4.68	71.54	36.38
[153]	80.90	26.30	95.90	10.50	84.20	23.00
[419]	83.50	25.20	94.90	10.30	86.10	22.00
[217]	83.80	22.30	96.50	8.70	84.50	21.50
[377]	77.80	29.20	96.40	8.90	85.30	23.90
[298]	53.50	48.00	81.40	26.00	73.80	32.80
[209]	82.10	ND	96.50	ND	87.20	ND
[234]	86.26	ND	96.06	ND	85.78	ND
[164]	ND	ND	97.80	ND	90.40	ND
[262]	ND	ND	96.20	ND	86.90	ND
[390]	ND	ND	96.80	ND	86.20	ND
[263]	ND	ND	82.80	ND.	84.30	ND
Promedio	**78.78**	**29.08**	**94.07**	**11.51**	**83.85**	**26.60**
Arquitectura Propuesta	**84.41**	**24.36**	**95.04**	**11.49**	**82.29**	**22.70**

Universitat d'Alacant
Universidad de Alicante

Tabla 4.4: Resultados de los experimentos de D-ADV-OC para el dataset PED 1. Se utiliza las funciones de loss OCC-SVDD y OCC-NN, con una combinación de clasificadores D-ADV, D-ADV+Context, D-ADV+CNN y D-ADV+CNN+Context para calcular la sensibilidad, especificidad, AUC y EER.

Dataset PED 1					
Func. Loss	**Método**	**Sensibilidad**	**Especificidad**	**AUC**	**EER**
OC-SVDD	D-ADV	73.84 %	73.87 %	81.03 %	26.14 %
OC-SVDD	**D-ADV+Context**	**73.84 %**	**73.87 %**	**81.03 %**	**26.14 %**
OC-SVDD	D-ADV+CNN	69.74 %	69.73 %	77.81 %	30.26 %
OC-SVDD	D-ADV+CNN+Context	70.31 %	70.30 %	78.03 %	29.69 %
OC-NN	D-ADV	74.29 %	74.20 %	83.40 %	25.75 %
OC-NN	**D-ADV+Context**	**75.43 %**	**75.95 %**	**84.41 %**	**24.36 %**
OC-NN	D-ADV+CNN	68.90 %	68.94 %	76.78 %	31.08 %
OC-NN	D-ADV+CNN+Context	69.84 %	69.84 %	77.00 %	30.16 %

Tabla 4.5: Resultados de los experimentos de D-ADV-OC para el dataset PED 2. Se utiliza las funciones de loss OCC-SVDD y OCC-NN, con una combinación de clasificadores D-ADV, D-ADV+Context, D-ADV+CNN y D-ADV+CNN+Context para calcular la sensibilidad, especificidad, AUC y EER.

Dataset PED 2					
Func. Loss	**Método**	**Sensibilidad**	**Especificidad**	**AUC**	**EER**
OC-SVDD	D-ADV	82.89 %	82.87 %	90.51 %	17.11 %
OC-SVDD	**D-ADV+Context**	**88.53 %**	**88.40 %**	**95.04 %**	**11.49 %**
OC-SVDD	D-ADV+CNN	71.36 %	71.27 %	80.04 %	28.66 %
OC-SVDD	D-ADV+CNN+Context	85.19 %	85.64 %	91.43 %	14.73 %
OC-NN	D-ADV	81.80 %	81.77 %	90.46 %	18.21 %
OC-NN	**D-ADV+Context**	**83.68 %**	**83.15 %**	**91.37 %**	**16.42 %**
OC-NN	D-ADV+CNN	73.24 %	73.20 %	80.54 %	26.77 %
OC-NN	D-ADV+CNN+Context	74.27 %	74.31 %	80.72 %	25.72 %

Tabla 4.6: Resultados de los experimentos de D-ADV-OC para el dataset Avenue. Se utiliza las funciones de loss OCC-SVDD y OCC-NN, con una combinación de clasificadores D-ADV, D-ADV+Context, D-ADV+CNN y D-ADV+CNN+Context para calcular la sensibilidad, especificidad, AUC y EER.

Dataset Avenue					
Func. Loss	**Método**	**Sensibilidad**	**Especificidad**	**AUC**	**EER**
OC-SVDD	D-ADV	75.62 %	75.62 %	82.17 %	24.38 %
OC-SVDD	**D-ADV+Context**	**77.29 %**	**77.30 %**	**82.29 %**	**22.70 %**
OC-SVDD	D-ADV+CNN	71.63 %	71.61 %	76.97 %	28.39 %
OC-SVDD	D-ADV+CNN+Context	71.93 %	71.87 %	78.84 %	28.11 %
OC-NN	D-ADV	74.30 %	74.29 %	81.06 %	25.70 %
OC-NN	**D-ADV+Context**	**75.57 %**	**75.17 %**	**82.90 %**	**24.73 %**
OC-NN	D-ADV+CNN	73.36 %	73.35 %	79.83 %	26.65 %
OC-NN	D-ADV+CNN+Context	74.35 %	74.34 %	81.29 %	25.66 %

Tabla 4.7: Resultados de los experimentos de D-ADV-OC para los valores promedio de los datasets PED 1, PED 2 y Avenue. Se utiliza las funciones de loss OCC-SVDD y OCC-NN, con una combinación de clasificadores D-ADV, D-ADV+Context, D-ADV+CNN y D-ADV+CNN+Context para calcular la sensibilidad, especificidad, AUC y EER.

Valores promedio (PED 1, PED 2, Avenue)					
Func. Loss	**Método**	**Sensibilidad**	**Especificidad**	**AUC**	**EER**
OC-SVDD	D-ADV	77.45 %	77.46 %	84.57 %	22.55 %
OC-SVDD	**D-ADV+Context**	**79.89 %**	**79.86 %**	**86.12 %**	**20.11 %**
OC-SVDD	D-ADV+CNN	70.91 %	70.87 %	78.27 %	29.10 %
OC-SVDD	D-ADV+CNN+Context	75.81 %	75.94 %	82.77 %	24.18 %
OC-NN	D-ADV	76.79 %	76.75 %	84.97 %	23.22 %
OC-NN	**D-ADV+Context**	**78.22 %**	**78.09 %**	**86.23 %**	**21.84 %**
OC-NN	D-ADV+CNN	71.83 %	71.83 %	79.05 %	28.17 %
OC-NN	D-ADV+CNN+Context	72.82 %	72.83 %	79.67 %	27.18 %

Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 5

Conclusiones

En este último capítulo se exponen las conclusiones del trabajo de tesis doctoral en donde se analizan las principales aportaciones realizadas a través de la arquitectura propuesta y la taxonomía de comportamientos presentada. Los hallazgos más relevantes encontrados en el estado de arte de los temas relacionados, los resultados de la investigación realizada, la interpretación de los resultados obtenidos en los diferentes experimentos realizados, y finalmente se incluye algunos los trabajos a futuro de corto, mediano y largo plazo que se pueden derivar de esta investigación.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

5.1. Conclusiones

En este trabajo de tesis doctoral se ha propuesto una arquitectura genérica basada en el análisis de movimientos locales que permite clasificar actividades de grupos de personas denominada D-ADV. Esta arquitectura, está compuesta de dos bloques principales, el descriptor de movimiento y el clasificador de actividad. Esta arquitectura se ha instanciado en dos casos que se ven presentes en el estado del arte, el D-ADV-MultiClass para clasificación multi-clase, y el D-ADV-OneClass para clasificación una-clase (one-class). También, se ha propuesto en esta tesis una taxonomía que permite clasificar las actividades y comportamientos desde una perspectiva bidimensional nivel de semántica/número de personas, que se ajusta más a los retos y soluciones actuales del estado del arte.

En el capítulo 2 de esta tesis se ha hecho un extenso estudio del estado del arte sobre el análisis del comportamiento humano. Se han revisado, en primer lugar, actuales propuestas de clasificación de niveles de comportamiento, o semántica como aquí se redefine. Se ha visto que no hay un consenso claro en los niveles, y se ha detectado que la nomenclatura no es clara. Es por ello que utilizamos el término "nivel de semántica" para definir los distintos niveles de complejidad de los comportamientos. También se han revisado los principales trabajos de machine learning y de deep learning para el análisis de comportamiento humano, centrado en grupos y multitudes. Del mismo modo, se ha hecho realizado un estudio de los principales datasets con datos del tema relativo a la tesis, con datos de grupo para técnicas de machine learning y deep learning.

Tras el estudio del estado del arte se ha propuesto una taxonomía que define un espacio bidimensional que permite describir las propuestas y retos en el análisis de las actividades y comportamientos humanos con dos elementos claves, el número de personas que puede ir desde una que representa a un individuo, hasta muchas que representan una multitud, y por otro lado, está el grado de semántica clasificado en bajo, medio y alto y, para este trabajo, basado en el tiempo de duración de la actividad. Siendo el nivel bajo en donde se ubican los movimientos elementales y con un poca semántica, además con un tiempo corto de duración. En el nivel

Universitat d'Alacant
Universidad de Alicante

medio los comportamientos un poco más complejos o combinación de los del nivel inferior, con un tiempo más prolongado de duración, que podría ser del orden de minutos. Finalmente, en el nivel más alto estarían los comportamientos complejos con mayor tiempo de duración y alto grado de semántica, como casos que duran horas o días.

En este documento de tesis se ha propuesto una arquitectura genérica para el reconocimiento de actividades de grupo basado en el descriptor de trayectoria , el cual hemos llamado D-ADV. El descriptor de trayectoria es una variante del Activity Descriptor Vector que permite describir la trayectoria en base a un vector de movimientos locales acumulados. La variante considera cualquier movimiento en la imagen en lugar de hacer uso de las trayectorias específicas del individuo o del grupo proporcionando generalidad en la entrada, y permitiendo su uso con datasets reducidos incluso utilizando redes profundas ya que se reduce el espacio de búsqueda. Además, permite generalización en cuanto al número de individuos en la escena, ya que el descriptor de trayectorias aísla la red de la imagen original de entrada, haciendo que solo se entrenen movimientos locales generales. El movimiento aparente es calculado por el flujo óptico en celdas distribuidas espacialmente según la imagen de entrada de la secuencia. Permite generar dos imágenes que contienen la descripción del movimiento y la aparición de los sujetos en la escena llamadas UDF y LRF por tratarse de imágenes acumuladas de los movimientos hacia arriba (Up), abajo (Down) y frecuencia (Frecuency), e izquierda (Left), derecha (Right) y la frecuencia también. Por otro lado, se ha instanciado la arquitectura D-ADV para el caso de clasificación normal/anormal con la particularidad de utilizar técnicas de clasificación una-clase (One-Class Classification). Esta propuesta es usada para detectar comportamientos normales y anormales, y se basa en tres streams (UDF,LRF) y, un stream adicional y una imagen de la escena para detectar el contexto. En este caso se usa fusión tardía con una capa densa. En este trabajo, se ha utilizado el aprendizaje por transferencia para no entrenar toda la arquitectura aprovechando los segmentos previamente entrenados y, entrenar solamente los segmentos de la arquitectura donde se concatenan los streams.

En el capítulo 4 se han realizados experimentos para validar la arqui-

tectura con las dos instancias D-ADV-MC y D-ADV-OC utilizando los datasets públicos ampliamente utilizados, como son BEHAVE, INRIA y CAVIAR para multi-class, y para one-class los datasets Ped 1, Ped 2 y Avenue. Los resultados experimentales muestran la capacidad de la arquitectura para clasificar las actividades de los grupos presentados en los datasets. Además, se demuestra que la arquitectura es capaz de tener buenos resultados utilizando datasets con poca cantidad de datos. En este caso, no a partir de la imagen sino de la representación del movimiento.

5.2. Trabajos a futuro

Es necesario proponer arquitecturas de aprendizaje más generales que puedan referirse a diferentes niveles de semántica y que sean independientes del grupo y de la multitud, en otras palabras, independientemente del número de individuos en el conjunto de personas analizado. En otros casos, los métodos deberían ser capaces de detectar automáticamente cuando un grupo se convierte en una multitud o viceversa, inclusive para detectar automáticamente cuándo una multitud se dispersa y se convierte en varios grupos. En consecuencia, la combinación de métodos generales para diferentes semánticas e independientes del número de personas sería el mayor reto. Para dar continuidad al de tema de investigación acerca del comportamiento humano, se ha establecido objetivos a corto mediano y largo plazo.

Trabajos a corto plazo: avanzar en la investigación sobre la arquitectura propuesta con las instancias de one-class y multi-class, realizando algunos afinamientos a los parámetros del modelo, con el objetivo de mejorar los resultados y compararlos con otros enfoques que propongan soluciones similares al problema abordado.

Trabajos a mediano plazo: para este periodo se ha pensado realizar algunos ajustes adicionales a nivel de etiquetas en datasets del mismo tipo de los ya utilizados (ShanghaiTech, The WorldExpo'10, New York Grand Central Station), para analizar los resultados de precisión obtenidos.

Trabajos a largo plazo: se tiene previsto incluir streams adicionales a la arquitectura para abordar otro tipo de comportamientos en donde exista

interacción con objetos, lugares u otro tipo de variable que pueda intervenir en el tipo de comportamiento de grupos.

5.3. Aportaciones

Como aporte a la investigación en el tema de Visión por Computador y puntualmente en el Análisis del Comportamiento Humano se ha publicado trabajos en los congresos IWANN2017, IJCNN2018, IJCNN2020 y SOCO2020, la revista IJCVIP2017. Adicionalmente se está desarrollando un artículo adicional con los resultados obtenidos en la experimentación realizada en este trabajo.

A continuación se presenta el resumen de los trabajos publicados:

- **Machine learning methods from group to crowd behaviour analysis:** IWANN2017
  En este trabajo se resumen los principales métodos para reconocimiento de comportamiento humano de grupos y multitudes, se incluyen los conjuntos de datos utilizados y se describe el nivel de semántica de cada trabajo de acuerdo a la definición en una escala propuesta en base a otros trabajos. La escala sobre la semántica define tres niveles para describir el comportamiento humano(movimientos, acciones y actividades) y adicionalmente se consideran dos aspectos para multitudes(seguimiento y contar personas).

- **A Compilation of Methods and Datasets for Group and Crowd Action Recognition** IJCVIP2017
  En este artículo se examina el estado del arte centrándose en las técnicas de aprendizaje automático y visión por computador. Además, se concluye que en la literatura se encuentra una falta de revisión que compare el nivel de abstracción en cuanto a la duración de las actividades. En este trabajo se presenta una revisión de los métodos y técnicas basadas en el aprendizaje automático para clasificar el comportamiento del grupo en secuencia de imágenes. Además, se describen los diferentes niveles de semántica y el número de personas en el grupo. Este trabajo está directamente relacionado con el

capítulo 2 de esta tesis.

- **A short review of deep learning methods for understanding group and crowd activities:** IJCNN2018
  En esta publicación se describen varias técnicas de Análisis de Comportamiento Humano basadas en Aprendizaje Profundo principalmente, se describen los conjuntos de datos y las Redes Neuronales utilizadas. En este trabajo se realiza una propuesta para detectar comportamiento humano de grupos y multitudes en donde intervienen algunas variables como el número de personas en escena y el tiempo de duración de los movimientos, acciones, actividades o comportamientos que pueden tener un grupo de personas o multitud.

- **Deep Learning Architecture for Group Activity Recognition using Description of Local Motions:** IJCNN2020
  En este artículo se propone una arquitectura de visión por ordenador capaz de aprender y reconocer las actividades de grupo utilizando los movimientos de la misma en la escena. Se basa en el ADV, un descriptor capaz de representar la información de la trayectoria de una secuencia de imágenes como una colección de los movimientos locales que ocurren en regiones específicas de la escena. Una versión evolucionada este descriptor genera imágenes de entrada de una red neuronal convolucional de dos streams (LRF, UDF) capaz de clasificar de forma eficiente las actividades de un grupo. Además el uso del análisis de trayectoria que permite una comprensión sencilla y de alto nivel de las actividades de grupos complejos, aprovecha las características de aprendizaje profundo para el reconocimiento de clases múltiples utilizando multi-class classification. Este trabajo está relacionado con el capítulo 3 y 4, en las partes relativas a la clasificación multi-clase.

- **A deep learning architecture for recognizing abnormal activities of groups using context and motion information:** Aceptado en SOCO2020
  En este trabajo se propone una arquitectura de visión computadorizada capaz de aprender y reconocer las actividades de un grupo utilizando los movimientos del grupo en la escena. Se basa en el ADV

Universitat d'Alacant
Universidad de Alicante

y variantes de este descriptor capaz de representar la información de la trayectoria de una secuencia de imágenes como una colección de los movimientos locales que ocurren en regiones específicas de la escena. El uso de aprendizaje profundo en este trabajo, permite el desarrollo de una arquitectura robusta de tres streams (LRF,UDF) y una imagen de la escena a modo de contexto para el reconocimiento de clases, en este trabajo se usa one-class classification para detectar si el comportamiento del grupo es normal o anormal. Este artículo está directamente relacionado con el capítulo 3 y 4, con las partes correspondientes a la clasificación one-class.

Universitat d'Alacant
Universidad de Alicante

Lista de Acrónimos

AAEs *Adversarial Autoencoders*

AAL *Ambient Assisted Living*

Aes *Autoencoders*

ADL *Activity Daily Live*

ADV *Activity Descriptor Vector*

AI *Artificial Intelligence*

ANN *Artificial Neural Network*

BP *Backpropagation*

BoW *Bag of Words*

CAEs *Convolutional Auto-Encoders*

CCTV *Closed Circuit Television*

COAes *Contractive Autoencoders*

COCO *Common Objects in Context*

CNN *Convolutional Neural Network*

DAD *Deep Anomaly Detection*

DAE *Denoising Auto-Encoders*

DT *Decision Trees*

D-ADV *Deep Activity Descriptor Vector*

D-ADV-MC *Deep Activity Descriptor Vector Multi Class*

D-ADV-OC *Deep Activity Descriptor Vector One Class*

DHM *Deep Hybrid Model*

DL *Deep Learning*

DNN *Deep Neural Networks*

FPN *Feature Pyramid Network*

Universitat d'Alacant
Universidad de Alicante

GADV *Group Activity Descriptor Vector*

GAN *Generative Adversarial Network*

GMM *Gaussian Mixed Model*

GPU *Graphics Processing Unit*

GPU-MPCNN *GPU-Based MPCNN*

GRUs *Gated Recurrent Unit Neural Networks*

HAR *Human Activity Recognition*

HBA *Human Behavior Analysis*

HOF *Histogram Optical Flow*

HMM *Hidden Markov Model*

IoT *Internet of the Things*

LDA *Linear Discriminant Analysis*

LRF *Left Right Frecuency*

LSTMs *Long Short-Term Memory Networks*

MC *Multi Class*

MCC *Multi Class Classification*

MCMC *Monte Carlo Markov Chains*

MDP *Markov Decision Process*

ML *Machine Learning*

MP *Max-Pooling*

MPCNN *Max-Pooling CNN*

NB *Naive Bayes*

NGAS *Neural GAS*

NN *Neural Network*

OC *One Class*

OCC *One Class Classification*

OC-NN *One Class Neural Network*

OVA *One Versus All*

OVR *One Versus Rest*

ReLU *Rectified Linear Unit*

RGB *Red, Green, Blue*

RGBD *Red, Green, Blue, Deep*

RBM *Restricted Boltzmann Machine*

Universitat d'Alacant
Universidad de Alicante

RNNs *Recurrent Neural Networks*

ROI *Region of Interest*

SDAEs *Stacked DAEs*

SL *Supervised Learning*

SOM *Self-Organizing Map*

SSOM *Supervised Self-Organizing Map*

SVM *Support Vector Machine*

UDF *Up Down Frecuency*

UL *Unsupervised Learning*

VAEs *Variational Autoencoders*

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Bibliografía

[1] Abati, D., Porrello, A., Calderara, S., and Cucchiara, R. (2019). Latent space autoregression for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 481–490. 67

[2] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*. 50

[3] Acampora, G., Foggia, P., Saggese, A., and Vento, M. (2012). Combining neural networks and fuzzy systems for human behavior understanding. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 88–93. IEEE. 5

[4] Adam, A., Rivlin, E., Shimshoni, I., and Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560. 52

[5] Afifi, M., Ali, Y., Amer, K., Shaker, M., and ElHelw, M. (2019). Robust real-time pedestrian detection in aerial imagery on jetson tx2. *arXiv preprint arXiv:1905.06653*. 22

[6] Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16. 28

[7] Al-Raziqi, A. and Denzler, J. (2016). Unsupervised Framework for Interactions Modeling between Multiple Objects. *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 4(Visigrapp):509–516. 48

[8] Ali, S. and Shah, M. (2005). A supervised learning framework for generic object detection in images. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1347–1354. IEEE. 7

[9] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., and Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3):292. 8

[10] Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, 19:1–9. 71

[11] Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, (3):299–307. 30

[12] Amer, M. R., Lei, P., and Todorovic, S. (2014). Hirf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision*, pages 572–585. Springer. 22

[13] Ammar, S., Anjum, M., Rounak, T., Islam, M., Islam, T., et al. (2019). *Using deep learning algorithms to detect violent activities*. PhD thesis, BRAC University. 22

[14] An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center, Tech. Rep.* 38

[15] Andrade, E. L., Blunsden, S., and Fisher, R. B. (2006). Modelling crowd scenes for event detection. In *18th international conference on pattern recognition (ICPR'06)*, volume 1, pages 175–178. IEEE. 49

[16] Atmosukarto, I., Ghanem, B., Ahuja, S., Muthuswamy, K., and Ahuja, N. (2013). Automatic recognition of offensive team formation in american football plays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 991–998. 22

[17] Autores, V. (1999). Ia, ml, dl. urlhttps://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/. 7

[18] Azorín-López, J., Saval-Calvo, M., Fuster-Guilló, A., and García-Rodríguez, J. (2013). Human behaviour recognition based on trajectory analysis using neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. 29, 66, 74, 76, 78

[19] Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., Cazorla, M., and Signes-Pont, M. T. (2016). Group activity description and recognition based on trajectory analysis and neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1585–1592. 29, 74, 79, 100, 101

[20] Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., and Orts-Escolano, S. (2015a). Self-organizing activity description map to represent and classify human behaviour. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. 66

[21] Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., Garcia-Rodriguez, J., and Orts-Escolano, S. (2015b). Self-organizing activity description map to represent and classify human behaviour. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. 74, 76

[22] Azorin-Lopez, J., Saval-Calvo, M., Fuster-Guillo, A., and Oliver-Albert, A. (2014). A predictive model for recognizing human behaviour based on trajectory representation. *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1494–1501. 48

[23] Azorin-López, J., Saval-Calvo, M., Fuster-Guilló, A., and Oliver-Albert, A. (2014). A predictive model for recognizing human behaviour based on trajectory representation. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1494–1501. IEEE. 29, 66, 74

[24] Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., and Savarese, S. (2017). Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4315–4324. 22, 61

[25] Baird, H. S. (1992). Document image defect models. In *Structured Document Image Analysis*, pages 546–556. Springer. 30

[26] Balakrishnama, S. and Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18:1–8. 28

[27] Banerjee, T., Yefimova, M., Keller, J. M., Skubic, M., Woods, D. L., and Rantz, M. (2017). Exploratory analysis of older adults' sedentary behavior in the primary living area using kinect depth data. *Journal of Ambient Intelligence and Smart Environments*, 9(2):163–179. 41

[28] Bao, Y., Ishii, N., and Du, X. (2004). Combining multiple k-nearest neighbor classifiers using different distance functions. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 634–641. Springer. 72

[29] Barekatain, M., Martí, M., Shih, H.-F., Murray, S., Nakayama, K., Matsuo, Y., and Prendinger, H. (2017). Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. 50

[30] Baro, X., Gonzalez, J., Fabian, J., Bautista, M. A., Oliu, M., Jair Escalante, H., Guyon, I., and Escalera, S. (2015). Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9. 50

[31] Barron, J. and Klette, R. (2002). Quantitative color optical flow. In *Object recognition supported by user interaction for service robots*, volume 4, pages 251–255. IEEE. 66

[32] Batchuluun, G., Kim, Y., Kim, J., Hong, H., and Park, K. (2016). Robust behavior recognition in intelligent surveillance environments. *Sensors*, 16(7):1010. 22

[33] Bay, S. D. (1998). Combining nearest neighbor classifiers through multiple feature subsets. In *ICML*, volume 98, pages 37–45. Citeseer. 72

[34] Benetazzo, F., Ferracuti, F., Freddi, A., Giantomassi, A., Iarlori, S., Longhi, S., Monteriù, A., and Ortenzi, D. (2015). Aal technologies for independent life of elderly people. In *Ambient Assisted Living*, pages 329–343. Springer. 6

[35] Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press. 72

[36] Biswas, S. and Gall, J. (2018). Structural recurrent neural network (srnn) for group activity analysis. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1625–1632. IEEE. 22

[37] Blunsden, S. and Fisher, R. (2010). The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4. 46, 48, 49, 50, 51, 52, 87, 93

[38] Boominathan, L., Kruthiventi, S. S., and Babu, R. V. (2016). Crowdnet: a deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 640–644. ACM. 22, 57

[39] Borja, L. F., Azorin-Lopez, J., and Saval-Calvo, M. (2017). A compilation of methods and datasets for group and crowd action recognition. *International Journal of Computer Vision and Image Processing (IJCVIP)*, 7(3):40–53. 28, 65

[40] Bour, P., Cribelier, E., and Argyriou, V. (2019). Crowd behavior analysis from fixed and moving cameras. In Alameda-Pineda, X., Ricci, E., and Sebe, N., editors, *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 289 – 322. Academic Press. 80

[41] Bozan, K. and Berger, A. (2019). Revisiting the technology challenges and proposing enhancements in ambient assisted living for the elderly. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 41

[42] Brostow, G. J. and Cipolla, R. (2006). Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601. IEEE. 49

[43] Bryson, A. E. (1961). A gradient method for optimizing multi-stage allocation processes. In *Proc. Harvard Univ. Symposium on digital computers and their applications*, volume 72. 30

[44] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167. 72, 73

[45] Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970. 50

[46] Caetano, C. A., De Melo, V. H. C., dos Santos, J. A., and Schwartz, W. R. (2017). Activity recognition based on a magnitude-orientation stream network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 47–54. IEEE. 22

[47] Camplani, M., Paiement, A., Mirmehdi, M., Damen, D., Hannuna, S., Burghardt, T., and Tao, L. (2016). Multiple human tracking in rgb-depth data: a survey. *IET computer vision*, 11(4):265–285. 48

[48] Chaaraoui, A., Padilla-López, J., Ferrández-Pastor, F., Nieto-Hidalgo, M., and Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context. *Sensors*, 14(5):8895–8925. 21, 40

[49] Chaaraoui, A. A., Climent-Pérez, P., and Flórez-Revuelta, F. (2012a). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888. 21, 26, 27, 40, 59

[50] Chaaraoui, A. A., Climent-Pérez, P., and Flórez-Revuelta, F. (2012b). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888. 28, 65

[51] Chaaraoui, A. A. and Flórez-Revuelta, F. (2013). Human action recognition optimization based on evolutionary feature subset selection. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, pages 1229–1236. ACM. 21, 40

[52] Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*. 12, 13, 14, 67

[53] Chalapathy, R., Menon, A. K., and Chawla, S. (2018a). Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*. 88

[54] Chalapathy, R., Toth, E., and Chawla, S. (2018b). Group anomaly detection using deep generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 173–189. Springer. 13

[55] Chan, A. B., Liang, Z.-S. J., and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE. 51, 52

[56] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15. 9, 10, 12, 14

[57] Chang, M. C., Krahnstoever, N., Lim, S., and Yu, T. (2010). Group level activity recognition in crowded environments across multiple cameras. *Proceedings - IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2010*, (February):56–63. 49

[58] Chen, K., Loy, C. C., Gong, S., and Xiang, T. (2012). Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3. 47, 52

[59] Chen, S., Feng, Z., Lu, Q., Mahasseni, B., Fiez, T., Fern, A., and Todorovic, S. (2014). Play type recognition in real-world football video. In *IEEE Winter Conference on Applications of Computer Vision*, pages 652–659. IEEE. 22

[60] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. 44, 45

Universidad de Alicante

[61] Cho, N.-G., Kim, Y.-J., Park, U., Park, J.-S., and Lee, S.-W. (2015). Group activity recognition with group interaction zone based on relative distance between human objects. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(05):1555007. 100, 101

[62] Choi, W., Chao, Y.-W., Pantofaru, C., and Savarese, S. (2014). Discovering groups of people in images. In *European conference on computer vision*, pages 417–433. Springer. 46, 48

[63] Choi, W. and Savarese, S. (2012). A unified framework for multi-target tracking and collective activity recognition. *Computer Vision–ECCV 2012*, pages 215–230. 22, 61

[64] Choi, W., Shahid, K., and Savarese, S. (2009). What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1282–1289. IEEE. 49

[65] Ciabattoni, L., Ferracuti, F., Freddi, A., Ippoliti, G., Longhi, S., Monteriù, A., and Pepa, L. (2016). Human indoor localization for aal applications: An rssi based approach. In *Italian Forum of Ambient Assisted Living*, pages 239–250. Springer. 6

[66] Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851. 30

[67] Cireşan, D., Meier, U., Masci, J., and Schmidhuber, J. (2011). A committee of neural networks for traffic sign classification. In *The 2011 international joint conference on neural networks*, pages 1918–1921. IEEE. 30

[68] Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*. 30

[69] Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer. 30

[70] Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220. 30

[71] Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*. 30

[72] Climent-Pérez, P., Mauduit, A., Monekosso, D. N., and Remagnino, P. (2014). Detecting events in crowded scenes using tracklet plots. *Proceedings of the International Conference on Computer Vision Theory and Applications*. 49

[73] Conigliaro, D., Rota, P., Setti, F., Bassetti, C., Conci, N., Sebe, N., and Cristani, M. (2015). The s-hock dataset: Analyzing crowds at the stadium. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2047. 52, 57

[74] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297. 72, 73

[75] Côté-Allard, U., Fall, C. L., Drouin, A., Campeau-Lecours, A., Gosselin, C., Glette, K., Laviolette, F., and Gosselin, B. (2019). Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(4):760–771. 6

[76] Craus, M. and Aflori, C. (2005). Association rules discovery using grid services. *SETIT*, pages 1–5. 7

[77] Crawford, V. P. (1974). Learning the optimal strategy in a zero-sum game. *Econometrica: Journal of the Econometric Society*, pages 885–891. 39

[78] Cucchiara, R., Grana, C., Prati, A., and Vezzani, R. (2004). Probabilistic posture classification for human-behavior analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35(1):42–54. 5

[79] Cupillard, F., Brémond, F., and Thonnat, M. (2002). Tracking groups of people for video surveillance. *Video-Based Surveillance Systems*. 49

[80] Dadlani, P., Gritti, T., Shan, C., de Ruyter, B., and Markopoulos, P. (2014). Sopresent: An awareness system for connecting remote households. In *European Conference on Ambient Intelligence*, pages 67–79. Springer. 41

[81] Dai, H.-N., Zheng, Z., and Zhang, Y. (2019). Blockchain for internet of things: A survey. *arXiv preprint arXiv:1906.00245*. 7

[82] Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387. 43

[83] Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736. 56

[84] de Alcântara, M. F., Moreira, T. P., and Pedrini, H. (2013). Motion silhouette-based real time action recognition. In *Iberoamerican Congress on Pattern Recognition*, pages 471–478. Springer. 21, 40

[85] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699. 6

[86] Deng, Z., Vahdat, A., Hu, H., and Mori, G. (2016). Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In

Universitat d'Alacant
Universidad de Alicante

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4772–4781. 4, 22, 61

[87] Deng, Z., Zhai, M., Chen, L., Liu, Y., Muralidharan, S., Roshtkhari, M. J., and Mori, G. (2015). Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191*. 4

[88] Devanne, M., Papadakis, P., et al. (2019). Recognition of activities of daily living via hierarchical long-short term memory networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3318–3324. IEEE. 41

[89] Dey, R. and Salem, F. M. (2017). Gate-variants of gated recurrent unit (gru) neural networks. *arXiv preprint arXiv*. 36, 57

[90] Ding, W., Liu, K., Fu, X., and Cheng, F. (2016). Profile hmms for skeleton-based human action recognition. *Signal Processing: Image Communication*, 42:109–119. 29, 75

[91] Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761. 52

[92] Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109. 47, 52

[93] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons. 72

[94] Ehsani, K., Bagherinezhad, H., Redmon, J., Mottaghi, R., and Farhadi, A. (2018). Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060. 43

[95] Ellis, A. and Ferryman, J. (2010). Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 135–142. IEEE. 46, 48, 49, 51, 52

[96] Ess, A., Leibe, B., Schindler, K., and Van Gool, L. (2008). A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 46, 48

[97] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338. 43, 44, 52

[98] Fabio, M., Pini, S., Borghi, G., Vezzani, R., and Cucchiara, R. (2019). Hand gestures for the human-car interaction: the briareo dataset. In *20th International Conference on Image Analysis and Processing*. 6

[99] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer. 66

[100] Fauzi, C., Sulistyo, S., et al. (2018). A survey of group activity recognition in smart building. In *2018 International Conference on Signals and Systems (ICSigSys)*, pages 13–19. IEEE. 28

[101] Feng, J., Zhang, S., and Xiao, J. (2019). Explorations of skeleton features for lstm-based action recognition. *Multimedia Tools and Applications*, 78(1):591–603. 29, 75

[102] Fisher, R. B. (2004). The pets04 surveillance ground-truth data sets. In *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–5. 46, 48, 49, 87, 93, 94

[103] Forrest, S. (2010). University of new mexico (unm) intrusion detection dataset. 46, 47, 48, 49

[104] Fradi, H. and Dugelay, J. L. (2016). Spatial and temporal variations of feature tracks for crowd behavior analysis. *Journal on Multimodal User Interfaces*, 10(4):307–317. 48

[105] Friedman, H., Friedman, J., and Friedman, J. (1996). Another approach to polychotomous classification. 71

[106] Frontoni, E., Mancini, A., and Zingaretti, P. (2014). Rgbd sensors for human activity detection in aal environments. In *Ambient Assisted Living*, pages 127–135. Springer. 6

[107] Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10):658–665. 30

[108] Fukushima, K. (2013). Artificial vision by multi-layered neural networks: Neocognitron and its advances. *Neural networks*, 37:103–119. 30

[109] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., and Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41 – 65. 65

[110] Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98. 28

[111] Ge, W. and Collins, R. T. (2009). Marked Point Processes for Crowd Counting. *IEEE Computer Vision and Pattern Recognition, 2009*, pages 2913–2920. 49

[112] Ge, W., Collins, R. T., and Ruback, R. B. (2012). Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016. 22, 49, 61

[113] Gers, F. A. and Schmidhuber, E. (2001). Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340. 36

[114] Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm. 30

Universitat d'Alacant
Universidad de Alicante

[115] Ghanem, B. and Ahuja, N. (2010). Maximum margin distance learning for dynamic texture recognition. In *European Conference on Computer Vision*, pages 223–236. Springer. 47

[116] Gibson, J. J. (1950). The perception of the visual world. 66

[117] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448. 43

[118] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587. 30, 43, 44, 45

[119] Gning, A., Mihaylova, L., Maskell, S., Pang, S. K., and Godsill, S. (2011). Group object structure and state estimation with evolving networks and Monte Carlo methods. *IEEE Transactions on Signal Processing*, 59(4):1383–1396. 49

[120] Gomez-Donoso, F., Escalona, F., Rivas, F. M., Cañas, J. M., and Cazorla, M. (2019). Enhancing the ambient assisted living capabilities with a mobile robot. *Computational intelligence and neuroscience*, 2019. 41

[121] Gong, S., Cristani, M., Yan, S., and Loy, C. C. (2014). Person Re-Identification. *Acvpr*. 49

[122] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680. 39

[123] Gorban, A., Idrees, H., Jiang, Y.-G., Zamir, A. R., Laptev, I., Shah, M., and Sukthankar, R. (2015). Thumos challenge: Action recognition with a large number of classes. 50

[124] Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., and Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098. 43

[125] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253. 50

[126] Gowsikhaa, D., Abirami, S., and Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765. 65

[127] Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (2017). The"something something"video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 3. 50

Universitat d'Alacant
Universidad de Alicante

[128] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM. 30

[129] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868. 30

[130] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610. 30

[131] Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552. 30

[132] Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer. 49

[133] Greitemeyer, T., Weiß, N., and Heuberger, T. (2019). Are everyday sadists specifically attracted to violent video games and do they emotionally benefit from playing those games? *Aggressive behavior*, 45(2):206–213. 22, 61

[134] Guan, Y. and Mao, W. (2019). Pedestrian virtual space based abnormal behavior detection. *IAENG International Journal of Computer Science*, 46(2). 22

[135] Ha, S., Yun, J.-M., and Choi, S. (2015). Multi-modal convolutional neural networks for activity recognition. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3017–3022. IEEE. 7

[136] Hajimirsadeghi, H. and Mori, G. (2015). Learning ensembles of potential functions for structured prediction with latent variables. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4059–4067. 41

[137] Hamidreza Rabiee, Javad Haddadnia, H. M. (2016). Emotion-Based Crowd Representation for Abnormality Detection Hamidreza. *International Journal on Artificial Intelligence Tools*. 28, 48

[138] Han, Y., Zhang, P., Zhuo, T., Huang, W., and Zhang, Y. (2018). Going deeper with two-stream convnets for action recognition in video surveillance. *Pattern Recognition Letters*, 107:83–90. 29, 75

[139] Hao, Y., Xu, Z.-J., Liu, Y., Wang, J., and Fan, J.-L. (2019). Effective crowd anomaly detection through spatio-temporal texture analysis. *International Journal of Automation and Computing*, 16(1):27–39. 22

Universidad de Alicante

[140] Hassner, T., Itcher, Y., and Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE. 46, 48

[141] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916. 43

[142] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 33, 39, 44

[143] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer. 31, 33

[144] Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., and Beslay, L. (2019). Faceqnet: Quality assessment for face recognition based on deep learning. *arXiv preprint arXiv:1904.01740*. 6

[145] Hinton, G. E. (1990). Connectionist learning procedures. In *Machine learning*, pages 555–610. Elsevier. 30

[146] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507. 30

[147] Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1). 30

[148] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 30

[149] Hospedales, T., Gong, S., and Xiang, T. (2009). A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1165–1172. IEEE. 52

[150] Hothorn, T. (2019). Cran task view: Machine learning & statistical learning. 7

[151] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425. 72, 73

[152] Hu, G., Cui, B., and Yu, S. (2018). Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention. *arXiv preprint arXiv:1811.04237*. 82, 84

[153] Hu, J., Zhu, E., Wang, S., Liu, X., Guo, X., and Yin, J. (2019). An efficient and robust unsupervised anomaly detection method using ensemble random projection in surveillance videos. *Sensors*, 19(19):4145. 103, 106

[154] Hu, Y., Chang, H., Nian, F., Wang, Y., and Li, T. (2016). Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38:530–539. 52

[155] Huang, C.-D., Wang, C.-Y., and Wang, J.-C. (2015). Human action recognition system for elderly and children care using three stream convnet. In *2015 International Conference on Orange Technologies (ICOT)*, pages 5–9. IEEE. 84

[156] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. 30, 31, 34

[157] Huang, H., Chong, Y., Nie, C., and Pan, S. (2019). Hand gesture recognition with skin detection and deep learning method. In *Journal of Physics: Conference Series*, volume 1213, page 022001. IOP Publishing. 6

[158] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154. 30

[159] Hussein, N., Gavves, E., and Smeulders, A. W. (2019a). Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263. 22, 60

[160] Hussein, N., Gavves, E., and Smeulders, A. W. (2019b). Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*. 22

[161] Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2016). A hierarchical deep temporal model for group activity recognition. pages 1971–1980. 4

[162] Idrees, H., Saleemi, I., Seibert, C., and Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554. 51, 52

[163] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. 31, 33

[164] Ionescu, R. T., Khan, F. S., Georgescu, M.-I., and Shao, L. (2019). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851. 106

[165] Iosifidis, A., Marami, E., Tefas, A., Pitas, I., and Lyroudia, K. (2015). The mobiserv-aiia eating and drinking multi-view database for vision-based assisted living. *Journal of Information Hiding and Multimedia Signal Processing*, 6(2):254–273. 50

[166] Jackman, S. (2019). Football shot detection using convolutional neural networks. 22, 61

[167] Jacques, J. C. S., Braun, A., Soldera, J., Musse, S. R., and Jung, C. R. (2007). Understanding people motion in video sequences using Voronoi diagrams. *Pattern Analysis and Applications*, 10(4):321–332. 49

Universitat d'Alacant
Universidad de Alicante

[168] Jagannath, J., Polosky, N., Jagannath, A., Restuccia, F., and Melodia, T. (2019). Machine learning for wireless communications in the internet of things: a comprehensive survey. *Ad Hoc Networks*, page 101913. 7

[169] Jalal, A., Kim, J. T., and Kim, T.-S. (2012). Human activity recognition using the labeled depth body parts information of depth silhouettes. In *Proceedings of the 6th International Symposium on Sustainable Healthy Buildings, Seoul, Korea*, volume 27. 40

[170] Jalal, A., Mahmood, M., and Hasan, A. S. (2019). Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 371–376. IEEE. 40

[171] Jhuang, H., Garrote, H., Poggio, E., Serre, T., and Hmdb, T. (2011). A large video database for human motion recognition. In *Proc. of IEEE International Conference on Computer Vision*, volume 4, page 6. 46

[172] Jiang, H. and Learned-Miller, E. (2017). Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 650–657. IEEE. 43

[173] Jiang, W. and Yin, Z. (2015). Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1307–1310. Acm. 7

[174] Jiang, Y.-G., Wu, Z., Wang, J., Xue, X., and Chang, S.-F. (2017). Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):352–364. 50

[175] Jing, L., Ye, Y., Yang, X., and Tian, Y. (2017). 3d convolutional neural network with multi-model framework for action recognition. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1837–1841. IEEE. 29, 75

[176] Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., and Sato, T. (2012). Viewpoint invariant collective activity recognition with relative action context. In *European Conference on Computer Vision*, pages 253–262. Springer. 41

[177] Kang, D., Dhar, D., and Chan, A. B. (2016). Crowd counting by adapting convolutional neural networks with side information. *arXiv preprint arXiv:1611.06748*. 22, 57

[178] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732. 50

[179] Karthikeyan, T. and Ravikumar, N. (2014). A survey on association rule mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(1):2278–1021. 7

[180] Ke, Q., Liu, J., Bennamoun, M., An, S., Sohel, F., and Boussaid, F. (2018). Computer vision for human–machine interaction. In Leo, M. and Farinella, G. M., editors, *Computer Vision for Assistive Healthcare*, Computer Vision and Pattern Recognition, pages 127 – 145. Academic Press. 79, 80

[181] Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954. 30

[182] Khan, S., Islam, N., Jan, Z., Din, I. U., and Rodrigues, J. J. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6. 7

[183] Khan, S. S. and Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374. 69

[184] Kilambi, P., Ribnick, E., Joshi, A. J., Masoud, O., and Papanikolopoulos, N. (2008). Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110(1):43–59. 49

[185] Kim, C. and Hwang, K.-B. (2008). Naive bayes classifier learning with feature selection for spam detection in social bookmarking. In *Proc. Europ. Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. 72

[186] Kiran, B. R., Thomas, D. M., and Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *arXiv preprint arXiv:1801.03149*. 37, 38

[187] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480. 28

[188] Koller, O., Camgoz, C., Ney, H., and Bowden, R. (2019). Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*. 82, 84

[189] Kopinski, T., Magand, S., Handmann, U., and Gepperth, A. (2015). A pragmatic approach to multi-class classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. 73

[190] Köpüklü, O., Gunduz, A., Kose, N., and Rigoll, G. (2019). Real-time hand gesture detection and classification using convolutional neural networks. *arXiv preprint arXiv:1901.10323*. 6

[191] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. 30, 31, 32, 33, 39, 44, 87

[192] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L.,

Universitat d'Alacant
Universidad de Alicante

and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc. 42, 52

[193] Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A. J., and Wiskott, L. (2012). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1847–1871. 84

[194] Kuehne, H., Arslan, A., and Serre, T. (2014). The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787. 21, 40, 55

[195] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE. 50

[196] Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., and Kim, K. J. (2017). A survey of deep learning-based network anomaly detection. *Cluster Computing*, pages 1–13. 11

[197] Lan, T., Wang, Y., and Mori, G. (2011). Discriminative figure-centric models for joint action localization and recognition. In *2011 International conference on computer vision*, pages 2003–2010. IEEE. 50

[198] Laptev, I. and Lindeberg, T. (2004). Velocity adaptation of space-time interest points. In *17th International Conference on Pattern Recognition (ICPR) Location: British Machine Vis Assoc, Cambridge, ENGLAND Date: AUG 23-26, 2004*, pages 52–56. IEEE conference proceedings. 50

[199] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 50

[200] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*. 30, 31, 34

[201] Lathuilière, S., Evangelidis, G., and Horaud, R. (2017). Recognition of group activities in videos based on single-and two-person descriptors. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 217–225. IEEE. 22

[202] Lau, B., Arras, K. O., and Burgard, W. (2010). Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics*, 2(1):19–30. 49

[203] Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750. 45

[204] Lebanoff, L. and Idrees, H. (2015). Counting in dense crowds using deep learning. 22

[205] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. 13

[206] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551. 30

[207] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404. 30

[208] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. 30, 32

[209] Lee, S., Kim, H. G., and Ro, Y. M. (2018). Stan: Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327. IEEE. 103, 106

[210] Li, J. and Song, J. (2016). Pedestrian counting via deep convolutional neural networks in crowded scene. In *2016 4th International Conference on Advanced Materials and Information Technology Processing (AMITP 2016)*. Atlantis Press. 22, 57

[211] Li, L.-J., Su, H., Fei-Fei, L., and Xing, E. P. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386. 52

[212] Li, R., Chellappa, R., and Zhou, S. K. (2009). Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2450–2457. IEEE. 22

[213] Li, W., Chang, M.-C., and Lyu, S. (2018). Who did what at where and when: Simultaneous multi-person tracking and activity recognition. *arXiv preprint arXiv:1807.01253*. 22, 61

[214] Li, W., Mahadevan, V., and Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32. 48

[215] Li, X. and Choo Chuah, M. (2017). Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2876–2885. 22

[216] Li, X., He, Y., and Jing, X. (2019a). A survey of deep learning-based human activity recognition in radar. *Remote Sensing*, 11(9):1068. 32, 35, 36, 37

[217] Li, Y., Cai, Y., Liu, J., Lang, S., and Zhang, X. (2019b). Spatio-temporal unity networking for video anomaly detection. *IEEE Access*, 7:172425–172432. 103, 106

[218] Liang, D., Fan, G., Lin, G., Chen, W., Pan, X., and Zhu, H. (2019). Three-stream convolutional neural network with multi-task and ensemble learning for 3d action

Universitat d'Alacant
Universidad de Alicante

recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0. 82, 84

[219] Liao, H., Xiang, J., Sun, W., Feng, Q., and Dai, J. (2011). An abnormal event recognition in crowd scene. *Proceedings - 6th International Conference on Image and Graphics, ICIG 2011*, (September 2011):731–736. 49

[220] Lim, H. J. (2015). *Crime reduction effects of open-street CCTVs in Cincinnati*. PhD thesis, University of Cincinnati. 4

[221] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*. 30, 34

[222] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125. 44

[223] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988. 39, 43, 84

[224] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer. 44, 45

[225] Lin, W., Sun, M.-T., Poovandran, R., and Zhang, Z. (2008). Human activity recognition for video surveillance. *IEEE International Symposium on Circuits and Systems*, (June):2737–2740. 28, 48

[226] Lio, V. (2019). La extensión de la videovigilancia en el territorio bonaerense. *Geograficando*, 15(1):e054–e054. 4

[227] Liu, H., Tu, J., and Liu, M. (2017a). Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv preprint arXiv:1705.08106*. 82, 84

[228] Liu, J., Tong, X., Li, W., Wang, T., Zhang, Y., Wang, H., Yang, B., Sun, L., and Yang, S. (2007). Automatic Player Detection, Labeling and Tracking in Broadcast Soccer Video. *Procedings of the British Machine Vision Conference 2007*, pages 3.1–3.10. 48

[229] Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K., and Kot, A. C. (2017b). Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599. 29, 75

[230] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer. 44, 84

[231] Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23. 72

[232] López, L. C. J. (2019). Hacia la transterritorialización de la política pública de videovigilancia en méxico. *ANUARIO DE ESPACIOS URBANOS, HISTORIA, CULTURA Y DISEÑO*, (25). 4

[233] Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727. 52, 93, 101

[234] Lu, Y., Kumar, K. M., shahabeddin Nabavi, S., and Wang, Y. (2019). Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE. 103, 106

[235] Luvizon, D. C., Tabia, H., and Picard, D. (2017). Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, 99:13–20. 29, 75

[236] Lyu, L., Bezdek, J. C., He, X., and Jin, J. (2019). Fog-embedded deep learning for the internet of things. *IEEE Transactions on Industrial Informatics*. 7

[237] Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE. 22, 47, 48, 52, 93, 101

[238] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*. 38

[239] Mallick, P. K., Ryu, S. H., Satapathy, S. K., Mishra, S., Nguyen, G. N., and Tiwari, P. (2019). Brain mri image classification for cancer detection using deep wavelet autoencoder-based deep neural network. *IEEE Access*, 7:46278–46287. 84

[240] Marsden, M., McGuinness, K., Little, S., and O'Connor, N. E. (2017). Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. *arXiv preprint arXiv:1705.10698*. 4

[241] Marszałek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*, pages 2929–2936. IEEE Computer Society. 50

[242] Martinetz, T. M., Berkovich, S. G., and Schulten, K. J. (1993). 'neural-gas' network for vector quantization and its application to time-series prediction. *IEEE transactions on neural networks*, 4(4):558–569. 28

[243] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer. 37

[244] Maymin, P. (2013). Acceleration in the nba: Towards an algorithmic taxonomy of basketball plays. MIT Sloan Sports Analytics Conference. 46

[245] McKenna, S. J., Jabri, S., Duric, Z., and Wechsler, H. (2000). Tracking interacting people. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 348–353. IEEE. 28

Universitat d'Alacant
Universidad de Alicante

[246] Medsker, L. and Jain, L. (2001). Recurrent neural networks. *Design and Applications*, 5. 35

[247] Mehran, R., Oyama, A., and Shah, M. (2009a). Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE. 52

[248] Mehran, R., Oyama, A., and Shah, M. (2009b). Abnormal Crowd Behaviour Detection using Social Force Model. *IEEE Conference on Computer Vision and Pattern Recognition*, (2):935–942. 49

[249] Micilotta, A. S., Ong, E.-J., and Bowden, R. (2005). Detection and tracking of humans by probabilistic body part assembly. In *BMVC*, number 1, pages 429–438. 28

[250] Moencks, M., De Silva, V., Roche, J., and Kondoz, A. (2019). Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset. *arXiv preprint arXiv:1901.02858*. 40

[251] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126. 24, 25, 26

[252] Moisen, G. (2008). Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, pages 582–588. 72

[253] Morris, B. T. and Trivedi, M. M. (2008). A survey of vision-based trajectory learning and analysis for surveillance. *IEEE transactions on circuits and systems for video technology*, 18(8):1114–1127. 66

[254] Moschevikin, A., Galov, A., Volkov, A., Mikov, A., Reginya, S., Voronov, R., Reut, O., Serezhina, M., Zaitsev, A., Lunkov, P., et al. (2015). Realtrac technology at the evaal-2013 competition. *Journal of Ambient Intelligence and Smart Environments*, 7(3):353–373. 41

[255] Moya, M. M., Koch, M. W., and Hostetler, L. D. (1993). One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N*, 93. 67

[256] Muhammad, K., Hussain, T., and Baik, S. W. (2018). Efficient cnn based summarization of surveillance videos for resource-constrained devices. *Pattern Recognition Letters*. 22

[257] Münch, D., Michaelsen, E., and Arens, M. (2012). Supporting fuzzy metric temporal logic based situation recognition by mean shift clustering. In *Annual Conference on Artificial Intelligence*, pages 233–236. Springer. 100, 101

[258] Nascimento, J. C. and Marques, J. S. (2006). Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4):761–774. 65

[259] Nasser, I. M. and Abu-Naser, S. S. (2019). Lung cancer detection using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS)*, 3(3):17–23. 84

[260] Nayan, N., Sahu, S. S., and Kumar, S. (2019). Detecting anomalous crowd behavior using correlation analysis of optical flow. *Signal, Image and Video Processing*, pages 1–9. 22, 61

[261] Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., Malík, P., and Hluchỳ, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, 52(1):77–124. 7

[262] Nguyen, T.-N. and Meunier, J. (2019a). Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283. 103, 106

[263] Nguyen, T. N. and Meunier, J. (2019b). Hybrid deep network for anomaly detection. *arXiv preprint arXiv:1908.06347*. 103, 106

[264] Niebles, J. C., Chen, C.-W., and Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pages 392–405. Springer. 50

[265] Ojha, S. and Sakhare, S. (2015). Image processing techniques for object tracking in video surveillance-a survey. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–6. IEEE. 65

[266] Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., and Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In *European conference on computer vision*, pages 28–39. Springer. 28

[267] Omae, Y., Mori, M., Akiduki, T., and Takahashi, H. (2019). A novel deep learning optimization algorithm for human motions anomaly detection. *INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING INFORMATION AND CONTROL*, 15(1):199–208. 5

[268] Onofri, L. and Soda, P. (2012). Combining video subsequences for human action recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 597–600. IEEE. 21, 40

[269] Pantic, M. and Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449. 28

[270] Papadimitriou, S., Mavroudi, S., Vladutu, L., Pavlides, G., and Bezerianos, A. (2002). The supervised network self-organizing map for classification of large data sets. *Applied intelligence*, 16(3):185–203. 28

Universitat d'Alacant
Universidad de Alicante

[271] Papadopoulos, S., Troncy, R., Mezaris, V., Huet, B., and Kompatsiaris, I. (2011). Social event detection at mediaeval 2011: Challenges, dataset and evaluation. In *MediaEval*. 52

[272] Parrilli, F. A. (2019). Ingeniería y sociedad de control: sobre diseño y videovigilancia pública. *Tecnología y Sociedad*, 1(2):11–37. 4

[273] Pece, A. E. (2002). From cluster tracking to people counting. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, pages 9–17. 28

[274] Pellegrini, S., Ess, A., Schindler, K., and Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE. 22, 46, 49

[275] Pellegrini, S., Ess, A., and Tanaskovic, M. (2010). Wrong turn-no dead end: a stochastic pedestrian motion model. *Computer Vision and*. 22, 49, 61

[276] Pereira, F. C. and Borysov, S. S. (2019). Machine learning fundamentals. In *Mobility Patterns, Big Data and Transport Analytics*, pages 9–29. Elsevier. 8

[277] Perera, P. and Patel, V. M. (2019). Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463. 67

[278] Perše, M., Kristan, M., Kovačič, S., Vučkovič, G., and Perš, J. (2009). A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, 113(5):612–621. 49

[279] Péteri, R., Fazekas, S., and Huiskes, M. J. (2010). Dyntex: A comprehensive database of dynamic textures. *Pattern Recognition Letters*, 31(12):1627–1632. 47

[280] Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., and Velastin, S. A. (2018). Learning to recognise 3d human action from a new skeleton-based representation using deep convolutional neural networks. *IET Computer Vision*, 13(3):319–328. 29, 75

[281] Piza, E. L., Welsh, B. C., Farrington, D. P., and Thomas, A. L. (2019). Cctv surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology & Public Policy*, 18(1):135–159. 4

[282] Pollack, J. B. (1989). Implications of recursive distributed representations. In *Advances in neural information processing systems*, pages 527–536. 30

[283] Prati, A., Shan, C., and Wang, K. I.-K. (2019). Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. *Journal of Ambient Intelligence and Smart Environments*, 11(1):5–22. 22, 61

[284] Purkait, P., Zhao, C., and Zach, C. (2017). Spp-net: Deep absolute pose regression with synthetic views. *arXiv preprint arXiv:1712.03452*. 43

[285] Puttemans, S. and Goedemé, T. (2015). Visual detection and species classification of orchid flowers. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 505–509. IEEE. 84

Universitat d'Alacant
Universidad de Alicante

[286] Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., and Van Gool, L. (2018). stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117. 22

[287] Qu, Y., Fang, Y., and Yan, F. (2019). Feature selection algorithm based on association rules. In *Journal of Physics: Conference Series*, volume 1168, page 052012. IOP Publishing. 7

[288] Rabiner, L. R. and Juang, B.-H. (1986). An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16. 28

[289] Ramasinghe, S. and Rodrigo, R. (2015). Action recognition by single stream convolutional neural networks: An approach using combined motion and static information. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 101–105. IEEE. 22

[290] Ramotsoela, D., Abu-Mahfouz, A., and Hancke, G. (2018). A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study. *Sensors*, 18(8):2491. 13

[291] Ranzato, M., Poultney, C., Chopra, S., and Cun, Y. L. (2007). Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144. 30

[292] Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., and Sebe, N. (2016). Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. *arXiv preprint arXiv:1610.00307*. 4

[293] Reddy, K. K. and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981. 50

[294] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788. 39

[295] Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. 39, 43, 84

[296] Ren, J., Reyes, N., Barczak, A., Scogings, C., and Liu, M. (2018). An investigation of skeleton-based optical flow-guided features for 3d action recognition using a multi-stream cnn model. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 199–203. IEEE. 82, 84

[297] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99. 43, 84

[298] Ribeiro, M., Lazzaretti, A. E., and Lopes, H. S. (2018). A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*, 105:13–22. 103, 106

[299] Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 833–840. Omnipress. 38

[300] Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. 72

[301] Rodrigues, F. and Pereira, F. (2017). Deep learning from crowds. *arXiv preprint arXiv:1709.01779*. 4

[302] Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011). Data-driven crowd analysis in videos. In *2011 International Conference on Computer Vision*, pages 1235–1242. IEEE. 48

[303] Rohrbach, M., Amin, S., Andriluka, M., and Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE. 50

[304] Ronfard, R., Schmid, C., and Triggs, B. (2002). Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714. Springer. 28

[305] Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., and Gurcan, M. N. (2013). Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, 4. 30

[306] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018a). Deep one-class classification. In *International Conference on Machine Learning*, pages 4393–4402. 13, 67

[307] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018b). Deep one-class classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402, Stockholmsmässan, Stockholm Sweden. PMLR. 88

[308] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. 39

[309] Ryoo, M. S. and Aggarwal, J. K. (2008). Recognition of high-level group activities based on activities of individual members. *2008 IEEE Workshop on Motion and Video Computing, WMVC*, 2008(January). 49

[310] Ryoo, M. S. and Aggarwal, J. K. (2009). Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In *ICCV*, volume 1, page 2. Citeseer. 50

Universitat d'Alacant
Universidad de Alicante

[311] Ryoo, M. S., Piergiovanni, A., Tan, M., and Angelova, A. (2019). Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*. 82, 84

[312] Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., and Klette, R. (2018a). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97. 52

[313] Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. (2018b). Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388. 67

[314] Sabri, N., Ibrahim, Z., and Rosman, N. N. (2016). K-means vs. fuzzy c-means for segmentation of orchid flowers. In *2016 7th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pages 82–86. IEEE. 84

[315] Salehinejad, H., Baarbe, J., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*. 35

[316] Sani, M. M., Kutty, S. B., Omar, H. A., and Isa, I. N. M. (2013). Classification of orchid species using neural network. In *2013 IEEE International Conference on Control System, Computing and Engineering*, pages 586–589. IEEE. 84

[317] Sathasivam, S. and Abdullah, W. A. T. W. (2008). Logic learning in hopfield networks. *arXiv preprint arXiv:0804.4075*. 35

[318] Saval-Calvo, M., Azorín-López, J., and Fuster-Guilló, A. (2013). Comparative analysis of temporal segmentation methods of video sequences. In *Robotic Vision: Technologies for Machine Learning and Vision Applications*, pages 43–58. IGI Global. 65

[319] Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., and Busch, C. (2019). Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026. 6

[320] Schmidhuber, J. (2013). My first deep learning system of 1991+ deep learning timeline 1962-2013. *arXiv preprint arXiv:1312.5548*. 29

[321] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117. 8

[322] Schmidhuber, J., Cireşan, D., Meier, U., Masci, J., and Graves, A. (2011). On fast deep nets for agi vision. In *International Conference on Artificial General Intelligence*, pages 243–246. Springer. 30

[323] Schuldt, C., Barbara, L., and Stockholm, S. (2004). Recognizing Human Actions : A Local SVM Approach * Dept . of Numerical Analysis and Computer Science. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 3:32–36. 49

[324] Sejnowski, T. J. and Rosenberg, C. R. (1987). Parallel networks that learn to pronounce english text. *Complex systems*, 1(1):145–168. 71

Universitat d'Alacant
Universidad de Alicante

[325] Sendo, K. and Ukita, N. (2019). Heatmapping of people involved in group activities. 22

[326] Shafiee, M. J., Chywl, B., Li, F., and Wong, A. (2017). Fast yolo: a fast you only look once system for real-time embedded object detection in video. *arXiv preprint arXiv:1709.05943*. 39, 84

[327] Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019. 54

[328] Shang, C., Ai, H., and Bai, B. (2016). End-to-end crowd counting via joint learning local and global count. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1215–1219. IEEE. 52

[329] Shao, J., Kang, K., Change Loy, C., and Wang, X. (2015). Deeply learned attributes for crowded scene understanding. pages 4657–4666. 4, 22, 61

[330] Shao, J., Loy, C. C., Kang, K., and Wang, X. (2016). Crowded scene understanding by deeply learned volumetric slices. *IEEE transactions on circuits and systems for video technology*, 27(3):613–623. 51, 52

[331] Shao, J., Loy, C. C., and Wang, X. (2014). Scene-Independent Group Profiling in Crowd. pages 2219–2226. 48

[332] Sharma, N., Jain, V., and Mishra, A. (2018). An analysis of convolutional neural networks for image classification. *Procedia computer science*, 132:377–384. 7

[333] Shehab, D. and Ammar, H. (2019). Statistical detection of a panic behavior in crowded scenes. *Machine Vision and Applications*, 30(5):919–931. 22

[334] Shen, C., Xie, R., Zhang, L., and Song, L. (2015). Small group people behavior analysis based on temporal recursive trajectory identification. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE. 48

[335] Shu, T., Todorovic, S., and Zhu, S.-C. (2017). Cern: Confidence-energy recurrent network for group activity recognition. *arXiv preprint arXiv:1704.03058*. 4, 22, 61

[336] Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3. 30

[337] Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576. 84

[338] Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 31, 33, 43

[339] Simonyan, K. and Zisserman, A. (2015). Two-stream convolutional networks for action recognition. In *Proceedings of the Neural Information Processing Systems (NIPS)*. 84

[340] Sindagi, V. A. and Patel, V. M. (2017a). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE. 22

[341] Sindagi, V. A. and Patel, V. M. (2017b). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*. 22, 57, 61

[342] Singh, B., Marks, T. K., Jones, M., Tuzel, O., and Shao, M. (2016). A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970. 50

[343] Singh, T. and Vishwakarma, D. K. (2019). Video benchmarks of human action datasets: a review. *Artificial Intelligence Review*, 52(2):1107–1154. 47

[344] Slotine, J.-J. E. and Yu, G. (2013). Fast pattern classification based on a sparse transform. US Patent 8,553,984. 7

[345] Smith, K., Gatica-Perez, D., and Odobez, J.-M. (2005). Using particles to track varying numbers of interacting people. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 962–969. IEEE. 28

[346] Solmaz, B., Moore, B. E., and Shah, M. (2012). Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2064–2070. 48, 52

[347] Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Thirty-first AAAI conference on artificial intelligence*. 29, 75

[348] Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R., and Wang, A. (2019). A novel violent video detection scheme based on modified 3d convolutional neural networks. *IEEE Access*, 7:39172–39179. 22

[349] Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*. 50

[350] Souza, J. T. d., Francisco, A. C. d., Piekarski, C. M., and Prado, G. F. d. (2019). Data mining and machine learning to promote smart cities: A systematic review from 2000 to 2018. *Sustainability*, 11(4):1077. 7

[351] Spinello, L. and Arras, K. O. (2011). People detection in rgb-d data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3838–3843. IEEE. 46, 48

[352] Steinberg, D. (2009). Cart: classification and regression trees. In *The top ten algorithms in data mining*, pages 193–216. Chapman and Hall/CRC. 72

Universitat d'Alacant
Universidad de Alicante

[353] Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., and Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111. 7

[354] Sun, L., Jia, K., Yeung, D.-Y., and Shi, B. E. (2015). Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605. 42

[355] Sung, J., Ponce, C., Selman, B., and Saxena, A. (2012). Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation*, pages 842–849. IEEE. 53

[356] Sylvester, J. J. (1857). A question in the geometry of situation. *Quarterly Journal of Pure and Applied Mathematics*, 1(1):79–80. 67, 68, 69

[357] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284. 30, 31, 33

[358] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9. 30, 31, 33

[359] Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590. 7

[360] Tian, Y., Luo, P., Wang, X., and Tang, X. (2015). Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087. 22, 57

[361] Tran, D. and Sorokin, A. (2008). Human activity recognition with metric learning. In *European conference on computer vision*, pages 548–561. Springer. 50

[362] Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473. 25, 26, 28

[363] Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer. 6

[364] Ullah, A., Muhammad, K., Haq, I. U., and Baik, S. W. (2019a). Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments. *Future Generation Computer Systems*, 96:386–397. 22, 60

[365] Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., and Baik, S. W. (2019b). Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472. 22

Universitat d'Alacant
Universidad de Alicante

[366] Vahora, S. and Chauhan, N. (2018). Group activity recognition using deep auto-encoder with temporal context descriptor. *International Journal of Next-Generation Computing*, 9(3). 41

[367] Vahora, S. A. and Chauhan, N. C. (2019). Deep neural network model for group activity recognition using contextual relationship. *Engineering Science and Technology, an International Journal*, 22(1):47–54. 41

[368] Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., and Murino, V. (2016). Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143(May 2016):11–24. 48

[369] Veta, M., Viergever, M., Pluim, J., Stathonikos, N., and van Diest, P. (2013). Miccai 2013 grand challenge on mitosis detection. 30

[370] Vishwakarma, D., Rawat, P., and Kapoor, R. (2015). Human activity recognition using gabor wavelet transform and ridgelet transform. *Procedia Computer Science*, 57:630–636. 41

[371] Vishwakarma, S. and Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009. 25, 26, 59

[372] Vu, H., Nguyen, T. D., Le, T., Luo, W., and Phung, D. (2019). Robust anomaly detection in videos using multilevel representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5216–5223. 103, 106

[373] Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11. 7, 36, 40

[374] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE. 54

[375] Wang, L., Ge, L., Li, R., and Fang, Y. (2017a). Three-stream cnns for action recognition. *Pattern Recognition Letters*, 92:33–40. 29, 75

[376] Wang, M., Ni, B., and Yang, X. (2017b). Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8. 4, 22

[377] Wang, S., Zeng, Y., Liu, Q., Zhu, C., Zhu, E., and Yin, J. (2018). Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 636–644. 104, 106

[378] Wang, X., Xu, H., Zheng, S., and Cheng, A. (2003). Evidential reasoning research on intrusion detection. In *Fifth International Symposium on Instrumentation and Control Technology*, volume 5253, pages 930–934. International Society for Optics and Photonics. 28

Universitat d'Alacant
Universidad de Alicante

[379] Wang, Y., Huang, K., and Tan, T. (2007). Human activity recognition based on r transform. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 50

[380] Wei, L. and Shah, S. K. (2017). Human activity recognition using deep neural network with contextual information. In *VISIGRAPP (5: VISAPP)*, pages 34–43. 41

[381] Weinland, D., Boyer, E., and Ronfard, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE. 50

[382] Welzl, E. (1991). Smallest enclosing disks (balls and ellipsoids). In *New results and new trends in computer science*, pages 359–370. Springer. 69

[383] Wu, C., Zhang, J., Sener, O., Selman, B., Savarese, S., and Saxena, A. (2016). Watch-n-patch: Unsupervised learning of actions and relations. *CoRR*, abs/1603.03541. 41, 53

[384] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., and Rehg, J. M. (2007). A scalable approach to activity recognition based on object use. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE. 25, 26

[385] Wu, J., Wang, L., Wang, L., Guo, J., and Wu, G. (2019). Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9964–9974. 22

[386] Xu, C., Chai, D., He, J., Zhang, X., and Duan, S. (2019a). Innohar: a deep neural network for complex human activity recognition. *IEEE Access*, 7:9893–9902. 22

[387] Xu, S., Li, S., Wen, R., and Huang, W. (2019b). Traffic event detection using twitter data based on association rules. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 543–547. 7

[388] Yang, B., Cao, X., Han, Z., and Qian, L. (2019). A machine learning enabled mac framework for heterogeneous internet-of-things networks. *IEEE Transactions on Wireless Communications*. 7

[389] Yazdi, M. and Bouwmans, T. (2018). New trends on moving object detection in video images captured by a moving camera: A survey. *Computer Science Review*, 28:157–177. 65

[390] Ye, M., Peng, X., Gan, W., Wu, W., and Qiao, Y. (2019). Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813. 104, 106

[391] Yi, S. (2016). *Pedestrian Behavior Modeling and Understanding in Crowds*. PhD thesis, The Chinese University of Hong Kong (Hong Kong). 52

[392] Yi, S., Li, H., and Wang, X. (2015a). Pedestrian Travel Time Estimation in Crowded Scenes Shenzhen Institutes of Advanced Technology , Chinese Academy of Sciences. pages 3137–3145. 22, 48

Universitat d'Alacant
Universidad de Alicante

[393] Yi, S., Li, H., and Wang, X. (2015b). Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496. 48

[394] Yi, S., Li, H., and Wang, X. (2015c). Understanding pedestrian behaviors from stationary crowd groups. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:3488–3496. 48

[395] Yin, Y., Yang, G., and Man, H. (2013a). Small human group detection and event representation based on cognitive semantics. *Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013*, (September 2013):64–69. 49

[396] Yin, Y., Yang, G., and Man, H. (2013b). Small human group detection and event representation based on cognitive semantics. In *2013 IEEE seventh international conference on semantic computing*, pages 64–69. IEEE. 100, 101

[397] Yoo, Y., Yun, K., Yun, S., Hong, J., Jeong, H., and Young Choi, J. (2016). Visual path prediction in complex scenes with crowded moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2668–2677. 52

[398] Young, M., Katell, M., and Krafft, P. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data & Society*, 6(2):2053951719868492. 4

[399] Yu, Z., Li, T., Yu, N., Gong, X., Chen, K., and Pan, Y. (2017). Three-stream convolutional networks for video-based person re-identification. *arXiv preprint arXiv:1712.01652*. 84

[400] Yun, I., Jung, C., Wang, X., Hero, A. O., and Kim, J. K. (2019). Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment. *IEEE Access*, 7:23027–23037. 22

[401] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer. 30, 31, 33

[402] ZENG, Q.-s., YU, M.-h., HE, W.-g., and LI, L. (2009). A new algorithm of action recognition. *Journal of Kunming University of Science and Technology (Science and Technology)*, (6):13. 28

[403] Zhang, C., Li, H., Wang, X., and Yang, X. (2015a). Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841. 22, 48, 57, 61

[404] Zhang, C., Sun, X., Dang, K., Li, K., Guo, X.-w., Chang, J., Yu, Z.-q., Huang, F.-y., Wu, Y.-s., Liang, Z., et al. (2019a). Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *The Oncologist*, 24(9):1159–1165. 84

[405] Zhang, C., Yang, X., Lin, W., and Zhu, J. (2012a). Recognizing human group behaviors with multi-group causalities. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2012*, pages 44–48. 49

Universitat d'Alacant
Universidad de Alicante

[406] Zhang, C., Yang, X., Lin, W., and Zhu, J. (2012b). Recognizing human group behaviors with multi-group causalities. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 44–48. IEEE. 100, 101

[407] Zhang, H. and Ling, C. X. (2003). A fundamental issue of naive bayes. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 591–595. Springer. 72

[408] Zhang, J., Zheng, Y., and Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, pages 1655–1661. 4

[409] Zhang, X., Lin, D., Zheng, J., Tang, X., Fang, Y., and Yu, H. (2019b). Detection of salient crowd motion based on repulsive force network and direction entropy. *Entropy*, 21(6):608. 22

[410] Zhang, X., Zou, J., He, K., and Sun, J. (2015b). Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1943–1955. 30, 39

[411] Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597. 47, 51, 52

[412] Zhang, Z., Tian, Z., and Zhou, M. (2018). Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. *IEEE Sensors Journal*, 18(8):3278–3289. 36

[413] Zhao, R., Xu, W., Su, H., and Ji, Q. (2019). Bayesian hierarchical dynamic model for human action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7733–7742. 22, 60

[414] Zhao, Z., Li, H., Zhao, R., and Wang, X. (2016). Crossing-line crowd counting with two-phase deep neural networks. In *European Conference on Computer Vision*, pages 712–726. Springer. 22

[415] Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., and Oliva, A. (2016). Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*. 84

[416] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464. 84

[417] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495. 84

[418] Zhou, B., Tang, X., and Wang, X. (2013). Measuring crowd collectiveness. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3049–3056. 48, 52

Universidad de Alicante

[419] Zhou, J. T., Du, J., Zhu, H., Peng, X., Liu, Y., and Goh, R. S. M. (2019). Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550. 104, 106

[420] Zhu, Y., Lan, Z., Newsam, S., and Hauptmann, A. (2018). Hidden two-stream convolutional networks for action recognition. In *Asian Conference on Computer Vision*, pages 363–378. Springer. 84

[421] Zhuang, N., Yusufu, T., Ye, J., and Hua, K. A. (2017). Group activity recognition with differential recurrent convolutional neural networks. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 526–531. IEEE. 47, 52

Universitat d'Alacant
Universidad de Alicante