INSTITUTO TECNOLÓGICO DE BUENOS AIRES
MAESTRÍA EN CIENCIA DE DATOS

RECONOCIMIENTO DE ACCIONES HUMANAS EN VIDEO, UTILIZANDO DETECCIÓN DE CONTORNOS Y DEEP LEARNING

Propuesta de Tesis

AUTOR: Ing. Rodrigo Carlos Fondato

DIRECTORA: Dra. María Juliana Gambini

PROPUESTA DE TESIS PARA OPTAR AL TÍTULO DE
MAGISTER EN CIENCIA DE DATOS

CIUDAD AUTÓNOMA DE BUENOS AIRES
Noviembre 2023

Palabras Clave

Reconocimiento de Acciones Humanas, Aprendizaje Automático, Redes profundas, Detección de Contornos, Segmentación de imágenes.

Contenido

Introducción

El reconocimiento de actividades humanas (HAR), consiste en identificar y detectar acciones simples y complejas realizadas por personas, en situaciones de la vida real, utilizando datos obtenidos por sensores (Singh et al., 2017).

Existen numerosas aplicaciones para éstas técnicas, entre las cuales se destacan videovigilancia (Ryoo, 2011), medicina (González-Ortega et al., 2014) o interfaz de aplicaciones, por ejemplo, para controlar un videojuego (Gerling et al., 2012).

Los modelos para detección de comportamiento humano difieren y presentan distintos desafíos, según la naturaleza de los datos de entrada que utilicen. Existen técnicas basadas en visión por computadora, cuya fuente de datos pueden ser videos capturados por cámaras RGB (videos convencionales) o por cámaras RGB-D, como los dispositivos Kinect, que capturan también la profundidad de cada punto, permitiendo el reconocimiento de acciones en un espacio 3D; pueden utilizar datos provenientes de sensores portátiles (por ejemplo acelerómetros de los dispositivos móviles) o ser multimodales, es decir, combinar múltiples fuentes de datos (Yadav et al., 2021).

Este trabajo analizará únicamente los modelos unimodales que usan segmentos de video convencionales, ya que las cámaras RGB son ampliamente utilizadas, lo cual implica una mayor oportunidad de adopción. Los modelos clásicos más comunes emplean técnicas de procesamiento de imágenes para obtener descriptores espacio-temporales como campos de flujo óptico (Efros et al., 2003; Chaudhry et al., 2009), vectores de movimiento (Zhang et al., 2018), histogramas de gradientes orientados (HOG) (Dalal and Triggs, 2005), puntos de interés espacio-temporales (STIPs) (Yan and Luo, 2012), entre otros, que sirven de entrada para un modelo de clasificación. Otros plantean al comportamiento humano como una secuencia de estados predecibles por modelos estocásticos (Robertson and Reid, 2006). Otros utilizan descriptores de bajo nivel para identificar acciones atómicas y luego combinarlas para reconocer comportamientos más complejos (jerárquicos) (Song et al., 2013). Existen también modelos basados en la evaluación de reglas lógicas preestablecidas (Morariu and Davis, 2011).

Los modelos anteriormente descritos suelen presentar distintas limitaciones. Algunas de las mencionadas por (Vrigkas et al., 2015) son las siguientes:

- **Técnicas basadas en descriptores espacio-temporales**: Sensibles al ruido y oclusiones, poca repetibilidad debido a características dispersas, reconocer actividades complejas es difícil, distancia entre características de bajo nivel y eventos de alto nivel.
- **Modelos estocásticos**: Se necesita un gran número de datos de entrenamiento, propenso al sobreajuste, aprender inferencias es complejo.
- **Modelos basados en reglas lógicas**: La generación de reglas es compleja, solo se

reconocen acciones atómicas, problemas con video secuencias largas.

Actualmente se han propuesto numerosas soluciones basadas en deep learning, debido a la excelente capacidad de generalización que poseen. Sin embargo, son modelos que pueden escalar fácilmente en complejidad y cantidad de parámetros, requiriendo un elevado poder de cómputo. Además pueden producir sobreajuste, y por ello es necesaria una gran cantidad de datos disponibles para el entrenamiento (Beddiar et al., 2020). También existe un compromiso entre la resolución espacial y la temporal, es decir, a mayor resolución espacial, menor es la cantidad de cuadros que podrá ser analizada para determinar la actividad (Asadi-Aghbolaghi et al., 2017).

En este trabajo se analizará la combinación de técnicas clásicas de detección de contornos y posterior utilización de un modelo de deep learning, bajo la hipótesis de que el uso de las siluetas humanas como única entrada de datos podrá reducir significativamente la cantidad de parámetros necesarios para la representación espacial, permitiendo ampliar el horizonte temporal y disminuir los requerimientos para el entrenamiento.

En las siguientes secciones se plantea el problema a resolver y las preguntas de investigación que conducen el proceso de revisión de la literatura; se especifican los objetivos; se presentan las técnicas clásicas y actuales utilizadas, limitaciones y desafíos; se plantea la hipótesis, se detalla la metodología adoptada, y finalmente se incluye un cronograma presentando las actividades a realizar y estimaciones de fechas para cada una.

Planteamiento del Problema

Las técnicas para la detección automática de acciones realizadas por personas en capturas de videos pueden dividirse, a grandes rasgos, en aquellas que utilizan herramientas de procesamiento de imágenes para obtener descriptores espacio-temporales y el posterior uso de modelos clásicos de machine learning para la clasificación, y en aquellas basadas en la explotación de grandes volúmenes de datos y la consecuente utilización de modelos de deep learning para etiquetar las acciones detectadas.

Estos últimos modelos, más investigados en la actualidad, permiten obtener descriptores más robustos que los primeros, pero requieren una gran cantidad de recursos computacionales (poder de procesamiento y memoria), tiempo y disponibilidad de datos para su entrenamiento, debido al tamaño de los archivos de video en memoria y la cantidad de parámetros que deben ser optimizados. Existe además, un compromiso entre el horizonte temporal y la resolución espacial. Es decir, a mayor resolución de la imagen de entrada, menor cantidad de cuadros podrán ser analizados por el modelo para efectuar la clasificación, y viceversa.

El problema que se intenta resolver es, para una determinada resolución de video de entrada, la expansión del horizonte temporal y la disminución de la cantidad de parámetros de un modelo basado en deep learning, con la consecuente reducción de la cantidad de datos y recursos computacionales necesarios para el entrenamiento, obteniendo a la vez las ventajas de robustez y rendimiento que otorgan este tipo de modelos, específicamente para el ámbito de reconocimiento de acciones humanas básicas.

2.1. Formulación del problema: Preguntas de Investigación

1. ¿ Qué bases de datos existen que contengan videos clasificados o no clasificados, muestren a una o más personas realizando distintas acciones y sean de acceso público?

    a) ¿Qué formato y calidad presentan los videos? ¿Son todos los segmentos de la misma duración, tamaño, formato y calidad?

    b) En caso de estar clasificados, ¿De qué manera están etiquetados?

2. ¿Qué modelos, basados en visión por computadora, son utilizados actualmente para reconocer acciones humanas en capturas de video?

    a) ¿Cuáles de ellos se basan en la utilización de machine y/o deep learning?

3. ¿Qué técnicas clásicas de procesamiento de imágenes se utilizan para construir descriptores útiles para la detección de actividades humanas y cuáles son las ventajas y limitaciones de cada uno?

    a) ¿Qué modelos existen que combinen éstas técnicas con el uso de deep learning?

4. ¿Qué modelos se utilizan actualmente para detectar contornos o siluetas, y que ventajas y limitaciones poseen?

   a) ¿Cuáles de éstos métodos se usan para el reconocimiento de actividades humanas?

5. ¿Qué técnicas combinan la detección de siluetas o contornos humanos y posterior uso de modelos de deep learning para clasificar actividades humanas en video?

2.2. Objetivo de la investigación

Combinar técnicas de detección de contornos y modelos basados en machine o deep learning para, a partir de capturas de video y en forma automática, clasificar acciones realizadas por personas, etiquetando cada cuadro con la actividad detectada, evaluando el rendimiento, consumo de recursos y tiempos de inferencia, y compararlos con modelos de deep learning que sean utilizados actualmente para este fin.

2.2.1. Objetivos Específicos

- Identificar, clasificar, adaptar y unificar bases de datos de acceso público que contengan registros de video donde se observen diferentes actividades realizadas por personas.
- Implementar un algoritmo de preprocesamiento que, dadas las capturas de video, permita construir características derivadas de la aplicación de técnicas de detección de contornos y/o segmentación de personas.
- Adaptar o combinar modelos que, dadas las características construidas como resultados del objetivo anterior, permitan clasificar las actividades realizadas por personas, en un segmento de video.
- Evaluar el desempeño de la herramienta desarrollada o adaptada sobre uno o más sets de datos de testeo, mediante la utilización de métricas para modelos de clasificación, y compararla con el rendimiento de técnicas existentes de deep learning que utilicen directamente las capturas de video como entrada de datos.

2.2.2. Justificación del estudio

Las técnicas de detección de acciones humanas en capturas de video que utilizan modelos de deep learning entrenados y aplicados cuadro por cuadro requieren el ajuste de una gran cantidad de parámetros para obtener un rendimiento útil, y por consiguiente necesitan un gran poder de cómputo y una elevada cantidad de datos de entrenamiento para reducir el sobreajuste. La utilización de métodos de detección de contornos y/o segmentación de personas para la construcción de características que puedan ser utilizadas como entrada a un modelo de machine o deep learning, podría reducir este costo.

El presente trabajo pretende estudiar el desempeño, uso de recursos y tiempos de inferencia de la técnica propuesta y evaluar ventajas y desventajas al compararlo con la utilización

de modelos basados en deep learning empleados actualmente para el mismo fin, otorgando como resultado, métricas concretas que permitan comparar objetivamente ambos enfoques.

El modelo resultante permitirá reconocer acciones humanas básicas en video y clasificarlas, en forma automática, utilizando únicamente videos RGB como entrada de datos, lo cual lo hace apto para aplicaciones de video vigilancia, médicas, entre otras.

Estado de la Cuestión

El reconocimiento de actividades humanas (HAR, por sus siglas en inglés), es la habilidad para interpretar movimientos y gestos efectuados por personas a través de sensores, para luego determinar qué acciones o actividades están realizando (Ann and Theng, 2014).

La detección y clasificación de estas actividades no es sencilla, ya que factores como un fondo de imagen complejo, la oclusión parcial de los cuerpos de las personas, cambios en la iluminación o escala de la imagen, dificultan el reconocimiento de los cuerpos humanos (Vrigkas et al., 2015).

3.1. Aplicaciones

Existen numerosas aplicaciones de ésta técnica. Citando solo algunos ejemplos, en el campo de la videovigilancia Ryoo propone un modelo para predecir actividades peligrosas o delictivas a partir de una secuencia de cuadros de video (Ryoo, 2011).

En el campo de la medicina, (González-Ortega et al., 2014) proponen un sistema para el monitoreo de ejercicios de rehabilitación cognitiva basado en el uso del dispositivo Kinect y procesamiento de datos 3D, realizando el seguimiento de partes del cuerpo de un paciente que permite medir y monitorear la relación entre los movimientos efectuados y los ejercicios indicados.

HAR también es utilizado como interfaz entre humanos y computadoras. Mediante gestos y movimientos se pueden generar instrucciones que permiten, por ejemplo, interactuar con un videojuego. (Gerling et al., 2012) proponen pautas para el desarrollo de juegos controlados a través del reconocimiento de movimientos en la totalidad del cuerpo para combatir el sedentarismo en adultos mayores.

3.2. Modelos para el reconocimiento de actividades humanas

(Vrigkas et al., 2015) clasifican a los modelos principalmente en unimodales y multimodales. Los unimodales utilizan datos relacionados a un único aspecto y obtenidos desde una única fuente, como por ejemplo el movimiento a través de una imagen, mientras que los multimodales emplean distintos tipos de características obtenidas por la recolección de múltiples fuentes de datos, por ejemplo imagen y sensores portátiles, que son luego fusionados para efectuar la clasificación.

3.2.1. Según la naturaleza de datos utilizados

Las técnicas empleadas varían según los sensores que son utilizados para recolectar los datos.

A grandes rasgos, podemos dividir a las fuentes de datos en:

- Cámaras RGB: Son las cámaras convencionales. Los modelos que utilizan solo este tipo de datos se enmarcan dentro de las técnicas de visión por computadora.
- Cámaras RGB-D: Son aquellas que, además de los canales RGB, incluyen un cuarto canal (D) que posee información sobre la profundidad de la imagen, es decir, reconocen las tres dimensiones espaciales. Un ejemplo es el dispositivo Kinect.
- Sensores portátiles: Corresponden a sensores presentes en dispositivos portátiles que las personas usan, y otorgan información sobre los movimientos que realizan. Generalmente son acelerómetros.
- Multimodales: Combinaciones de los anteriores. Cámaras y sensores, múltiples dispositivos portátiles, etc. (Yadav et al., 2021)

Este trabajo está enfocado en el primer grupo, es decir, técnicas de visión por computadora que usan cámaras RGB.

3.2.2. Técnicas basadas en el uso de características espacio-temporales

Las técnicas unimodales más comunes utilizan una secuencia de imágenes 2D en el tiempo, es decir un cubo 3D como única fuente de datos, y son los denominados por (Vrigkas et al., 2015) como métodos de "espacio-tiempo". Consisten generalmente en la construcción de características obtenidas de las imágenes como puntos de interés, vectores de movimientos, campos de flujo óptico, para luego ser clasificados por un método clásico de machine learning, generalmente SVM (Wang and Mori, 2010; Schuldt et al., 2004; Jhuang et al., 2007; Dalal and Triggs, 2005), o el vecino más cercano (Efros et al., 2003; Gorelick et al., 2007; Vrigkas et al., 2014).

Los métodos que emplean campos de flujo óptico permiten representar aproximadamente los movimientos de cada punto de una superficie 3D en un plano 2D, dada una secuencia temporal de imágenes (Beauchemin and Barron, 1995). Por ejemplo (Efros et al., 2003) desarrollaron descriptores basados en vectores de flujo óptico, y utilizando el algoritmo de vecino más cercano, una base de datos de secuencias de videos previamente etiquetados, y una medida de similitud establecida, pueden clasificar acciones realizadas por personas a gran distancia, es decir, cuyas figuras son representadas por una pequeña cantidad de píxeles, comparando nuevas secuencias con videos previamente etiquetados.

(Dollár et al., 2005) utilizan también descriptores basados en el flujo óptico. Con ello extraen puntos de interés y luego cuboides (es decir pixeles próximos en ambas dimensiones espaciales y la temporal). Cada cuboide es clasificado en tipos mediante el uso de k-means y

luego se construye un histograma con los tipos reconocidos para determinar la clasificación. Esté método no tiene en cuenta información de ubicación espacial ni temporal de cada cuboide.

(Yan and Luo, 2012), en cambio, tienen en cuenta la ubicación espacial y temporal extrayendo puntos de interés espacio-temporales (STIP) por regiones y realizando histogramas que consideran éstas ubicaciones.

3.2.3. Técnicas estocásticas y jerárquicas

Los métodos estocásticos consideran a los comportamientos humanos como una secuencia de estados estocásticamente predecibles utilizando, por ejemplo, modelos ocultos de Markov. Los modelos jerárquicos, por otro lado, descomponen una actividad en múltiples acciones de bajo nivel, y luego las combinan en capas que reconocen dinámicas a más alto nivel. Múltiples ejemplos muestran que ambas técnicas suelen estar asociadas, obteniendo descriptores de bajo nivel para cada cuadro y utilizando modelos estocásticos para combinarlos.

(Robertson and Reid, 2006) consideran que un determinado comportamiento humano depende de una secuencia estocástica de acciones. Combinan métodos paramétricos (Modelos ocultos de Markov) y no paramétricos (basados en similitudes con bases de datos de videos clasificados). Los métodos no paramétricos sirven para extraer distribuciones para las características de bajo nivel (velocidad, posición, descriptores locales), y los métodos paramétricos permiten extraer comportamientos a más alto nivel, combinando la información de pasos anteriores.

Otro ejemplo de esquema estocástico y jerárquico utiliza diferentes niveles de resolución temporal como entrada de datos para determinar el comportamiento a clasificar. Se usan CRFs (campos aleatorios condicionales) con variables latentes para capturar la dinámica de cada capa, y luego se agrupan observaciones vecinas en el tiempo en super observaciones en forma recursiva, calculando probabilidades condicionales en cada nivel con respecto a los anteriores (Song et al., 2013).

3.2.4. Técnicas basadas en reglas lógicas

Otros enfoques utilizan reglas para modelar un evento, descomponiendo una actividad en un conjunto de atributos o reglas primitivas. (Morariu and Davis, 2011), por ejemplo, detectan eventos complejos en el ámbito del basketball, incluyendo reglas que permiten desambiguar, entre otras cosas, que una acción es ofensiva o defensiva. Se emplean redes lógicas de Markov para realizar inferencias sobre las características de bajo nivel extraídas del video y predecir eventos de alto nivel, como un disparo al aro.

3.2.5. Técnicas basadas en detección de figuras

Los métodos basados en detección de figuras buscan reconocer las diferentes partes del cuerpo de las personas y detectar las poses, para luego utilizar ésta información para determinar la actividad que se está realizando.

(Yang et al., 2010) proponen un método para reconocer actividades humanas sobre imágenes estáticas. Extraen las poses humanas como variables latentes del modelo que luego son utilizadas para predecir la acción, es decir, la detección de poses y la clasificación se entrenan en forma conjunta. Para la detección de partes humanas usan Poselets, que describen partes del cuerpo considerando la perspectiva (Bourdev and Malik, 2009).

(Lillo et al., 2014) combinan un enfoque de detección de poses con una estructura jerárquica. Establecen tres niveles: en un primer nivel se detectan las poses a partir de características de bajo nivel, en un segundo nivel se combinan estas poses en acciones y finalmente un tercer nivel combina acciones en comportamientos más complejos. Una ventaja detallada es el manejo de oclusiones, ya que el modelo propuesto otorga mayor peso a las partes visibles.

3.2.6. Técnicas basadas en el uso de deep learning

Las técnicas de reconocimiento de actividades humanas basadas en deep learning son recientemente muy utilizadas debido a su buen rendimiento, producto de una robusta extracción de características y capacidad de generalización, pero requieren de una gran capacidad de cómputo y una enorme cantidad de datos para su entrenamiento (Beddiar et al., 2020).

(Beddiar et al., 2020) clasifican a éstas técnicas en generativas, discriminativas e híbridas. Las generativas son aquellas que utilizan aprendizaje no supervisado para representar distribuciones de datos no etiquetados con menor dimensionalidad, buscando replicar la distribución verdadera del set de datos. Las discriminativas comprenden a modelos supervisados (entrenados sobre set de datos etiquetados) que pueden predecir la clase de nuevos datos de entrada. Generalmente se basan en el uso de redes con múltiples capas que toman los datos de entrada y otorgan como resultado una categoría. Los híbridos utilizan una combinación de los dos enfoques anteriores.

(Asadi-Aghbolaghi et al., 2017), por otro lado, consideran que el desafío más grande es cómo lidiar con la dimensión temporal, y basado en ello, categorizan éstas técnicas en:

- El uso de filtros 3D en redes convolucionales: A las dos dimensiones espaciales de una imagen 2D se le suma la dimensión temporal. Las convoluciones son entonces calculadas tomando en cuenta la vecindad espacial y temporal.

- El uso de características de movimiento (como flujos ópticos 2D) que luego sirven de entrada para el modelo de deep learning.

- El uso de redes convolucionales para extraer características a nivel espacial combinado con un modelo de secuencias temporales como redes recurrentes, LSTMs y otros similares para modelar la dimensión temporal.

Como un ejemplo del primer caso, (Ji et al., 2013) proponen el uso de una red convolucional 3D directamente sobre los cuadros de video. Para resolver la presencia de múltiples personas, se utiliza primero un detector pre-entrenado para obtener cuadros delimitadores, y luego se aplica la red convolucional sobre cada uno, en múltiples imágenes adyacentes en la secuencia temporal, asumiendo que la persona sigue ubicada en la misma posición. Como limitación adicional se menciona el uso de entradas de 80x40 píxeles y 9 cuadros contiguos debido a los requerimientos de memoria por la gran cantidad de parámetros del modelo.

(Li et al., 2019), en cambio, proponen el uso de redes convolucionales 3D pero incluyendo bloques densos intermedios con una arquitectura de "cuello de botella" para disminuir la cantidad de parámetros.

Otro ejemplo es el propuesto por (Tran et al., 2015), que menciona la creación de características lo suficientemente genéricas para ser utilizadas para diferentes propósitos, construidas a partir de una red convolucional 3D. La red otorga como salidas descriptores de tan solo 10 dimensiones. Sin embargo cada capa convolucional utiliza kernels de 3x3x3, lo que implica el uso de secuencias temporales cortas para la clasificación.

Para poder procesar clips de video más largos, una de las soluciones propuestas emplea redes convolucionales 2D sobre cada cuadro y luego capas de pooling para unir los descriptores de cuadros adyacentes, o redes secuenciales como LSTM para modelar la relación temporal, basándose en que los parámetros son compartidos en todos los pasos temporales, lo cual reduce su cantidad (Ng et al., 2015). Este modelo corresponde al tercer tipo descrito por (Asadi-Aghbolaghi et al., 2017)

Otra solución para reducir los tiempos de entrenamiento y demanda computacional propone inicializar los pesos de la red convolucional 3D usando los pesos pre-entrenados de la red ImageNet 2D (Mansimov et al., 2015). El desafío en ésta propuesta es el de utilizar parámetros pre-entrenados con imágenes estáticas en un modelo que incluye la dimensión temporal.

3.2.7. Técnicas que combinan descriptores espacio-temporales y deep learning

(Simonyan and Zisserman, 2014) proponen usar dos redes convolucionales separadas que son combinadas en la salida. Una representa la relación espacial y la otra la temporal. Ésta última utiliza flujos ópticos entre los cuadros adyacentes en el tiempo para representar el movimiento de los píxeles como entrada de datos.

Continuando ésta línea, (Zhang et al., 2018) argumentan que el método descrito anteriormente no es apto para procesamiento en tiempo real debido a la demanda computacional del cálculo de campos de flujo óptico entre cada par de cuadros en la secuencia, y proponen utilizar vectores de movimiento extraídos directamente desde video comprimido.

(Peng and Schmid, 2015) proponen combinar características de trayectoria como histogramas de gradientes orientados (Dalal and Triggs, 2005), histogramas de flujos ópticos (Chaudhry et al., 2009), e histogramas basados en movimiento (Wang and Schmid, 2013) con una red convolucional (VGG19), concatenando los resultados y aplicando un SVM para la clasificación.

3.3. Oportunidades y desafíos actuales

Las cámaras RGB son dispositivos muy comunes, por este motivo los métodos basados en el uso de videos RGB como entrada de datos presentan una gran oportunidad de adopción. Por ejemplo, en el mundo se utiliza una enorme cantidad de cámaras RGB destinadas a video-vigilancia.

En contraposición, las cámaras RGB-D suelen usarse solo en determinados ámbitos, y las técnicas que usan dispositivos portátiles requieren de múltiples sensores ubicados en las diferentes partes del cuerpo para otorgar un rendimiento adecuado, lo cual puede ser muy intrusivo para las personas (Yadav et al., 2021).

Como se describió anteriormente, los modelos alimentados por secuencias de video RGB que utilizan descriptores obtenidos por la aplicación de técnicas clásicas como campos de flujo óptico, puntos de interés espacio-temporales o histogramas de gradientes orientados, tienen ciertas limitaciones para generalizar ya que dependen de aspectos como la iluminación, perspectiva y presentan desafíos cuando hay oclusiones. En algunos casos pueden ser computacionalmente intensivos para la inferencia (Zhang et al., 2018).

En el caso de los métodos basados en el uso de técnicas de deep learning, la gran cantidad de parámetros a entrenar puede resultar en un entrenamiento computacionalmente intensivo, y su rendimiento a la hora de clasificar dependerá de la cantidad de datos que se hayan utilizado, ya que un número de registros que no es lo suficientemente elevado, puede producir el sobreajuste de los parámetros.

Otro inconveniente con las técnicas de deep learning es el horizonte temporal que pueden procesar, debido al incremento de parámetros que conlleva expandirlo, con lo cual la clasificación de acciones suele producirse en un segmento breve de tiempo. Para reducir el número de parámetros e incrementar el horizonte temporal, existe un compromiso en el cual se propone reducir la resolución espacial (Asadi-Aghbolaghi et al., 2017).

Hipótesis

La construcción de descriptores resultantes de la aplicación de técnicas de detección de contornos y/o segmentación de personas sobre cada cuadro de video y posterior uso de modelos de machine o deep learning, permite reconocer y clasificar actividades básicas realizadas por personas en video con un rendimiento similar al obtenido mediante el uso de modelos de deep learning que son entrenados utilizando directamente los cuadros de video como entrada de datos, pero requiriendo menor tiempo, recursos de cómputo y datos para el entrenamiento.

4.1. Variables independientes

Variable	Descripción	Indicadores	Métricas
Cuadros de video	Cada imagen estática RGB presente en un segmento de video	Intensidad de gris por canal, por píxel en cada cuadro de video	Valores enteros entre 0 y 255.

Tabla 4.1: Variables Independientes

4.2. Variables dependientes

Variable	Descripción	Indicadores	Métricas
Rendimiento del modelo	Métricas de rendimiento para modelos de clasificación	Precisión Especificidad Sensibilidad F1-Score	Valores reales entre 0 y 1
Tiempo de cómputo	Tiempo necesario para entrenar el modelo	Cantidad de minutos cuantificados desde el inicio hasta la convergencia del entrenamiento del modelo	Valor real positivo
Tiempo de inferencia	Tiempo necesario para inferir un resultado a partir de un dato de entrada	Cantidad de segundos por cuadro de video a inferir	Valor real positivo
Recursos de cómputo	Cantidad de espacio en memoria necesario para entrenar el modelo y para la inferencia	Espacio en disco y en memoria RAM utilizados medidos en megabytes (MB)	Valores reales positivos

Tabla 4.2: Variables Dependientes

Metodología

5.1. Tipo de Investigación

Se realizará una investigación aplicada, adaptando e integrando modelos y herramientas existentes para estudiar concretamente la detección automática de acciones humanas básicas en segmentos de video en tiempo real, rendimiento y uso de recursos computacionales de cada técnica.

Se experimentará en un entorno controlado mediante una preselección, adaptación e integración de bases de datos, y se utilizarán métricas concretas para medir los aspectos mencionados, por lo cual será una investigación experimental y cuantitativa, obteniendo como resultado una descripción de las consecuencias observadas en base a los modelos utilizados.

5.2. Diseño de la investigación

A continuación se describe la organización de tareas, disponibilidad de datos y recursos, y técnicas y herramientas a utilizar para llevar a cabo la investigación propuesta.

5.2.1. Planificación de actividades

1. Como primer paso se realizará una revisión de la literatura, principalmente de artículos publicados en revistas científicas y conferencias. El objetivo es identificar las diferentes técnicas, modelos y herramientas utilizados actualmente para el reconocimiento de actividades humanas, y realizar un análisis del uso y rendimiento de técnicas de detección de contornos, en segmentos de video.

2. Se perfilarán las técnicas y modelos encontrados:

    a **Según el enfoque:** Uso o no de técnicas de deep learning, métodos clásicos, entrenamiento supervisado o no supervisado.

    b **Según el tratamiento de los datos de entrada:** La realización o no de un preprocesamiento para la obtención de características previo al entrenamiento del modelo y el tipo de características extraídas.

    c **Según los set de datos usados para la experimentación:** Datos etiquetados o no, generados artificialmente o no, tipos de actividades humanas estudiadas en cada base de datos, etc.

    d **Según las métricas obtenidas:** Cuáles se utilizaron, como fueron medidas.

3. Se explorarán y analizarán bases de datos de acceso público que contengan segmentos de videos que muestren personas realizando diversas actividades y hayan sido mencionadas y empleadas en publicaciones de calidad para la medición del rendimiento de modelos de detección.

4. Se construirá un único set de datos, que se utilizará para la experimentación. Este set de datos deberá ser lo más heterogéneo posible respecto al contenido (distintas áreas geográficas, distintos ángulos de visión, distinta iluminación, distinta composición de personas y objetos, etc.) y lo más homogéneo posible respecto a formato, calidad y duración de los segmentos de video.

5. Se seleccionará el modelo de detección de contornos y/o segmentación de personas a utilizar y se implementará el algoritmo que permitirá preprocesar las imágenes aplicando esta técnica sobre cada cuadro de cada video.

6. Se diseñará y entrenará el modelo que tomará como entrada los contornos mencionados en el punto anterior y clasificará los segmentos de video con la actividad detectada.

7. Se probará el modelo con un subconjunto de datos de testeo.

8. Se obtendrán métricas de rendimiento y uso de recursos respecto al entrenamiento y la inferencia.

9. Se seleccionarán y analizarán los modelos existentes que serán comparados con el modelo propuesto, y se obtendrán e instalarán las implementaciones.

10. Se ejecutarán los modelos del punto anterior y se compararán métricas con las obtenidas por el modelo en el paso 8.

11. Se propondrán cambios en el diseño de las características y del modelo y repetirán los pasos 5 a 9, para intentar conseguir mejores resultados.

5.2.2. Disponibilidad de los datos

Existen numerosas bases de datos de acceso público que poseen segmentos de video, mostrando diferentes actividades realizadas por personas.

A grandes rasgos pueden ser segmentadas en:

- **No clasificadas:** Archivos de video que contienen diferentes actividades humanas no etiquetadas.
- **Clasificadas a nivel de segmento de video:** Cada archivo de video está etiquetado.
- **Clasificadas a nivel cuadro:** Cada cuadro indica la actividad presente.
- **Clasificadas a nivel píxel:** Existe una etiqueta en cada píxel de cada cuadro, lo que permite ubicar la posición de las personas en la imagen.

5.2.3. Disponibilidad recursos tecnológicos

Los experimentos serán conducidos en un entorno cloud, con disponibilidad de uso de GPUs.

5.2.4. Personas involucradas

No es necesaria la presencia de un experto. La investigación será realizada únicamente por el autor.

5.2.5. Procesos de gestión de proyecto y datos

Para la gestión de los datos se utilizará la metodología CRISP-DM, siguiendo un enfoque iterativo de entendimiento del negocio y los datos, modelado y evaluación. De ésta manera se obtendrán periódicamente medidas de desempeño que guiarán correcciones en la toma de decisiones para la elección de las características a utilizar y modelado, con el propósito de incrementar la precisión de clasificación en cada iteración.

Para la gestión de las tareas se adoptarán algunos aspectos de la metodología SCRUM, aplicadas a un solo individuo: Dividir el proyecto en iteraciones, construir una lista de tareas y escoger aquellas más prioritarias en cada iteración, obteniendo un entregable al finalizar cada ciclo.

5.2.6. Población y Muestra

Se usarán segmentos de video que muestran actividades básicas realizadas por humanos. No se pretende limitar la ubicación geográfica ni temporal. Los segmentos serán clasificados en múltiples categorías, según las actividades a reconocer.

5.2.7. Técnicas y herramientas

Para la construcción de características:

- Detección de contornos, segmentación de imágenes.

Para la construcción del modelo:

- Modelos basados en el uso de machine o deep learning.

Para la gestión de datos no estructurados:

- Data Lake.

El lenguaje Python será la base para el desarrollo del modelo y evaluación de su desempeño. Se utilizará un data lake para almacenar y procesar los archivos de entrada (videos etiquetados), dado que son datos no estructurados de gran tamaño.

- **Gestión del data lake:** Herramientas similares a PySpark.
- **Procesamiento de datos:** Librerías de python de acceso público y uso frecuente para el procesamiento de datos tales como PySpark, Pandas, Numpy.
- **Preprocesamiento de imágenes:** Librerías de python y C++ de acceso público y uso frecuente para el procesamiento de imágenes tales como OpenCV.

- **Construcción de modelos de machine y deep learning:** Librerías de python para el modelado, entrenamiento y ejecución de herramientas basadas en machine o deep learning tales como Tensorflow, PyTorch, Keras, Scikit-Learn.

Se usará un software de código abierto y acceso público para el registro y administración de las tareas del proyecto, como por ejemplo, OpenProject.

Cronograma

Etapa	Actividad	Duración (semanas)	Fecha Inicio	Fecha Fin
Inicio del proyecto	Definición del proceso de gestión del proyecto.	2	13/11/23	27/11/23
	Selección de herramientas a utilizar para la gestión, instalación y puesta a punto.			
	Creación de una lista de tareas iniciales.			
Investigación de las técnicas y modelos utilizados	Definición del proceso de revisión de la literatura y fuentes a investigar.	1	27/11/23	04/12/23
	Armado de consultas en base a objetivos, y criterios de selección de artículos.	2	04/12/23	18/12/23
	Ejecución de de cada consulta en cada fuente y selección preliminar de artículos (lectura diagonal)	3	18/12/23	08/01/24
	Lectura en profundidad de los artículos y selección final	3	08/01/24	29/01/24
	Armado del registro con las características, métricas y bases de datos utilizados en cada artículo	2	29/01/24	12/02/24
	Identificación, clasificación, selección y obtención de bases de datos públicas.	1	12/02/24	19/02/24
	Análisis preliminar y preselección de los segmentos de video según categorías, calidad, duración y formato.	2	19/02/24	04/03/24

Obtención y armado de la base de datos	Análisis individual de cada archivo de video y selección final.	2	04/03/24	18/03/24
	Definición del proceso y herramientas necesarias para adaptar y homogeneizar los distintos archivos.	1	18/03/24	25/03/24
	Implementación y ejecución del proceso, obteniendo como resultado una única base de datos.	3	25/03/24	15/04/24
Experimentación	Selección e implementación del algoritmo de contornos activos.	4	15/04/24	13/05/24
	Diseño e implementación del modelo de reconocimiento de actividades humanas.	4	13/05/24	10/06/24
	Diseño de las métricas e implementación de la herramienta que las obtendrá.	2	10/06/24	24/06/24
	Entrenamiento y optimización del modelo.	2	24/06/24	08/07/24
	Revisión de diseño de características y el modelo, propuesta e implementación de mejoras. Obtención de nuevas métricas y elección del modelo ganador.	4	08/07/24	05/08/24
	Instalación y ejecución de modelos pre-entrenados existentes. Obtención de métricas.	3	05/08/24	26/08/24
Resultados	Análisis detallado de los resultados obtenidos y la comparación de modelos.	2	26/08/24	09/09/24
	Elaboración de conclusiones y finalización de informe.	1	09/09/24	16/09/24
Total		**44**	**13/11/23**	**16/09/24**

Tabla 6.1: Cronograma

Bibliografía

Ann, O. C. and Theng, L. B. (2014). Human activity recognition: A review. In *Proceedings - 4th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2014*, pages 389–393.

Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-Lopez, V., Baro, X., Guyon, I., Kasaei, S., and Escalera, S. (2017). A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. In *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pages 476–483.

Beauchemin, S. S. and Barron, J. L. (1995). The Computation of Optical Flow. *ACM Computing Surveys (CSUR)*, 27(3):433–466.

Beddiar, D. R., Nini, B., Sabokrou, M., and Hadid, A. (2020). Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41-42):30509–30555.

Bourdev, L. and Malik, J. (2009). Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1372.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pages 1932–1939.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 886–893.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings - 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS*, volume 2005, pages 65–72.

Efros, Berg, Mori, and Malik (2003). Recognizing action at a distance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 726–733. IEEE.

Gerling, K. M., Livingston, I. J., Nacke, L. E., and Mandryk, R. L. (2012). Full-body motion-based game interaction for older adults. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1873–1882.

González-Ortega, D., Díaz-Pernas, F. J., Martínez-Zarzuela, M., and Antón-Rodríguez, M. (2014). A Kinect-based system for cognitive rehabilitation exercises monitoring. *Computer Methods and Programs in Biomedicine*, 113(2):620–631.

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.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. Ieee.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Li, J., Jiang, X., Sun, T., and Xu, K. (2019). Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2019*. Institute of Electrical and Electronics Engineers Inc.

Lillo, I., Soto, A., and Niebles, J. C. (2014). Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 812–819.

Mansimov, E., Srivastava, N., and Salakhutdinov, R. (2015). Initialization Strategies of Spatio-Temporal Convolutional Neural Networks.

Morariu, V. I. and Davis, L. S. (2011). Multi-agent event recognition in structured scenarios. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3289–3296.

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 4694–4702.

Peng, X. and Schmid, C. (2015). Encoding feature maps of cnns for action recognition.

Robertson, N. and Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104(2-3 SPEC. ISS.):232–248.

Ryoo, M. S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1036–1043.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 1, pages 568–576.

Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., and Holzinger, A. (2017). Human activity recognition using recurrent neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10410 LNCS, pages 267–274.

Song, Y., Morency, L. P., and Davis, R. (2013). Action recognition by hierarchical sequence summarization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3562–3569.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 4489–4497.

Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, I. A. (2014). Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, 119:27–40.

Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558. Institute of Electrical and Electronics Engineers Inc.

Wang, Y. and Mori, G. (2010). Hidden part models for human action recognition: Probabilistic versus max margin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1310–1323.

Yadav, S. K., Tiwari, K., Pandey, H. M., and Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223.

Yan, X. and Luo, Y. (2012). Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier. *Neurocomputing*, 87:51–61.

Yang, W., Wang, Y., and Mori, G. (2010). Recognizing human actions from still images with latent poses. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2030–2037.

Zhang, B., Wang, L., Wang, Z., Qiao, Y., and Wang, H. (2018). Real-Time Action Recognition with Deeply Transferred Motion Vector CNNs. *IEEE Transactions on Image Processing*, 27(5):2326–2339.