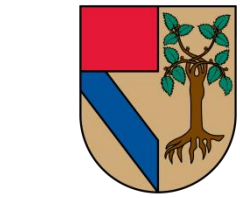


UNIVERSIDAD
PANAMERICANA®
Aguascalientes

FACULTAD DE INGENIERÍA

Un enfoque basado en la visión para la detección de caídas utilizando múltiples cámaras y Redes Neuronales Convolucionales: Un caso de estudio en UP-Fall Detection Data-set

TESIS

QUE PRESENTA

Ricardo Abel Espinosa Loera

PARA OBTENER EL GRADO DE

MAESTRÍA EN CIENCIAS

CON RECONOCIMIENTO DE VALIDEZ OFICIAL DE ESTUDIOS DE LA SECRETARÍA DE EDUCACIÓN PÚBLICA, DE ACUERDO CON EL Nº 2007574 DE FECHA 29 DE JUNIO DE 2007

DIRECTORES DE TESIS

Dr. Hiram Eredin Ponce Espinosa

Dr. José Sebastián Gutiérrez Calderón

Aguascalientes, Ags. Diciembre 2019

Contenido

1.Resumen

Actualmente, el reconocimiento automático de caídas humanas es un tema de investigación importante para la visión por computadora y la comunidad de la inteligencia artificial. Para el análisis de imágenes, es común usar un enfoque basado en visión para la detección de caídas y sistemas de clasificación debido al aumento exponencial actual en el uso de cámaras. Además, las técnicas de deep learning han revolucionado las técnicas basadas en visión. Han sido consideradas robustas y confiables en la detección y clasificación de problemas, principalmente usando Redes Neuronales Convolucionales (CNN). Recientemente, nuestro grupo de investigación lanzo un nuevo Data Set multimodal para la detección de caídas (Up-Fall Detecction dataset), y se requieren diferentes estudios de enfoques de modalidades para la detección y clasificación de caídas. Centrándonos solo en un enfoque basado en visión, en este articulo presentamos un sistema de detección de caídas basado en 2D CNN como método de inferencia y varias cámaras. Este enfoque analiza imágenes en marcos de ventana de tiempo fijo que extraen características utilizando un método de flujo óptico que obtiene información de movimiento relativo entre dos imágenes consecutivas. Para resultados experimentales, probamos este enfoque en nuestro dataset público. Los resultados mostraron que nuestra propuesta de enfoque basado en la visión múltiple detecta caídas humanas que alcanzan un 95.64% de precisión con una arquitectura de red CNN simple en comparación con otros métodos de vanguardia.

2.Abstract

The automatic recognition of human falls is currently an important topic of research for the computer vision and artificial intelligence communities. In image analysis, it is common to use a vision-based approach for fall detection and classification systems due to the recent exponential increase in the use of cameras. Moreover, deep learning techniques have revolutionized vision-based approaches. These techniques are considered robust and reliable solutions for detection and classification problems, mostly using convolutional neural networks (CNNs). Recently, our research group released a public multimodal dataset for fall detection called the UP-Fall Detection dataset, and studies on modality approaches for fall detection and classification are required. Focusing only on a vision-based approach, in this paper, we present a fall detection system based on a 2D CNN inference method and multiple cameras. This approach analyzes images in fixed time windows and extracts features using an optical flow method that obtains information on the relative motion between two consecutive images. We tested this approach on our public dataset, and the results showed that our proposed multi-vision-based approach detects human falls and achieves an accuracy of 95.64% compared to state-of-the-art methods with a simple CNN network architecture.

3.Introducción

El reconocimiento de la actividad humana (HAR) en el monitoreo y seguimiento de la salud de las personas es recientemente un tema interesante que ha estado creciendo dentro de la comunidad de investigación, especialmente para detectar caídas humanas entre las personas mayores. Las caídas pueden provocar lesiones, daños corporales, fracturas, etc. De hecho, las caídas son, a nivel mundial, la segunda causa principal de lesiones no intencionales y muertes relacionadas con lesiones entre adultos de 65 años de edad y mayores [1]. Aproximadamente 28-35% de las personas de 65 años o mas caen cada año, aumentando al 32-42% para los mayores de 70 años [2]. Las caídas con frecuencia causan dependencias funcionales en los ancianos. Además, muchas de estas muertes relacionadas son el resultado de una larga puesta como un periodo prolongado de tiempo en el que la víctima permanece inmóvil en el suelo.

Hay muchos tipos de caídas. Oneill et al [4] divide las caídas humanas por dirección: adelante, atrás y hacia un lado. Por ejemplo, las caídas hacia adelante son las mas comunes, con un 38% en hombres menores de 65 y un 62% en hombres mayores a esta edad. De forma similar, en las mujeres, las caídas están presentes con 62% en mujeres menores de 65 años y 60% en mujeres mayores de esta edad.

En 1987, el grupo de trabajo internacional de Kellogg [3] sobre la prevención de caídas en los ancianos definido una caída como caer involuntariamente al suelo o en un nivel mas bajo y que no sea como consecuencia de un golpe violento, perdida de conciencia, aparición repentina de parálisis como en un derrame cerebral o una crisis epiléptica. Una caída humana generalmente comienza con un corto periodo de caída libre. Esto hace que la amplitud de la aceleración caiga significativamente por debajo del umbral de 1G. Esto representa el periodo de tiempo en que tiene lugar la caída real. La caída de detenerse y causa aceleración y pico en el gráfico. La amplitud que luego cruza un umbral superior sugiere una caída [5].

Se ha demostrado que las consecuencias médicas de una caída dependen mucho del tiempo de respuesta y rescate. En este sentido, los sistemas de detección de caídas pueden mejorar el tiempo de respuesta a la atención médica y disminuir las consecuencias médicas de las caídas.

Debido a un avance extraordinario y al aumento de la investigación en sistemas de sensores integrados, dispositivos móviles y microelectrónica, los sistemas de Internet de las cosas (IoT) permiten a las personas interactuar continuamente con esta tecnología. Además, implica el acceso a una gran cantidad de datos sobre sus acciones diarias para poder realizar sistemas de detección de caídas y para permitir una asistencia rápida y adecuada a las personas mayores.

Existen muchos enfoques para los sistemas de detección de caídas, incluidas las estrategias basadas en sensores, visión y multimodales. Los enfoques basados en sensores hacen uso del ambiente, dispositivos inteligentes y sensores portátiles para proporcionar información importante como aceleración, ausencia / presencia, etc. Por otro lado, la estrategia basada en la visión utiliza imágenes, como entrada principal, como: reconstrucciones 3D del entorno, secuencias simples de video 2D RGB con una o varias cámaras, o imágenes de profundidad adquiridas de sensores de profundidad 3D. Los enfoques multimodales recopilan toda la información posible que proporcionan las cámaras, micrófonos, sensores portátiles, sensores ambientales, dispositivos inteligentes, entre otros, y combinan toda esta información para mejorar la detección y clasificación de caídas de una manera más factible.

Los métodos analíticos y de machine learning son dos enfoques principales para detectar actividades y caídas [6]. Los métodos analíticos resuelven la detección de caídas utilizando algoritmos de umbral. Por ejemplo, al caer, frecuentemente, una persona se golpea con el suelo o un obstáculo. Este "choque de impacto" resulta en un cambio intenso de la aceleración en términos de la dirección de la trayectoria. Este cambio de direccionalidad puede detectarse mediante un valor umbral. En este tipo de métodos, la tarea más difícil es adaptar las detecciones a diferentes tipos de caídas o diferentes personas, ya que los umbrales son diferentes por persona y / o por tipo de caída [6]. Para enfrentar este problema, existen otras estrategias como la normalización de patrones [61] o los algoritmos basados en correlación [62]. Actualmente, las investigaciones recientes recomiendan elegir el umbral utilizando algoritmos de optimización [47].

Por otro lado, los métodos de machine learning han ido adquiriendo más popularidad gracias a la flexibilidad de los algoritmos para diferentes sujetos y tipos de caídas [63].

Algunas de las técnicas más conocidas de aprendizaje supervisado usadas para detención de caídas por sus siglas en ingles son: Multi-Layer Perceptron (MLP) [42], Support Vector Machines (SVM) [39], Hidden Markov Models (HMM), decision trees, random forest, k-Nearest Neighbors (KNN) [41], and Convolutional Neural Networks (CNN) [40]; el ultimo como método de deep learning.

Hoy en día, las técnicas de deep Learning han cambiado y mejorado la forma de abordar los problemas de visión por computadora. Con respecto a CNN, su característica principal considera aprender automáticamente las características de los datos de entrenamiento, haciendo posible una extracción automática de características para las imágenes. CNN se ha aplicado ampliamente en múltiples problemas de procesamiento de imágenes como en [48] en el que los autores utilizan deep learning para detectar accidentes utilizando el flujo óptico como método de extracción de características y luego se prueban en videos reales. En [49], se entrenó un CNN utilizando imágenes directamente para clasificar las lesiones de piel y detectar el cáncer con AUC de 0,96. Además, la CNN se ha utilizado en sistemas de detección de caídas con un enfoque basado en sensores con una precisión del 92,3% [50] y con un enfoque basado en dispositivos portátiles con AUC igual a 0,75 [51].

Los enfoques basados en visión (vision-based) para sistemas de detección de caídas han sido abordados de exitosamente con deep learning. Por ejemplo, Núñez-Marcos et al [19] implementó una CNN para evitar la ingeniería manual de características, permitiendo que las capas convolucionales de su sistema extraigan las características más importantes de las imágenes obteniendo una sensibilidad y especificidad del 94%. La CNN para un enfoque basado en la visión también se implementó en [52] en la que los autores usan una CNN 3D con videos de entrada de la cinemática de las personas, con lo que logran una precisión del 100% evaluada en diferentes conjuntos de datos.

Además, debido a la cantidad de información proporcionada por las cámaras, un reconocimiento de caída humana se convierte en una tarea alcanzable. Hoy en día, las cámaras usan sistemas de monitoreo de salud en escuelas, hospitales, hogares de ancianos, etc. Además, se han adoptado porque representan una solución de bajo costo y también son fáciles de instalar [64].

Recientemente, nuestro grupo de investigación lanzó un conjunto de datos multimodales público para la detección de caídas, llamado UP-Fall Detection Dataset [24]. Los datos se obtuvieron de diferentes fuentes de información: sensores portátiles, sensores ambientales y cámaras. Sin embargo, las diferentes técnicas y habilidades requeridas para

construir y configurar un sistema de detección de caídas multimodal hacen que sea difícil implementarlo en el mundo real. Además, los sensores portátiles y ambientales están condicionados por el sujeto y el entorno, lo que dificulta su portabilidad.

Además, los sistemas de detección de caídas que se basan en cámaras RGB individuales a menudo dependen del punto de vista, según [67]. Esto plantea la necesidad de nuevos conjuntos de datos cuando una cámara se mueve a diferentes puntos de vista y, en particular, a diferentes alturas. Para enfrentar este problema, los diferentes puntos de vista de la cámara en una colección de conjuntos de datos pueden ayudar a identificar si un método dado tiene o no una propiedad independiente del punto de vista. Para este fin, un sistema de detección de caídas debe ser confiable, independientemente de la posición del sujeto, mientras cae, con respecto a la cámara.

De lo anterior, este trabajo presenta un sistema de detección de caídas basado en un método de inferencia 2D CNN y múltiples cámaras. Como describimos más adelante, este enfoque analiza imágenes en marcos de ventana de tiempo fijo que extraen características usando un método de flujo óptico que obtiene información de movimiento relativo entre dos imágenes consecutivas de grabaciones de video adquiridas de cámaras en diferentes puntos de vista. Para la experimentación, probamos este enfoque en nuestro UP-Fall Detection data set. Los resultados mostraron que nuestro enfoque basado en la visión múltiple propuesto detecta las caídas humanas utilizando una arquitectura de red CNN simple, logrando un rendimiento competitivo en comparación con otros métodos de vanguardia. Además, es comparable con el rendimiento logrado con un enfoque multimodal.

A pesar de que CNN se ha utilizado en el pasado en sistemas de detección de caídas con un buen rendimiento utilizando un conjunto de datos particular, Casilari et al [53], concluyeron que estos sistemas deberían ser entrenados y probados con diferentes conjuntos de datos debido a la cantidad diferente de muestras, diferentes tipos de caídas o series temporales diferentes que realizan cualquier tipo de caída. En ese sentido, la implementación de CNN en un enfoque basado en la visión de múltiples cámaras, específicamente para el UP-Fall Detection data set, podría aumentar la vanguardia de los sistemas de detección de caídas.

Las principales contribuciones de este trabajo consideran: (i) el uso de múltiples cámaras con CNN para la detección y clasificación de caídas, (ii) la implementación de este enfoque

en el UP-Fall Detection data set, y (iii) el rendimiento competitivo comparable con otros conocidos métodos de aprendizaje supervisado. Hasta donde sabemos, solo hay un trabajo [59] que combina CNN con un enfoque basado en la visión de múltiples cámaras para reconocer las caídas humanas. En contraste con nuestra propuesta, los autores en [59] usan una estrategia de votación de los resultados de salida de cámaras independientes; mientras que los nuestros usan la información de todas las cámaras en el mismo modelo de machine learning.

El resto del articulo está organizado de la siguiente manera. Primero, revisamos y analizamos diferentes enfoques para los sistemas de detección de caídas, centrándonos principalmente en soluciones basadas en la visión. Luego, presentamos una descripción del UP-Fall Detection data set, y después de eso, presentamos la propuesta en detalle. Más adelante, explicamos la experimentación e incluimos los resultados y las discusiones de esta propuesta. Finalmente, las conclusiones.

4. Sistemas de detección de caídas

Los sistemas HAR y de detección de caídas se han convertido en tareas difíciles, y hay varias formas de lograrlos debido a los muchos enfoques diferentes propuestos en la literatura. Por ejemplo, Lara et al [8] y Noury [6] dividen la taxonomía HAR en tres enfoques generales dependiendo de la fuente de información: externa, ponible y detección de video. De ellos, hay estudios de casos relacionados con enfoques basados en sensores [9], basados en visión [10], basados en teléfonos inteligentes [11] y basados en multimodales [12] para abordar el reconocimiento de caídas en humanos, como se describe a continuación.

4.1. Sistemas de detección de caídas basados en sensores

Con el aumento y la accesibilidad a la tecnología de sensores móviles, los sistemas de detección de caídas se han diseñado para fines del mundo real. La actividad humana se puede rastrear, monitorear y etiquetar datos provenientes de diferentes tipos de sensores en varias ubicaciones en el medio ambiente y en el cuerpo humano. Una aplicación importante de un enfoque basado en sensores es detectar actividades anormales de sensores ponibles en áreas determinadas [9].

Luego, se pueden aplicar métodos de detección de actividad anormal para seguir continuamente los movimientos de cada individuo para verificar si las actividades de la persona están fuera de lo normal [9]. En [21] utilizando sensores triaxiales y SVM como método de inferencia, los autores lograron una precisión del 98,33%. O en [65] usando aceleración y ángulo de Euler (guiñada, cabeceo y balanceo) que logran 100% de precisión, sensibilidad y especificidad. Sin embargo, algunas desventajas de las redes heterogéneas de sensores provienen del hecho de que las actividades humanas generalmente involucran más de una parte del cuerpo. Además, varios estudios fisiológicos y biomecánicos han demostrado que la mayoría de las actividades humanas que se realizan día a día son inherentemente multimodales [13]. Por lo tanto, se requieren diferentes tipos de sensores para recopilar datos.

4.2. Sistemas de detección de caídas basados en dispositivos portátiles

Los enfoques basados en dispositivos portátiles son soluciones comunes para la detección de caídas, aprovechando las tecnologías portátiles debido a su bajo costo, capacidad de seguimiento en vivo y tamaños pequeños. Por ejemplo, en [46] se usó un dispositivo Shimmer para recopilación y transmisión de los datos. El dispositivo portátil se colocó en

el pecho obteniendo un 98.8% de precisión utilizando diferentes modelos de machine learning. En [47], los autores utilizaron una banda ponible colocada en la muñeca, logrando una especificidad de 0,95 y una sensibilidad de 0,83 mediante el reconocimiento de picos basado en el umbral con SVM para la clasificación, en el que optimizaron el mejor valor de umbral para diferentes conjuntos de datos.

En dispositivos portátiles y teléfonos inteligentes, el consumo de energía es un problema difícil de abordar, ya que requieren estar continuamente encendidos para rastrear la información de los sujetos. La vida útil de los dispositivos portátiles y los teléfonos inteligentes se limita a la capacidad de la batería, y la recarga constante es necesaria para evitar el seguimiento constante de las actividades del paciente [46].

4.3. Sistemas de detección de caídas basados en teléfonos inteligentes

Hoy en día, los teléfonos inteligentes contienen múltiples sensores integrados y demasiada capacidad de procesamiento que crece con el paso de los años. Los teléfonos inteligentes pueden medir los movimientos del usuario fuera de un controlador de una manera no intrusiva. Los sistemas de detección de caídas basados en teléfonos inteligentes utilizan sensores de teléfonos inteligentes, p. giroscopio, acelerómetro triaxial o altímetro, para lograrlos en un largo período de tiempo. Por ejemplo, los estudios de casos que utilizan este enfoque se pueden encontrar en [43], en el que los autores utilizan un acelerómetro triaxial basado en un teléfono inteligente con características estadísticas en el dominio del tiempo. Then, they applied principal component analysis method for feature selection, and finally they inferred outputs with MLP obtaining 92% of accuracy. Another example is the work of Vilarinho et al [44] that combined smartphone and smartwatch sensors, using threshold-based techniques and pattern recognition algorithms for recognizing falls with 63% of accuracy and daily activities with 78% of accuracy.

4.4. Sistemas multimodales de detección de caídas

La adquisición de datos es una tarea importante en los sistemas de detección y clasificación de caídas, principalmente sobre sensores ambientales, dispositivos portátiles, cámaras, micrófonos, etiquetas RFID, entre muchos otros que se pueden utilizar para la tarea de reconocimiento. El uso de ponibles no es capaz de distinguir una gran cantidad de actividades humanas complejas y / o de grano fino, dificultad similar en los sensores ambientales para el contexto. En este sentido, los enfoques multimodales pueden combinar más de una fuente de datos para obtener mucha más información sobre el entorno y el usuario. Estos enfoques hacen posible la detección y clasificación de caídas al aprovechar diferentes modos selectivos de detección en la amplia gama de fuentes [12].

Debido a que los enfoques multimodales comprenden muchas fuentes diferentes de datos de sujetos y entornos, existen algunas debilidades como se informa en [60]: (i) mucha información requiere aplicar técnicas más robustas de extracción de características y selección de características, así como consideraciones en los enfoques de Machine learning para diferentes tipos de datos de entrada, haciendo que el sistema de detección de caídas sea una tarea computacionalmente costosa y difícil, y (ii) múltiples sensores con una ubicación compleja en el cuerpo (y en el medio ambiente) podrían causar mayores costos, dificultades prácticas de despliegue y molestias, especialmente para las personas mayores.

4.5. Sistemas de detección de caídas basados en visión

Este trabajo se centra principalmente en métodos basados en la visión. Tradicionalmente, los sistemas de detección de caídas se han abordado utilizando técnicas de visión por computadora y procesamiento de imágenes en marcos de ventanas para clasificar las actividades. Con los avances recientes, los sensores no invasivos de imágenes en profundidad producen imágenes profundas de alta calidad. Esta información también se analiza para el seguimiento humano, el monitoreo y los sistemas de reconocimiento de usuarios [14-16], y también para monitorear y reconocer las actividades diarias de los sujetos en ambientes interiores [17].

La mayoría de los enfoques basados en la visión se han trabajado con cámaras RGB simples, cámaras web, sistemas de cámaras de movimiento o incluso Kinect [18]. Por ejemplo, el uso de este último para la detección de caídas ha aumentado dado que puede obtener información en 3D, como la pose humana o el seguimiento de las extremidades [18].

Las estrategias clásicas de detección y clasificación de caídas basadas en la visión consisten en cinco fases [4], como sigue: (1) adquisición de datos de secuencias de video, (2) extracción de características de imágenes, (3) selección de características y (4) aprendizaje e inferencia. Existen múltiples técnicas de machine learning utilizadas en la literatura, como SVM [35] o random forest [36]. Zerrouki et al [17] propuso un sistema de detección de caídas basado en la variación de la forma de la silueta humana en el monitoreo de la visión y SVM para identificar posturas. Luego, utilizaron HMM para clasificar los datos en eventos de caída y no caída. Rougier et al [7] siguió la silueta de la persona junto con las secuencias de video. La deformación de la forma se cuantificó a

partir de estas siluetas en base a los métodos de análisis de forma. Finalmente, se detectaron caídas de las actividades diarias utilizando modelos de mezcla gaussiana (GMM).

Los sistemas basados en visión se pueden abordar por dos categorías: sistemas monoculares y sistemas de cámaras múltiples. En los sistemas de detección de caídas basados en monoculares, depende de un punto de vista. Mover una cámara a diferentes puntos de vista requeriría recopilar nuevos datos de entrenamiento para ese punto de vista específico y una nueva calibración de la cámara. Sin embargo, estos sistemas pueden fallar debido a la oclusión de objetos entre el objetivo y la cámara. Zhang et al [66] propuso múltiples dispositivos Kinect para lograr ese problema utilizando su propio conjunto de datos OCCU que fue creado con caídas ocluidas y no ocluidas. Kwolek et al [21] extrajo mapas de profundidad sobre el entorno y la silueta de la persona en combinación con acelerómetros de 3 ejes y SVM como técnica de machine learning. En términos de sistemas de detección de caídas multicámara, Thome y Miguet [22] propusieron utilizar un HMM para distinguir las caídas de las actividades para caminar. Las características extraídas para el análisis de movimiento se obtuvieron de una rectificación de imagen métrica en cada vista. Anderson et al [23] analizó los estados de los objetos 3D, llamados vóxel de una persona, obtenidos de dos cámaras. Todos estos trabajos construyen modelos 3D con múltiples cámaras para reconstruir el entorno. Esta tarea es particularmente difícil porque las cámaras deben calibrarse para calcular correctamente la información 3D. También presenta problemas en la sincronización de secuencias de video de cada cámara, lo que hace que sea más difícil de implementar que un enfoque monocular.

Por lo tanto, desde el punto de vista de la implementación de estos sistemas, las cámaras múltiples 2D suelen ser una mejor opción, principalmente debido al bajo costo y la facilidad de implementación. También es importante resaltar que las cámaras ya están instaladas en muchos lugares públicos, como aeropuertos, tiendas y centros de atención para personas mayores, que también pueden ocuparse para sistemas de detección de caídas. Por esas razones, las cámaras múltiples 2D son relevantes para el dominio de la aplicación de detección de caídas.

4.6. Sistemas de detección de caídas basados en la visión usando CNN

Los trabajos recientes sobre los sistemas de reconocimiento de caídas se han aprovechado del éxito de Machine learning en tareas de reconocimiento y clasificación utilizando imágenes regulares, imágenes profundas, imágenes infrarrojas, etc. El Machine learning CNN trabaja buscando características relevantes en imágenes evitando las tareas de ingeniería de características y proporcionando una extracción automática de características muy versátil dependiendo de su arquitectura de capas convolucionales e inferencia [25].

Por ejemplo, algunos de los trabajos recientes sobre sistemas de detección de caídas reportados en la literatura consideran [70] en el que los autores usan filtros basados en reglas antes de una capa convolucional de entrada que combina la salida de capa convolucional con características de flujo óptico para elegir una mejor entrada para la fase de inferencia de su arquitectura CNN 3D, logrando una precisión del 92.67%. En [72], los autores usan imágenes infrarrojas (IR) y una CNN 3D para encontrar características en tres canales de color en situaciones reales, teniendo en cuenta una información de imagen espacio-temporal, logrando un 85% de precisión en las secuencias de video de prueba.

Existen múltiples trabajos que utilizan CNN en sistemas de detección de caídas basados en visión monocular con excelentes resultados [37] [50-52]. Además, hay varios trabajos que utilizan un enfoque de múltiples cámaras con diferentes modelos clásicos Machine learning u otros algoritmos [54-58] y solo hay un trabajo que utiliza múltiples cámaras y CNN en el sistema de detección de caídas [59].

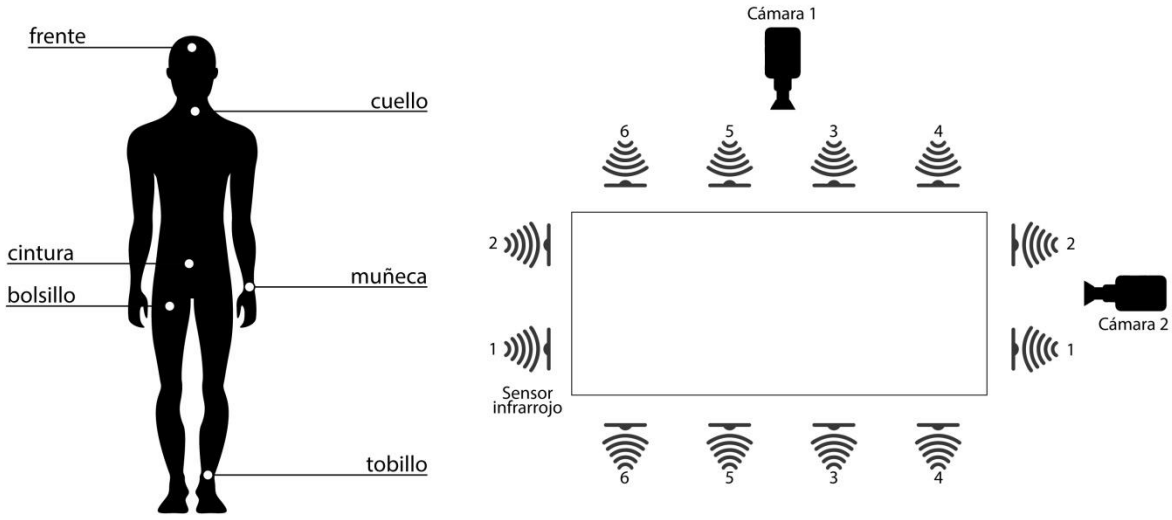
5. Descripción del conjunto de datos

En esta investigación, utilizamos un conjunto de datos públicos llamado UP-Fall Detection [24]. Este conjunto de datos se realizó con la información de 17 jóvenes voluntarios sanos sin ningún tipo de deterioro (9 hombres y 8 mujeres) que van desde los 18-24 años de edad, la altura media de 1,66 y un peso medio de 66,8 kg realizan 11 actividades y 3 ensayos por actividad, seis actividades diarias humanas simples y cinco tipos diferentes de caídas humanas utilizando un enfoque multimodal, con sensores portátiles, sensores ambientales y dispositivos de visión. El conjunto de datos consolidado y el conjunto de datos de características están disponibles públicamente.

Las actividades y caídas almacenadas en este conjunto de datos se resumen en la Tabla 1. Todos los datos se recolectaron utilizando 14 dispositivos: cinco sensores portátiles de Mbientlab MetaSensor que recopilan datos sin procesar del acelerómetro de 3 ejes, giroscopio de 3 ejes y sensor de luz ambiental; se ocupó un auricular NeuroSky MindWave de electroencefalograma (EEG) para medir la señal de ondas cerebrales sin procesar de su sensor de canal EEG único ubicado en la frente; como sensores sensibles al contexto, instalamos seis sensores infrarrojos como una cuadrícula a 0.40 m sobre el piso de la habitación, para medir los cambios en la interrupción de los dispositivos ópticos (donde 0 significa interrupción y 1 sin interrupción); y, por último, dos cámaras Microsoft LifeCam Cinema se ubicaron a 1,82 m sobre el piso, una para una vista lateral y la otra para una vista frontal en relación con el sujeto. Para obtener más información sobre el UP-Fall Detection data set, consulte la referencia [24].

Tabla 1. Actividades realizadas por sujetos, adaptada de [24].

Actividad ID	Descripción	Duración	Abreviación
1	Caer hacia adelante usando las manos	10s	FH
2	Caer hacia adelante usando las rodillas	10s	FF
3	Caer hacia atrás	10s	FB
4	Caer de lado	10s	FS
5	Caer sentado en una silla vacía	10s	FE
6	Camina	60s	W
7	En pie	60s	S
8	Sentado	60s	ST
9	Recolectando un objeto del piso	10s	P
10	Saltando	30s	J
11	Tendido	60s	L



(a) (b)

Figura 1. Distribución de sensores. (a) sensores vestibles y casco EEG en el cuerpo humano. (b) Distribución de sensores de ambiente y multicámaras. Adaptado de [24].

En este trabajo, solo usamos la información de las dos cámaras en el conjunto de datos, aprovechando las distribuciones de múltiples cámaras. En ese sentido, nuestro objetivo es implementar un sistema de detección y clasificación de caídas utilizando múltiples cámaras, y comparar su rendimiento cuando solo se utiliza un enfoque monocular. Para este fin, CNN se ocupará como modelo de clasificación.

6. Descripción de la propuesta

En este trabajo, adoptamos el flujo de trabajo tradicional para sistemas de detección de caídas [8] que consta de los siguientes pasos: (i) recopilación de datos, (ii) ventanas, (iii) extracción de características y (iv) aprendizaje e inferencia. Se muestra en la Figura 2.

Figura 2. Flujo de trabajo tradicional para los sistemas de detección de caídas.

6.1. Recopilación de datos

Una de las fases más desafiantes en el flujo de trabajo tradicional para sistemas de detección de caídas y, en general, problemas de Machine learning es la tarea de recopilación de datos. Más recientemente, las técnicas de Deep learning requieren una gran cantidad de datos para ser entrenados y probados correctamente y otros factores como el número de individuos, sus características físicas, el número de individuos con características diversas en términos de género, edad, altura, peso y condiciones de salud [8].

Como explicamos en la sección Descripción del conjunto de datos, utilizamos el UP-Fall detección dataset para lograr este paso de recopilación de datos. Para resumir, el conjunto de datos contiene información sobre 17 sujetos jóvenes que realizan 11 actividades diferentes, incluidas 5 caídas y 6 actividades. Para este trabajo, utilizamos la información recopilada de dos cámaras RGB ubicadas en diferentes puntos de vista (vistas lateral y frontal), capturando imágenes de los sujetos [24].

6.2. Ventanas

Los enfoques de ventanas en los sistemas de detección de caídas normalmente se utilizan para segmentar series temporales de caídas realizadas. La segmentación corresponde al

proceso de dividir las señales del sensor en segmentos de datos más pequeños. Este proceso se ha realizado de diferentes maneras en el campo de reconocimiento de actividad y en los sistemas de detección de caídas. La mayoría de las técnicas de segmentación podrían clasificarse en tres grupos, ventanas definidas por actividad, ventanas definidas por eventos y ventanas deslizantes [68].

Adoptamos un enfoque de ventanas deslizantes para capturar la dependencia temporal entre muestras. En este caso, dividimos todos los datos en ventanas de tiempo de longitud fija, para cada actividad y caída. Nuestra implementación utiliza ventanas de un segundo con 0,5 segundos de superposición. El resultado de este paso fueron múltiples series de imágenes de 1 segundo de longitud de ventana que se procesarán en el siguiente paso.

6.3. Extracción de características

La extracción de características es un método general en el que se intenta desarrollar una transformación del espacio de entrada en el subespacio de baja dimensión que conserva la mayor parte de la información relevante para mejorar el análisis de datos [69]. En los sistemas de detección de caídas, existen múltiples técnicas que dependen del tipo de datos para extraer información relevante, en los enfoques basados en la visión, el algoritmo de flujo óptico proporciona información muy rica sobre los movimientos aparentes en las imágenes, y se han utilizado en múltiples investigaciones que utilizaron la combinación entre CNN y flujo óptico como una característica extraída de las imágenes [70], [37].

Para la extracción de características, cada marco de ventana se procesa previamente para obtener información que podría proporcionar información suficiente para describir la actividad realizada. En este caso de estudio, el algoritmo de flujo óptico [24], [26] se utilizó como características visuales extraídas de cada cámara. Este algoritmo nos ayuda a obtener los desplazamientos aparentes entre dos marcos de ventanas y, de ese modo, distinguir movimientos y direcciones sin tener en cuenta las características estáticas de la imagen. Las características obtenidas son los movimientos relativos horizontales y verticales de píxeles en las imágenes, U y V, respectivamente [26]. La combinación resultante, D, de estos valores corresponde a la magnitud del movimiento relativo, como se muestra en la ecuación (1), donde las imágenes de matriz resultantes son del mismo tamaño que las imágenes de la ventana original, es decir, 640 x 480 píxeles en nuestro caso de estudio.

$$D_{i,j} = \sqrt{U^2{}_{i,j} + V^2{}_{i,j}} \quad (1)$$

Para evitar características ambientales y de color, calculamos cada imagen de la ventana en escala de grises, y luego ingresamos esta información de flujo óptico, como extracción de características, en una CNN. En este sentido, las capas convolucionales de nuestro modelo solo se enfocan para aprender características relevantes de los movimientos de píxeles en nuestras imágenes [25]. Finalmente, para minimizar el esfuerzo computacional, redimensionamos todas las imágenes de la ventana de 640 x 480 píxeles a 38 x 51 píxeles de tamaño. Este paso se realizó para cada cámara por separado.

6.4. Aprendizaje e inferencia

En la literatura [6], hay múltiples formas de lograr esta fase, dos de ellas son el machine learning y recientemente, los algoritmos de deep learning. Este paso busca entrenar y probar el resultado de la ingeniería de características para clasificar o predecir la caída realizada con entradas de sensores ambientales, sensores portátiles y, en este trabajo en particular, desde un enfoque basado en la visión de múltiples cámaras.

En deep learning CNN ha revolucionado la forma de abordar los problemas de visión por computadora debido al descubrimiento automático de la representación de la estructura en grandes conjuntos de datos. Este método ha mejorado dramáticamente el estado del arte en el procesamiento de imágenes [25].

Sin embargo, encontrar una arquitectura adecuada de la CNN es una tarea difícil [25]. En ese sentido, la literatura ha reportado múltiples tipos de arquitecturas de red, dependiendo de la resolución de problemas. Por ejemplo, en los últimos años, se han reportado muchas estructuras de red para el reconocimiento de imágenes y problemas de clasificación como: AlexNet [27], ClarifaiNet [28], GoogLeNet [29] y VGGNet [30]. Todas estas redes han demostrado ser eficientes en sus propios dominios problemáticos; pero también, pueden usarse como modelos pre-entrenados para que los usuarios puedan reducir la cantidad de tiempo al volver a entrenarlos para otra tarea específica. Sin embargo, estas son arquitecturas complejas que podrían mejorarse.

En este trabajo, diseñamos un CNN con tres capas convolucionales y tres capas 2D de agrupación máxima para la extracción de características, y luego, tres capas completamente conectadas para la detección de caídas. En ese sentido, fijamos el tamaño de todas las imágenes en 38 x 51 píxeles, se fijaron tres capas completamente conectadas y se eligieron debido a capas completamente conectadas, la restricción de tamaño fijo proviene solo de las capas completamente conectadas, que existen en una etapa más profunda de la red [71]. La CNN recibe las magnitudes D, calculadas a partir de U y V, que se convirtieron en imágenes en escala de grises con un tamaño de 38x51 píxeles, que representan las características de flujo óptico extraídas. Luego, estas imágenes llegan a la capa de entrada que consta de 128 filtros de convolución con un tamaño de núcleo de 3x3. La segunda capa convolucional tiene 128 filtros y el mismo tamaño de núcleo, y la tercera capa con 64 filtros convolucionales con el mismo núcleo. Esta arquitectura de capa convolucional se seleccionó mediante validación cruzada y utilizando la métrica de puntaje F1, como se muestra en la Tabla 2. Después de cada capa convolucional, se emplean capas 2D de agrupación máxima para sintetizar convoluciones de salida. Después de eso, estos resultados se ingresan en tres capas completamente conectadas, es decir, 64 unidades lineales rectificadas (ReLU) en la primera capa, 128 ReLU en la segunda 254 ReLu en la tercera y finalmente en la capa softmax 2D con una salida. Este último se emplea para realizar la detección de caídas: clases de caída (1) o no caída (0). La Figura 3 muestra la representación de la CNN propuesta.

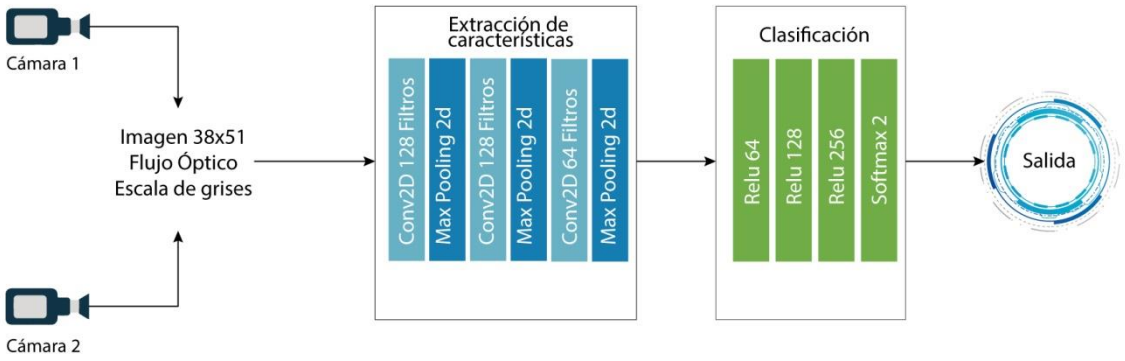


Figura 3. Nuestra propuesta de arquitectura de CNN para el sistema para detección de caídas con multicámaras.

Tabla 2. Cross-validation para la arquitectura de capas de convolución de nuestra CNN.

Arquitectura CNN	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-Score (%)
64 64 64	95.40	86.28	83.76	97.54	95.40
64 64 128	95.27	88.26	80.34	98.03	84.11
64 64 256	94.90	83.67	83.57	96.99	83.62
64 128 64	94.62	82.36	83.27	96.71	82.81
64 128 128	94.66	86.49	77.83	97.76	81.94
64 128 256	95.15	85.74	82.35	97.46	84.14
64 256 64	94.32	85.86	76.00	97.69	80.63
64 256 128	94.92	86.45	79.91	97.69	83.05
64 256 256	94.90	91.18	74.48	98.67	81.98
128 64 64	95.17	86.21	82.11	97.58	84.11
128 64 128	94.80	96.02	97.89	78.02	96.95
128 64 256	94.79	97.07	96.74	84.18	96.91
128 128 64	95.64	96.91	97.95	83.08	97.43
128 128 128	95.44	96.19	98.49	78.87	97.33
128 128 256	95.05	97.88	96.22	88.70	97.04
128 256 64	94.28	96.32	96.92	79.91	96.62
128 256 128	94.51	97.00	96.47	83.82	96.74
128 256 256	95.19	96.84	97.48	82.78	97.16
256 64 64	94.81	96.16	97.76	78.81	96.95
256 64 128	94.26	96.25	96.97	79.54	96.61
256 64 256	94.38	96.34	97.03	80.03	96.68
256 128 64	94.75	96.19	97.64	79.05	96.91
256 128 128	94.72	97.63	96.08	87.36	96.85
256 128 256	94.40	96.66	96.71	81.86	96.68
256 256 64	94.10	96.31	96.71	79.91	96.51
256 256 128	94.57	96.36	97.24	80.09	96.80
256 256 256	94.09	96.19	96.83	79.18	96.51

Como se describió anteriormente, el UP-Fall Detection data set está integrado por información de 17 sujetos que realizaron 11 actividades / caídas diferentes en tres ensayos diferentes para cada actividad. Para entrenar a la CNN, dividimos los datos tomando las pruebas 1 y 2 para cada actividad y materia como conjunto de entrenamiento (67%), y la prueba 3 para cada actividad y materia como conjunto de prueba (33%). El conjunto de datos de entrenamiento comprendía de 42,000 imágenes en escala de grises de tamaño 38 x 51 con el flujo óptico como preprocesamiento; mientras que el conjunto de datos de prueba comprendía de 21,000 imágenes en escala de grises

con el mismo flujo óptico preprocesador. Para fines de entrenamiento, entrenamos durante 50 épocas, utilizando el optimizador Adam y la función de pérdida de entropía cruzada binaria, como se define en la Ecuación (2) donde p es la predicción de la red y t es la verdad fundamental.

$$loss\ (p,t) = -(t * log\ log\ (p)\ + (1-t) * log\ (1-p)) \quad (2)$$

7. Experimentación

Para analizar nuestra propuesta, se realizaron los siguientes experimentos en tres ramas: (i) experimentos para probar nuestro modelo CNN y compararlo con los métodos clásicos de machine learning SVM, RF, MPL, kNN, (ii) experimentos para comparar enfoques monoculares con sistemas de detección de caídas basados en visión multicámara y (iii) probar nuestra propuesta no solo para la detección, sino también en la clasificación de actividades y caídas utilizando el enfoque basado en visión multicámara.

En estos experimentos, utilizamos conjuntos de datos de entrenamiento y prueba con información que proporcionan las dos cámaras. Decidimos usar la información de una cámara por modelo, y luego, usar las cámaras de punto de vista lateral y frontal al mismo tiempo [8]. Para la ventana, se emplearon períodos de tiempo fijos de 1 segundo con solapamiento de 0,5 segundos. Las imágenes fueron tratadas como escala de grises e implementación flujo óptico como extracción de características. Cambiamos el tamaño de las imágenes a 38 x 51 píxeles, y realizamos un punto de referencia entre los métodos clásicos de machine learning (es decir, SVM, MLP, RF y kNN) y el CNN representado en la Figura 3.

Estos experimentos tienen como objetivo explorar y comparar el rendimiento entre sistemas de detección de caídas basados en visión monocular con sistemas de detección de caídas basados en visión multicámara, y también hacer un punto de referencia de los métodos clásicos de machine learning y CNN para la detección de caídas utilizando este último enfoque.

Para medir el rendimiento de nuestro trabajo, utilizamos cinco métricas: precisión, sensibilidad, especificidad, precisión y puntaje F1. Como se muestra en las Ecuaciones (3) - (7) donde TP se refiere a positivos verdaderos, TN a negativos verdaderos, FP a falsos positivos y FN a falsos negativos [32].

$$Exactitud = \frac{TP+TN}{2TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TN+FP} \quad (4)$$

$$Sensitividad = \frac{TP}{TP+FN} \quad (5)$$

$$Especificdad = \frac{TN}{TN+FP} \quad (6)$$

$$F1 - score = 2 * \frac{precision\ x\ sensitivity}{precision+sensitivity} \quad (7)$$

Con este fin, todos estos experimentos se implementaron en Python 3.7.3 utilizando el framework sklearn para las técnicas clásicas de Machine learning y el framework keras para que CNN aproveche su gestión de GPU [31].

8. Resultados y discusión

Los resultados experimentales se describen en esta sección. Después de eso, se presenta una discusión del análisis.

8.1. Detección de caídas utilizando modelos convencionales de machine learning.

Primero, realizamos un experimento utilizando las funciones ópticas basadas en el flujo de ambas cámaras al mismo tiempo (Cam1 y Cam2). Entrenamos cuatro modelos convencionales de machine learning: SVM, RF, MLP y KNN, como se describió anteriormente. La Tabla 3 muestra la configuración de metaparámetros para estos modelos.

Tabla 3. Parámetros usados para el entrenamiento en la clasificación de modelos.

Clasificador	Parámetros
SVM	kernel = “radial basis function” kernel coefficient = 1/num_features c = 1 shrinking = 1 tolerance = 0.001
RF	minimum samples split = 2 minimum samples leaf = 1 estimators = 2 bootstrap = 1
MLP	activation function = “reLU” hidden layers = 100 penalty parameter = 0.0001 batch size = min(200, num_samples) shuffle = 1 initial learning rate = 0.001 tolerance = 0.0001 exponential decay (first moment) = 0.9 exponential decay (second moment) = 0.999 regularization coefficient = 0.000000001 solver = “stochastic gradient” maximum epochs = 10
KNN	neighbors = 5 leaf size = 30 distance metric = “euclidean”

Para este experimento, creamos los modelos usando 67% para entrenamiento y 33% para datos de prueba. La Tabla 4 resume los resultados de rendimiento utilizando las características visuales extraídas en ventanas de 1 segundo de longitud con 0,5 segundos de superposición.

De la Tabla 4, se puede observar que los modelos convencionales de machine learning no pueden predecir las caídas humanas con un buen rendimiento en términos de precisión, precisión, sensibilidad, especificidad o puntaje F1. Actualmente, KNN parece ser el mejor rendimiento basado en la métrica de puntaje F1 (15.27%). En términos de precisión, SVM se desempeña mejor con 32.40%. Al final, estos modelos de machine learning lograron una precisión promedio de 29.77%. A partir de estos resultados, podríamos suponer que los métodos convencionales de machine learning que utilizan ventanas y extracción de características, como se explicó anteriormente, no son lo suficientemente sólidos. Para

mejorar este rendimiento, consideramos implementar CNN, como se describe a continuación.

Tabla 4. Desempeño obtenido por los modelos de machine learning tradicional.

Modelo	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-Score (%)
SVM	32.40	14.03	14.10	90.03	14.0649
RF	29.30	14.45	14.30	91.26	14.3746
MPL	30.08	9.05	11.03	93.65	9.9423
KNN	27.30	16.32	14.35	90.96	15.2717

8.2. Detección de caídas usando CNN

En este experimento, entrenamos tres modelos diferentes de CNN: (i) un modelo de CNN que usa características visuales desde la vista lateral (Cam1), (ii) un modelo de CNN que usa características visuales solo desde la vista frontal (Cam2) y (iii) un modelo CNN que utiliza características visuales de ambas cámaras al mismo tiempo.

El resumen de resultados se incluye en la Tabla 5. Como se puede observar, podemos ver que el rendimiento es muy similar en cualquiera de las combinaciones. En comparación, la vista lateral (Cam1) es ligeramente mejor que la vista frontal, como se esperaba [7]. Sin embargo, Cam2 muestra menos especificidad (79.67%) que Cam1 (81.58%), lo que podría conducir a una clasificación errónea. Además, la combinación de ambas vistas mantiene el rendimiento de salida de la vista lateral. Esto es importante porque si ocurre la oclusión en algunas de las cámaras, será factible que la detección se realice con una sola cámara, como lo respalda la literatura [7]. Por otro lado, hicimos un experimento usando la arquitectura CNG VGG-16 con imágenes de 2 cámaras (frontal y lateral), los resultados que se muestran en la Tabla 5 indican que nuestra propuesta tiene un rendimiento significativamente mejor que la arquitectura CNN VGG-16 que usa UP-Fall. Por lo tanto, concluimos que nuestro sistema de detección de caídas basado en visión multicámara tiene un rendimiento aceptable, en contraste con los modelos convencionales de Machine learning y el uso de la arquitectura VGG-16 CNN, además de que evita el problema de oclusión siempre que no perdamos de vista del sujeto.

Tabla 5. Desempeño de nuestra propuesta de CNN utilizando vista lateral, vista frontal y ambas vistas.

Data	Método	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-Score (%)
(Cam1) Vista lateral	Propuesta CNN	95.24	96.68	97.72	81.58	97.20
(Cam2) Vista frontal	Propuesta CNN	94.78	96.30	97.57	79.67	96.93
(Cam1 & Cam 2)	Propuesta CNN	95.64	96.91	97.95	83.08	97.43
(Cam1 & Cam 2)	VGG-16 CNN	84.44	84.44	100	0	91.56

Por otro lado, comparamos nuestro método propuesto en contraste con otros sistemas de detección de caídas basados en la visión de múltiples cámaras reportados en la literatura [15], [35], [37], considerando que estos últimos se implementaron utilizando métodos convencionales de machine learning. Para esta comparación, utilizamos la base de datos basada en visión multicámara, llamada conjunto de datos Multicam [34]. Este conjunto de datos comprende 24 actuaciones en las que 22 ensayos tienen al menos una caída humana y los dos restantes contienen eventos de confusión. Cada actuación ha sido grabada desde 8 vistas diferentes. El mismo escenario se utiliza para todos los videos, con algunas reasignaciones de muebles [34]. Para fines de entrenamiento de nuestra propuesta, seleccionamos dos puntos de vista (vistas lateral y frontal) de este conjunto de datos, dividiendo el entrenamiento (67%) y las pruebas (33%). La Tabla 6 resume los resultados de rendimiento en términos de sensibilidad y especificidad, como se informa en la literatura [33], [35], [37].

Como se muestra en la Tabla 6, se ve que nuestro método propuesto puede ser competitivo en términos de vanguardia, principalmente sobre la sensibilidad. Además, nuestro método puede manejar la detección de caídas utilizando dos cámaras, en contraste con las ocho cámaras utilizadas en los otros enfoques. Además, la arquitectura de red de nuestra propuesta (Figura 3) es muy simple en comparación con otros trabajos. Por ejemplo, Núñez-Marcos en [37] utilizó una arquitectura VGG-16 modificada para recibir entradas, los autores en [33] ocuparon PCA para extraer características y SVM para la clasificación, y en [35] autores presentaron un promedio móvil multivariado ponderado exponencialmente (MEWMA) y SVM con 2 pasos para la clasificación (ver Tabla 6). En ese sentido, nuestro sistema tiene un buen rendimiento, teniendo en cuenta su tiempo mucho menor para la capacitación y la simplicidad de su arquitectura.

Tabla 6. Comparación de nuestra propuesta contra otros Sistemas de detección de caídas con multicámaras reportados en el estado del arte usando Multicam dataset.

Propuesta	Método	Sensibilidad (%)	Especificidad (%)	Cámaras
Wang et al. [33]	SVM	89.20	90.30	8
Wang et al. [35]	SVM	93.70	92.00	8
Núñez et al. [37]	VGG-16 CNN	99.00	96.00	8
Nuestra propuesta (Combinado)	Propuesta CNN	97.95	83.08	2

8.3. Actividades diarias y Clasificación de caídas usando CNN

Por último, realizamos un experimento para actividades diarias y clasificación de caídas utilizando nuestra propuesta. En este caso, se tuvo en cuenta cada actividad y tipo de caída registrada en el UP-Fall Detection data set, por lo que la CNN se convirtió en un clasificador de varias clases, como se muestra en la Tabla 1.

Aplicamos nuestra propuesta usando ambas cámaras (Cam1 y Cam2) y los resultados, en contraste con el rendimiento obtenido en [24] usando el mismo conjunto de datos, se muestran en la Tabla 7. Como se muestra, nuestra propuesta es ligeramente peor que los resultados del enfoque multimodal presentados por Martínez-Villaseñor et al [24]. Este es un resultado esperado ya que un enfoque multimodal (es decir, sensores portátiles, casco EEG y cámaras) es mejor que una modalidad única como la nuestra. También es importante notar que el puntaje F1 en ambos enfoques es similar, 72.94% para nuestra propuesta y 72.80% para el enfoque multimodal. A partir de los resultados presentados en la Tabla 7, el rendimiento obtenido por nuestra propuesta puede considerarse competitivo (por ejemplo, puntaje F1 similar), más fácil de implementar (es decir, debido a la cantidad de sensores) y menos molesto (es decir, sensores portátiles), en comparación con el enfoque multimodal informado en [24].

Tabla 7. Desempeño de nuestra propuesta de CNN para clasificación de caídas y actividades diarias.

Data	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-Score (%)
Nuestra propuesta	82.26	74.25	71.67	77.48	72.94
Martínez-Villaseñor, et al. [24]	95.00	77.70	69.90	99.50	72.80

9. Discusión

El sistema de detección y clasificación de caídas basado en la visión multicámara propuesto ofrece una solución comparable a los métodos de vanguardia. Los resultados respaldan la evidencia sobre: el poder predictivo de nuestro sistema de detección de caídas propuesto utilizando dos puntos de vista (97.43% de la puntuación F1), el rendimiento superior de los métodos convencionales de machine learning (SVM, RF, MLP y KNN) utilizando características basadas en flujo óptico, el uso de un menor número de cámaras, con un rendimiento similar, que otras reportadas en el estado del arte (97.00% de sensibilidad y 80.00% de especificidad), y un rendimiento similar (70.81% de la puntuación F1) comparable a un enfoque multimodal (72,80% de la puntuación F1).

De lo anterior, las ventajas de nuestra propuesta se pueden señalar de la siguiente manera. Los enfoques multicámara ofrecen soluciones robustas que reconocen las caídas, aunque cuando se produce una oclusión en un punto de vista, siempre que una cámara mantenga el foco en el sujeto. Esto se puede observar en nuestra propuesta en la Tabla 5 que informa un rendimiento similar al usar una cámara u otra (vista lateral o frontal), o incluso ambas. Además, nuestra propuesta ofrece una arquitectura CNN simple (Figura 3) y un menor costo computacional de implementación. Debido a la naturaleza basada en la visión de nuestro enfoque, un buen punto para discutir es la invasión de la privacidad debido a la video vigilancia constante. En este sentido, nuestros trabajos evitan este problema analizando solo la información relevante sobre la caída en las imágenes utilizando la información de flujo óptico calculada a partir de la secuencia de video. Por lo tanto, la privacidad de la persona no se ve afectada debido a que los datos utilizados para reconocer una caída no contienen información personal.

Por otro lado, es importante tener en cuenta algunas limitaciones de nuestra propuesta durante su uso. Un enfoque basado en la visión siempre está sujeto a la calidad de la imagen capturada, la posición de las cámaras y la presencia del sujeto interesado. Además, los problemas de privacidad deben abordarse antes de la implementación. A menos que esto sea una limitación, como se dijo antes y que se complementa, las imágenes originales tomadas de las cámaras no deben almacenarse; solo deben usarse para extraer las características de flujo óptico. Sin embargo, la omnipresencia sigue siendo un inconveniente importante ya que las cámaras siempre obtienen videos de los sujetos. Además, también debe destacarse la complejidad computacional en términos de memoria y procesamiento de tiempo. De hecho, esto dificulta que un sistema de detección de caídas en tiempo real sea escalable [8].

En términos de la cantidad de muestras recuperadas de las caídas y actividades realizadas en el UP-Fall dataset, se analizaron 42,958 muestras de entrenamiento dispuestas en ventanas de 1s y se emplearon 21,038 muestras de prueba también dispuestas en ventanas de 1s en nuestros experimentos. Los resultados obtenidos fueron el competitivos relativos de vanguardia en las tareas de detección (Tabla 5 y Tabla 6) y clasificación (Tabla 7).

Además, es importante analizar la edad de los sujetos que realizaron caídas y actividades al construir el UP-Fall Detection dataset utilizado en este trabajo. Este conjunto de datos se realizó con información de 17 sujetos jóvenes sanos sin ningún impedimento (9 hombres y 8 mujeres) con edades comprendidas entre 18 y 24 años. Sin embargo, en [73], se muestra que el uso de un conjunto de datos construido solo por jóvenes no tiene pruebas de discrepancias significativas con las personas mayores. En ese sentido, consideramos que nuestro enfoque puede aplicarse en situaciones reales, consideradas como trabajo futuro.

Con este fin, los resultados experimentales mostraron que nuestra propuesta es competitiva en comparación con lo más avanzado en enfoques basados en visión multicámara para sistemas de detección, y también es competitiva en la clasificación de caída (Tabla 6), incluso en contraste a un enfoque multimodal como se informó en [24].

10. Conclusiones

En este artículo, presentamos un sistema de detección y clasificación de caídas basado en la visión de múltiples cámaras que aprovecha CNN. Además, combinamos los modelos CNN con características visuales extraídas de secuencias de imágenes utilizando el método de flujo óptico. En este trabajo, utilizamos el UP-Fall Detection dataset como caso de estudio. Realizamos diferentes experimentos para: comparar nuestra propuesta con modelos convencionales de machine learning, analizar el rendimiento de nuestra propuesta en enfoques basados en la visión de una o varias cámaras, y también extender nuestro modelo para la clasificación de caídas.

A partir de los resultados experimentales, llegamos a la conclusión de que nuestro sistema de clasificación y detección de caídas basado en la visión multicámara supera a los métodos convencionales de machine learning, ahorra procesos de computación debido a la arquitectura CNN simple y es competitivo con el estado del arte y enfoques multimodales.

Por último, los trabajos futuros que consideren implementar este enfoque en un sistema de vida asistida en el mundo real, y analizar y proponer mejoras a los problemas de privacidad, omnipresencia, cambios en las condiciones ambientales y oclusión. Además, consideraremos probar nuestro sistema en una situación real.

11. Bibliografía

[1] Department of Health and Human Services. Fatalities and injuries from falls among older adults - United States, 1993-2003 and 2001- 2005. pages 1221–1224, November 2006. Morbidity and Mortality Weekly Report.

[2] Schneider, M. (2011). Introduction to public health. Sudbury, MA: Jones and Bartlett.

[3] Lord, S. R., Sherrington, C., Menz, H. B., & Close, J. C. (n.d.). Strategies for prevention. Falls in Older People,173-176. doi:10.1017/cbo9780511722233.011

[4] Oneill, T. W., Varlow, J., Silman, A. J., Reeve, J., Reid, D. M., Todd, C., & Woolf, A. D. (1994). Age and sex influences on fall characteristics. Annals of the Rheumatic Diseases,53(11), 773-775. doi:10.1136/ard.53.11.773

[5] Bourke, A., & Lyons, G. (2008). A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. Medical Engineering & Physics,30(1), 84-90. doi:10.1016/j.medengphy.2006.12.001

[6] Noury, N., Fleury, A., Rumeau, P., Bourke, A., Laighin, G. O., Rialle, V., & Lundy, J. (2007). Fall detection - Principles and Methods. 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. doi:10.1109/iembs.2007.4352627

[7] Rougier, C., Meunier, J., St-Arnaud, A., & Rousseau, J. (2011). Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. IEEE Transactions on Circuits and Systems for Video Technology,21(5), 611-622. doi:10.1109/tcsvt.2011.2129370

[8] Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. 2013, 15, 1192–1209.

[9] Yin, J., Yang, Q., & Pan, J. (2008). Sensor-Based Abnormal Human-Activity Detection. IEEE Transactions on Knowledge and Data Engineering,20(8), 1082-1090. doi:10.1109/tkde.2007.1042

[10] Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., & Qiu, Y. (2013). Exploring Techniques for Vision Based Human Activity Recognition: Methods, Systems, and Evaluation. Sensors, 13(2), 1635-1650. doi:10.3390/s130201635

[11] Dungkaew, T., Suksawatchon, J., & Suksawatchon, U. (2017). Impersonal smartphone-based activity recognition using the accelerometer sensory data. 2017 2nd International Conference on Information Technology (INCIT). doi:10.1109/incit.2017.8257856

[12] Bharti, P. (2017). Complex activity recognition with multi-modal multi-positional body sensing. Journal of Biometrics & Biostatistics,08(05). doi:10.4172/2155-6180-c1-005

[13] Chetty, G., White, M., Singh, M., & Mishra, A. (2014). Multimodal activity recognition based on automatic feature discovery. 2014 International Conference on Computing for Sustainable Global Development (INDIACom). doi:10.1109/indiacom.2014.6828039

[14] Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. 2008, 18, 1473–1488.

[15] Raty, T.D. Survey on contemporary remote surveillance systems for public safety. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 2010, 40, 493–515.

[16] Albanese, M.; Chellappa, R.; Moscato, V.; Picariello, A.; Subrahmanian, V.S.; Turaga, P.; Udrea, O.A constrained probabilistic petri net framework for human activity detection in video. IEEE Trans. Multimed. 2008, 10, 1429–1443.

[17] Zerrouki, N., & Houacine, A. (2017). Combined curvelets and hidden Markov models for human fall detection. Multimedia Tools and Applications, 77(5), 6405-6424. doi:10.1007/s11042-017-4549-5

[18] E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier, “Fall detection with multiple cameras: An occlusion resistant method based on 3-D silhouette vertical distribution,” IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 2, pp. 290–300, 2011.

[19] Núñez-Marcos A, Azkune G, Arganda-Carreras I (2017) Vision-based fall detection with convolutional neural networks. Wireless communications and mobile computing 2017: https://doi.org/10.1155/2017/9474806

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 770–778, July 2016.

[21] B. Kwolek and M. Kepski, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” Computer Methods and Programs in Biomedicine, vol. 117, no. 3, pp.489–501, 2014.

[22] Thome, N., Miguet, S., & Ambellouis, S. (2008). A Real-Time, Multiview Fall Detection System: A LHMM-Based Approach. IEEE Transactions on Circuits and Systems for Video Technology,18(11), 1522-1532. doi:10.1109/tcsvt.2008.2005606

[23] Anderson, D., Luke, R. H., Keller, J. M., Skubic, M., Rantz, M., & Aud, M. (2009). Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Computer Vision and Image Understanding,113(1), 80-89. doi:10.1016/j.cviu.2008.07.006

[24] Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., & Peñafort-Asturiano, C. (2019). UP-Fall Detection Dataset: A Multimodal Approach. Sensors,19(9), 1988. doi:10.3390/s19091988

[25] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[26] S. S. Beauchemin and J. L. Barron, “The Computation of Optical Flow,” ACM Computing Surveys, vol. 27, no. 3, pp.433–466, 1995.

[27] A. Krizhevsky, I. Sutskever, andG. E.Hinton, “Imagenet classification with deep convolutional neural networks,” in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS ’12), pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In ECCV, pages 818–833, 2014.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 1, 2, 3

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014. 1, 2

[31] C. Francois and et al., “Keras,” 2015, https://github.com/fchollet/keras.

[32] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management,45(4), 427-437. doi:10.1016/j.ipm.2009.03.002

[33] Wang, S., Chen, L., Zhou, Z., Sun, X., & Dong, J. (2015). Human fall detection in surveillance video based on PCANet. Multimedia Tools and Applications,75(19), 11603-11613. doi:10.1007/s11042-015-2698-y

[34] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. DIRO-Université de Montréal, Tech. Rep, 1350, 2010.

[35] Charfi, I.; Miteran, J.; Dubois, J.; Atri, M.; Tourki, R. Definition and Performance Evaluation of a

Robust SVM Based Fall Detection Solution. SITIS 2012, 12, 218–224.

[36] Kozina, S., Gjoreski, H., Gams, M., & Luštrek, M. (2013). Efficient Activity Recognition and Fall Detection Using Accelerometers. Communications in Computer and Information Science Evaluating AAL Systems Through Competitive Benchmarking,13-23. doi:10.1007/978-3-642-41043-7_2

[37] K.Wang, G. Cao, D. Meng,W. Chen, andW. Cao, "Automatic fall detection of human in video using combination of features," in Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, pp. 1228–1233, China, December 2016.

[38] Blanc-Talon, J. (2006). Advanced concepts for intelligent vision systems: 8th International Conference, ACIVS 2006: Antwerp, Belgium, September 18-21, 2006: Proceedings. Berlin: Springer.

[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Training computationally efficient smartphone-based human activity recognition models," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8131 LNCS, pp. 426–433, 2013.

[40] Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., & Dürichen, R. (2017). CNN-based sensor fusion techniques for multimodal human activity recognition. Proceedings of the 2017 ACM International Symposium on Wearable Computers - ISWC 17. doi:10.1145/3123021.3123046

[41] L. C. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, "Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity," in 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5250–5253, 2008.

[42] Jalal, A., Kamal, S., & Kim, D. (2014). A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. Sensors,14(7), 11735-11759. doi:10.3390/s140711735

[43] Torres-Huitzil, C., & Nuno-Maganda, M. (2015). Robust smartphone-based human activity recognition using a tri-axial accelerometer. 2015 IEEE 6th Latin American Symposium on Circuits & Systems (LASCAS). doi:10.1109/lascas.2015.7250435

[44] Vilarinho, T.; Farshchian, B.; Bajer, D.G.; Dahl, O.H.; Egge, I.; Hegdal, S.S.; Lønes, A.; Slettevold, J.N.; Weggersen, S.M. A combined smartphone and smartwatch fall detection system. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 26–28 October 2015; pp. 1443–1448.

[45] Vavoulas, G., Pediaditis, M., Chatzaki, C., Spanakis, E. G., & Tsiknakis, M. (n.d.). The MobiFall Dataset. Artificial Intelligence,1218-1231. doi:10.4018/978-1-5225-1759-7.ch048

[46] Kerdjidj, O., Ramzan, N., Ghanem, K., Amira, A., & Chouireb, F. (2019). Fall detection and human activity classification using wearable sensors and compressed sensing. Journal of Ambient Intelligence and Humanized Computing. doi:10.1007/s12652-019-01214-4

[47] Khojasteh, S., Villar, J., Chira, C., González, V., & Cal, E. D. (2018). Improving Fall Detection Using an On-Wrist Wearable Accelerometer. Sensors,18(5), 1350. doi:10.3390/s18051350

[48] Bortnikov, M., Khan, A., Khattak, A. M., & Ahmad, M. (2019). Accident Recognition via 3D CNNs for Automated Traffic Monitoring in Smart Cities. Advances in Intelligent Systems and Computing Advances in Computer Vision, 256-264. doi:10.1007/978-3-030-17798-0_22

[49] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. & Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks, Nature, 542 (7639), 115-118, doi:10.1038/nature21056

[50] A. H. Fakhrulddin, X. Fei, and H. Li. Convolutional neural networks (cnn) based human fall detection on body sensor networks (bsn) sensor data. In 2017 4th ICSAI, Nov 2017.

[51] Nait Aicha, A., Englebienne, G., van Schooten, K. S., Pijnappels, M., & Kröse, B. (2018). Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry. Sensors (Basel, Switzerland), 18(5), 1–14. https://doi.org/10.3390/s18051654

[52] Lu, N., Wu, Y., Feng, L., & Song, J. (2019). Deep Learning for Fall Detection: Three-Dimensional CNN Combined With LSTM on Video Kinematic Data. IEEE Journal of Biomedical and Health Informatics,23(1), 314-323. doi:10.1109/jbhi.2018.2808281

[53] Casilari, E., Santoyo-Ramón, J., & Cano-García, J. (2017). Analysis of Public Datasets for Wearable Fall Detection Systems. Sensors, 17(7), 1513. doi:10.3390/s17071513

[54] Shieh, W., & Huang, J. (2012). Falling-incident detection and throughput enhancement in a multi-camera video-surveillance system. Medical Engineering & Physics,34(7), 954-963. doi:10.1016/j.medengphy.2011.10.016

[55] Mousse, M. A., Motamed, C., & Ezin, E. C. (2016). Percentage of human-occupied areas for fall detection from two views. The Visual Computer,33(12), 1529-1540. doi:10.1007/s00371-016-1296-y

[56] Zhang, S., Li, Z., Wei, Z., & Wang, S. (2016). An automatic human fall detection approach using RGBD cameras. 2016 5th International Conference on Computer Science and Network Technology (ICCSNT). doi:10.1109/iccsnt.2016.8070265

[57] Hekmat, M., Mousavi, Z., & Aghajan, H. (2016). Multi-view Feature Fusion for Activity Classification. Proceedings of the 10th International Conference on Distributed Smart Camera - ICDSC 16. doi:10.1145/2967413.2967434

[58] Su, S., Wu, S., Chen, S., Duh, D., & Li, S. (2015). Multi-view fall detection based on spatio-temporal interest points. Multimedia Tools and Applications,75(14), 8469-8492. doi:10.1007/s11042-015-2766-3

[59] Kong, Y., Huang, J., Huang, S., Wei, Z., & Wang, S. (2019). Learning spatiotemporal representations for human fall detection in surveillance video. Journal of Visual Communication and Image Representation,59, 215-230. doi:10.1016/j.jvcir.2019.01.024

[60] Koshmak, G., Loutfi, A., & Linden, M. (2016). Challenges and Issues in Multisensor Fusion Approach for Fall Detection: Review Paper. Journal of Sensors,2016, 1-12. doi:10.1155/2016/6931789

[61] Wu, Y., Su, Y., Hu, Y., Yu, N., & Feng, R. (2019). A Multi-sensor Fall Detection System Based on Multivariate Statistical Process Analysis. Journal of Medical and Biological Engineering, 39(3), 336–351. https://doi.org/10.1007/s40846-018-0404-z

[62] Wu, Y., Su, Y., Hu, Y., Yu, N., & Feng, R. (2019). A Multi-sensor Fall Detection System Based on Multivariate Statistical Process Analysis. Journal of Medical and Biological Engineering, 39(3), 336–351. https://doi.org/10.1007/s40846-018-0404-z

[63] Mubashir, M., Shao, L., & Seed, L. (2013). A survey on fall detection: Principles and approaches. Neurocomputing, 100, 144–152. https://doi.org/10.1016/j.neucom.2011.09.037

[64] Dong, Z., Li, F., Ying, J., & Pahlavan, K. (2018). Indoor motion detection using Wi-Fi channel state information in flat floor environments versus in staircase environments. Sensors (Switzerland), 18(7). https://doi.org/10.3390/s18072177

[65] Mao, A., Ma, X., He, Y., & Luo, J. (2017). Highly portable, sensor-based system for human fall monitoring. Sensors (Switzerland), 17(9). https://doi.org/10.3390/s17092096

[66] Zhang, Z., Conly, C., & Athitsos, V. (2014). Evaluating Depth-Based Computer Vision Methods for Fall Detection under Occlusions. 196–207. https://doi.org/10.1007/978-3-319-14364-4_19

[67] Zhang, Z., Conly, C., & Athitsos, V. (2015). A survey on vision-based fall detection. Proceedings of the 8th ACM international conference on Pervasive technologies related to assistive environments. ACM, 2015. http://dx.doi.org/10.1145/2769493.2769540

[68] Banos, O., Galvez, J. M., Damas, M., Pomares, H., & Rojas, I. (2014). Window Size Impact in Human Activity Recognition. Sensors, 14(4), 6474–6499. doi: 10.3390/s140406474

[69] Khalid, S.; Khalil, T.; Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In Proceedings of the Science and Information Conference (SAI), London, UK, 27–29 August 2014.

[70] Hsieh, Y. Z., & Jeng, Y. L. (2018). Development of Home Intelligent Fall Detection IoT System Based on Feedback Optical Flow Convolutional Neural Network. IEEE Access, 6, 6048–6057. doi: 10.1109/access.2017.2771389

[71] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Spatial pyramid pooling in deep convolutional networks for visual recognition,’’ in Proc. 13th Eur. Conf. Comput. Vis., 2014, pp. 346–361.

[72] Akula, N. V. A., Shah, A. K., & Ghosh, R. (2018). A spatio-temporal deep learning approach for human action recognition in infrared videos. Optics and Photonics for Information Processing XII. doi: 10.1117/12.2502993

[73] Sucerquia, A., López, J. D., & Vargas-Bonilla, F. (2018). Real-Life/Real-Time Elderly Fall Detection with a Triaxial Accelerometer. doi: 10.20944/preprints201711.0087.v3

12. Artículo publicado

A Vision-Based Approach for Fall Detection Using Multiple Cameras and Convolutional Neural Networks: A Case Study Using the UP-Fall Detection Dataset

Ricardo Espinosa[a], Hiram Ponce[b,*], Sebastián Gutiérrez[a], Lourdes Martínez-Villaseñor[b], Jorge Brieva[b], Ernesto Moya-Albor[b]

[a]Universidad Panamericana Aguascalientes. Facultad de Ingeniería.
Rusticos Calpulli 101, Aguascalientes, 20290, México.
respinosa@up.edu.mx, jsgutierrez@up.edu.mx

[b]Universidad Panamericana. Facultad de Ingeniería.
Augusto Rodin 498, Ciudad de México, 03920, México.
hponce@up.edu.mx, lmartine@up.edu.mx, jbrieva@up.edu.mx, emoya@up.edu.mx

Abstract

The automatic recognition of human falls is currently an important topic of research for the computer vision and artificial intelligence communities. In image analysis, it is common to use a vision-based approach for fall detection and classification systems due to the recent exponential increase in the use of cameras. Moreover, deep learning techniques have revolutionized vision-based approaches. These techniques are considered robust and reliable solutions for detection and classification problems, mostly using convolutional neural networks (CNNs). Recently, our research group released a public multimodal dataset for fall detection called the UP-Fall Detection dataset, and studies on modality approaches for fall detection and classification are required. Focusing only on a vision-based approach, in this paper, we present a fall detection system based on a 2D CNN inference method and multiple cameras. This approach analyzes images in fixed time windows and extracts features using an optical flow method that obtains information on the relative motion between two consecutive images. We tested this approach on our public dataset, and the results showed that our proposed multi-vision-based approach detects human falls and achieves an accuracy of 95.64% compared to state-of-the-art methods with a simple CNN network architecture.

Keywords: Human Activity Recognition, Human Fall Detection, Machine Learning, Healthcare, Computer Vision

*Corresponding author

1. Introduction

Human activity recognition (HAR) in the monitoring and tracking of human health is an interesting topic that has recently been growing within the research community, especially in the detection of falls among elderly people. Falls can cause injuries, bodily harm, fractures, etc. In fact, globally, falls are the second leading cause of unintentional injury and injury-related deaths among adults 65 years of age and older [1]. "Approximately 28–35% of people aged 65 and over fall each year, increasing to 32–42% for those over 70 years of age" [2]. Falls frequently cause functional dependencies in the elderly. Additionally, many fall-related deaths result from a long "laying time", which is defined as the extended period of time in which the victim remains immobile on the ground.

There are many types of falls. Oneill et al. [4] divides human falls by direction, namely, forward, backward and to the side. For instance, falls forward are the most common falls, with 38% occurring in men younger than 65 and 62% occurring in men older than 65. Similarly, falls forward occur 62% of the time in women younger than 65 and 60% of the time in women older than 65.

In 1987, the Kellogg International Working Group [3] on the prevention of falls in the elderly defined a fall as an unintentional ground impact or some lower level for reasons other than sustaining a violent blow, loss of consciousness, or the sudden onset of paralysis, as in stroke or epileptic seizure. A human fall typically starts with a short freefall period. This freefall causes the acceleration to significantly decrease below the 1G threshold. This freefall represents the period of time when the actual fall is taking place. The fall must stop, and a fall causes acceleration and a spike in the graph. The amplitude crossing an upper threshold suggests a fall [5].

It has been proven that the medical consequences of a fall are highly contingent upon the response and rescue time. In this sense, fall detection systems can improve the response time of medical professionals and decrease the medical consequences of falls.

Due to the extraordinary advances in and increased research on embedded sensor systems, mobile devices and microelectronics, Internet of Things (IoT) systems allow people to continually interact with technology. Additionally, large amounts of data about a person's daily actions are needed so that fall detection systems can allow for the rapid and appropriate assistance of elderly people.

There are many different types of fall detection systems, including sensor-based, vision-based and multimodal-based systems. For instance, sensor-based approaches make use of ambient, smart devices and wearable sensors to provide important information, such as acceleration, absence/presence of individuals, etc., while vision-based strategies use images, such as 3D reconstructions of the environment, simple 2D RGB video sequences with one or multiple cameras, or depth images acquired from 3D depth sensors, as the main input. Multimodal-based approaches collect all the information possible from cameras, microphones, wearable sensors, ambient sensors, and smart devices, among others, and

they combine all this information to improve the fall detection and classification results in a practical manner.

Analytical and machine learning methods are two main approaches for detecting activities and falls [6]. Analytical methods detect falls using threshold algorithms. For example, when falling, a person hits either the ground or an obstacle. This impact results in an intense reversal of the acceleration in terms of trajectory. This change in directionality can be detected by a threshold value. With these types of methods, the most difficult task is adapting the detection to different types of falls or to different people since thresholds differ by person and/or by the type of fall [6]. To address this problem, there are other strategies, such as pattern normalization [61] and correlation-based algorithms [62], and recent investigations report the use of optimization algorithms to choose the threshold [47].

Furthermore, machine learning methods have been gaining popularity due to their flexibility to different subjects and types of falls [63]. The most well-known supervised learning techniques used for fall detection systems include multi-layer perceptron (MLP) [42], support vector machine (SVM) [39], the hidden Markov model (HMM), decision trees, random forest, k-nearest neighbors (KNN) [41], and the convolutional neural network (CNN) [40], which is a deep learning method.

Deep learning techniques are currently changing and improving the methods used to address computer vision problems. CNNs automatically learn features from training data, thus creating a feasible automatic feature extraction method for images. CNNs have been widely applied in image processing problems; for example, in reference [48], the authors use deep learning to detect accidents using the optical flow as the feature extraction method and then test this method using real videos. In [49], a single CNN was trained by using images to directly classify skin lesions and to detect cancer, and it achieved an area under the receiver operating characteristic curve (AUC) of 0.96%. Moreover, CNNs have been used in fall detection systems with a sensor-based approach with up to 92.3% accuracy [50] and with a wearable approach with an AUC equal to 0.75 [51].

Regarding vision-based approaches for fall recognition systems, deep learning has been successfully applied. For example, Nez-Marcos et al. [19] implemented a CNN to avoid manual feature engineering; the convolutional layers of the system extracted the most important features of the images, and a sensitivity and specificity of 94% was obtained. A CNN for a vision-based approach was also implemented in [52], in which the authors used a 3D CNN with input videos of peoples' kinematics and achieved 100% accuracy when evaluated on different datasets.

Recently, our research group released a public multimodal dataset for fall detection called the UP-Fall Detection dataset [24]. The data were gathered from different sources of information, i.e., wearable sensors, ambient sensors and cameras. Until now, we have studied this dataset using a multimodal approach [24]. However, the technical expertise and skills required for building and setting a multimodal fall detection system make it difficult to implement in the real world. Moreover, wearable and ambient sensors are conditioned by the environment, thus making portability difficult to achieve. Therefore, we are interested

in creating a vision-based fall detection system using this dataset and the video recordings from multiple cameras.

Additionally, fall detection systems based on single RGB cameras are often viewpoint-dependent, according to [67]. This issue raises the need for new datasets when a camera is moved to different viewpoints and, in particular, to different heights. To address this issue, using different camera viewpoints in a dataset can help to identify whether a given method is independent of viewpoint. To this end, a fall detection system must be reliable regardless of the position of the subject while falling with respect to the camera.

Based on the above, this work presents a fall detection system based on a 2D CNN inference method and multiple cameras. As we describe later, this approach analyzes images using fixed time windows and extracts features using an optical flow method that obtains the information of the relative motion between two consecutive images from video recordings acquired from cameras with different viewpoints. We tested this approach with our public UP-Fall Detection dataset, and the results showed that our proposed multi-vision-based approach detects human falls using a simple CNN network architecture, achieving results that are comparable to those of state-of-the-art methods. In addition, the performance of the proposed approach is comparable to the performance achieved using a multimodal approach.

Even though CNN has previously been used in fall detection systems with good performance using a particular dataset, Casilari et al. [53] concluded that these systems should be trained and tested with different datasets due to the different numbers of samples, different types of falls and different time series for any type of fall. Thus, the implementation of CNN in the multicamera vision-based approach, specifically for the UP-Fall Detection dataset, might improve state-of-the-art fall detection systems.

The main contributions of this work are as follows: (i) the usage of multiple cameras with CNN for fall detection and classification, (ii) the implementation of this approach with the UP-Fall Detection database, and (iii) the comparison of the performance of this approach with those of well-known supervised learning methods. To the best of our knowledge, only one study [59] combines a CNN with a multicamera vision-based approach to recognize human falls. In contrast to our proposal, in [59], a voting strategy based on the output results from independent cameras is used; our approach uses the information from all the cameras in the same machine learning model.

The rest of the paper is organized as follows. First, we review and analyze different fall detection systems, focusing mainly on vision-based solutions. Next, we present a description of the UP-Fall Detection dataset. Then, we present the proposal in detail. We also explain the experiment as well as the results and then discuss the proposal. Finally, the conclusions are presented.

2. Fall Detection Systems

HAR and fall detection are difficult tasks, and there are several ways to complete these tasks due to the numerous approaches proposed in the literature. For instance, Lara et al. [8]

and Noury [6] divide the HAR taxonomy into three general approaches, i.e., external, wearable and video sensing, depending on the information source. Using these approaches, there are case studies related to sensor-based [9], vision-based [10], smartphone-based [11], and multimodal-based [12] strategies to address human fall recognition, as described below.

2.1. Sensor-Based Fall Detection Systems

With the increasing accessibility of mobile sensor technology, fall detection systems have been designed for real-world purposes. Human activity can be tracked, monitored and labeled using data from different types of sensors at various locations in the environment and in the human body. An important application of sensor-based approaches is the detection of abnormal activities from wearable sensors [9]. Abnormal activity detection methods can be applied to continuously track the movements of an individual to determine whether the person's activities are abnormal [9]. In [21], using triaxial sensors and SVM as an inference method, the authors achieved 98.33% accuracy. In [65], using acceleration and the Euler angle (yaw, pitch, and roll), the authors achieved 100% accuracy, sensitivity, and specificity. Nevertheless, some disadvantages of heterogeneous sensor networks are because human activities typically involve more than one body part. Moreover, several physiological and biomechanical studies have shown that most daily human activities are inherently multimodal [13]. Thus, different types of sensors are required for data collection.

2.2. Wearable Fall Detection Systems

Wearable approaches are common solutions for fall detection, as they take advantage of the low cost, online tracking capability and small sizes of wearable technologies. For example, in [46], a Shimmer device was used for acquisition and transition data. The wearable device was placed on the chest and obtained 98.8% accuracy using different machine learning models. In [47], the authors used a wearable band placed on the wrist and achieved 0.95 specificity and 0.83 sensitivity using threshold-based peak recognition with SVM for classification; they optimized the threshold values for different datasets.

In wearables and smartphones, energy consumption is a difficult problem to solve since these devices need to be continuously worn to obtain tracking information from subjects. The lifetimes of wearables and smartphones are limited to the capacity of the battery, and constant recharging is necessary, which prevents the constant tracking of the patient's activities [46].

2.3. Smartphone-Based Fall Detection Systems

Currently, smartphones contain multiples integrated sensors and a large processing capacity, which has grown over the years. Smartphones can measure the movements of a controller in a nonintrusive way. Smartphone-based fall detection systems use smartphone sensors, e.g., gyroscopes, triaxial accelerometers, and altimeters, to collect data over long periods of time. Case studies using this approach can be found in [43], in which the authors use a smartphone-based triaxial accelerometer with statistical time-domain features. Then, the authors applied principal component analysis for feature selection and inferred the

outputs with MLP, obtaining an accuracy of 92%. Another example is the work of Vilarinho et al. [44], which combined smartphone and smartwatch sensors and used threshold-based techniques and pattern recognition algorithms to recognize falls with an accuracy of 63% and daily activities with an accuracy of 78%.

2.4. Multimodal-Based Fall Detection Systems

Data acquisition, mainly from ambient sensors, wearables, cameras, microphones, and radio frequency identification (RFID) tags, among others, is an important task of fall detection and classification systems. Wearables are not able to distinguish a large number of fine-grained and/or complex human activities, as they have difficulty differentiating between similar activities; thus, ambient sensors are needed for context awareness. In this regard, multimodal-based approaches can combine more than one source of data to obtain information about both the environment and the user. These approaches make fall detection and classification feasible by leveraging different modes of sensing from a wide range of sources [12].

Because multimodal approaches comprise many different sources of data from subjects and environments, there are some weaknesses, as reported in [60]: (i) many information types require the application of robust feature extraction and feature selection techniques and considerations regarding machine learning approaches for different types of input data, making fall detection computationally expensive and difficult to perform, and (ii) multiple sensors with complex placement on the body (and in the environment) could lead to high costs, practical deployment difficulties, and obtrusiveness, especially for elderly people.

2.5. Vision-Based Fall Detection Systems

This work mainly focuses on vision-based methods. Traditionally, fall detection systems have been implemented by using computer vision and image processing techniques to classify activities. With recent advancements, in-depth, noninvasive imaging sensors produce high-quality images. This information is also analyzed for human tracking, monitoring and user recognition systems [14, 15, 16] and for monitoring and recognizing the daily activities of subjects in indoor environments [17].

Most of the vision-based approaches have used simple RGB cameras, web cameras, motion camera systems, or even Kinect [18]. The use of Kinect for fall detection has increased because it can obtain 3D information such as human poses or limb positions [18].

The classic vision-based fall detection and classification strategies consist of four phases [4]: (1) data acquisition from video sequences, (2) feature extraction from images, (3) feature selection and (4) learning and inference. Multiple machine learning techniques are used in the literature, such as SVM [35] or random forest [36]. Zerrouki et al. [17] proposed a fall detection system based on human silhouette variations in vision monitoring and SVM to identify postures. Then, these authors used HMM to classify the data into fall and non-fall events. Rougier et al. [7] tracked the person's silhouette along with the video sequences. Shape deformation was then quantified from these silhouettes based on shape

analysis methods. Finally, falls were detected from daily activities using Gaussian mixture models (GMMs).

Vision-based systems can be divided into two categories: monocular systems and multicamera systems. Monocular-based fall detection systems depend on one viewpoint. Moving a camera to different viewpoints requires collecting new training data for that specific viewpoint and calibrating the camera sensor. However, these systems can fail because of occluding objects between the target and camera. Zhang et al. [66] proposed using multiple Kinect devices to solve that problem, as their OCCU dataset included both occluded and nonoccluded falls. Kwolek et al. [21] extracted depth maps about the environment and the person's silhouette in combination with 3-axis accelerometers and SVM as a machine learning technique. In terms of multicamera fall detection systems, Thome and Miguet [22] proposed using an HMM to distinguish falls from walking activities. The features extracted for the motion analysis were obtained from a metric image rectification in each view. Anderson et al. [23] analyzed the states of 3D objects, called the voxel of a person, obtained from two cameras. All these works construct 3D models with multiple cameras to reconstruct the environment. This task is particularly difficult because the cameras need to be calibrated to correctly compute the 3D information, which presents issues regarding the synchronization of the video sequences of each camera, making it more difficult to implement than a monocular-based approach.

Thus, from the point of view of the deployment of these systems, 2D multiple cameras are usually a good option, mainly because of their low cost and ease of implementation. It is also important to emphasize that cameras are already installed in many public places, such as airports, shops, and elderly care centers, and these cameras can also be used for fall detection systems. Thus, 2D cameras are relevant for the fall detection application domain.

Multiple studies use CNN on monocular vision-based fall detection systems with excellent results [37, 50, 51, 52]. Moreover, several works use a multicamera approach with different classical machine learning models or other algorithms [54, 55, 56, 57, 58], and only one work uses a multicamera approach and CNN in a fall detection system [59].

2.6. Vision-Based Fall Detection Systems using CNN

Recent works on fall recognition systems have taken advantage of the success of deep learning for recognition and classification tasks using regular images, deep images, infrared images, etc. Deep learning CNN searches for the relevant features in images, avoiding the feature engineering tasks and providing versatile automatic feature extraction, depending on the architecture of its convolutional and inference layers [25].

Some recent works on fall detection systems are reported in reference [70], in which the authors use rule-based filters before an input convolutional layer, combining the convolutional layer output with optical flow features to choose a good input for the inference phase of its 3D CNN architecture; these method achieved 92.67% accuracy. In [72], the authors use infrared (IR) images and a 3D CNN to find features on three color channels in real situations, taking into consideration the spatiotemporal image information; this method achieved an 85% accuracy on test video sequences.

3. Dataset Description

In this research, we use a public dataset called UP-Fall Detection [24]. This dataset was made using 17 healthy subjects without any impairments (9 males and 8 females) ranging from 18–24 years of age, with a mean height of 1.66 ± 0.0530 m and a mean weight of 66.8 ± 12.1182 kg; the subjects performed 11 activities and 3 trials per activity, including six simple human daily activities and five different types of human falls using a multimodal approach, with wearable sensors, ambient sensors, and vision devices. The consolidated dataset and the feature dataset are publicly available.

The activities and falls stored in this dataset are summarized in Table 1. All data were collected using the following 14 devices: five Mbientlab MetaSensor2 wearable sensors that collected the raw data from a 3-axis accelerometer, a 3-axis gyroscope, and an ambient light sensor; one electroencephalograph (EEG) NeuroSky MindWave headset was used to measure the raw brainwave signal from its unique EEG channel sensor located at the forehead; as context-aware sensors, we installed six infrared sensors as a grid 0.40 m above the floor of the room to measure the changes in interruption of the optical devices (where 0 means interruption and 1 means no interruption); and two Microsoft Life-Cam Cinema cameras, one for a lateral view and the other for a frontal view in relation to the subject, were located at 1.82 m above the floor, which is higher than the mean height of the subjects. The falls were performed from right to left according to the viewpoint of Camera 1 (lateral view). All experiments were recorded by positioning the subject at the center of the view in both cameras and at the same distance from both cameras. That is, Camera 1 and Camera 2 were 2.10 m and 1.90 m away from the center point of the mattress, respectively. Additionally, all images contain at most one subject, so multiple people were not simultaneously recorded for this dataset. All these devices were located as shown in Figure 1. For more information about the UP-Fall Detection dataset, see reference [24].

In this work, we use only the information from two cameras, which run at 18 fps, from the dataset, taking advantage of the multiple camera distributions.

Here, we aim to implement a fall detection and classification system using multiple cameras and to compare its performance when using only a monocular-based approach. To this end, CNN will be used as the classification model.

4. Description of the Proposal

In this work, we adopted the traditional workflow for fall detection systems [8], which consists of the following steps: (i) data collection, (ii) windowing, (iii) feature extraction, and (iv) learning and inference. The workflow is shown in Figure 2.

Table 1: Activities performed by subjects, adapted from [24].

Activity ID	Description	Duration (s)	Abbreviation
1	Falling forward using hands	10	FH

2	Falling forward using knees	10	FF
3	Falling backward	10	FB
4	Falling sideward	10	FS
5	Falling while attempting to sit sitting in an empty chair	10	FE
6	Walking	60	W
7	Standing	60	S
8	Sitting	60	ST
9	Picking up an object	10	P
10	Jumping	30	J
11	Laying	60	L

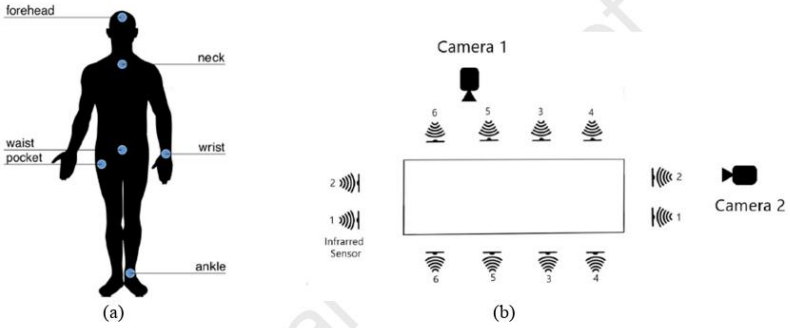


Figure 1: Distribution of the sensors. (a) Wearable sensors and EEG helmet on the human body. (b) Layout of ambient sensors and multiple cameras. Adapted from [24].

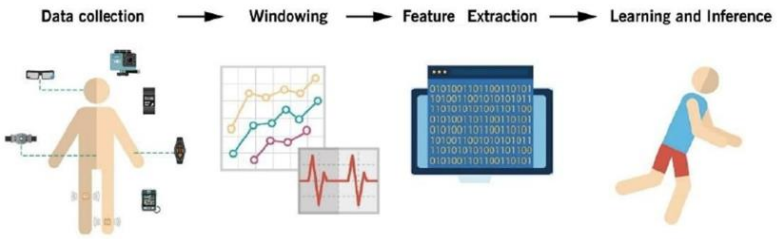


Figure 2: Traditional workflow for fall detection systems.

4.1. Data Collection

One of the most challenging phases in the traditional workflow for fall detection systems and machine learning problems in general is data collection. Recently, to be correctly trained and tested, deep learning techniques require large amounts of data and other factors, such as the number of individuals, the physical characteristics of the individuals, and the number of individuals with diverse characteristics in terms of gender, age, height, weight, and health conditions [8].

As we explained in the Dataset Description section, we use the UP-Fall detection dataset. To summarize, the dataset contains information related to 17 young subjects performing 11 different activities, including 5 falls and 6 other activities. For this work, we use the information gathered from two RGB cameras positioned at different viewpoints (lateral and front views) [24].

4.2. Windowing

The windowing approaches in fall detection systems are normally used to segment the time series of performed falls. Segmentation is the process of dividing the sensor signals into smaller data segments. This process has been performed in different ways in the activity recognition field and fall detection systems. Segmentation techniques can be categorized into three main groups: activity-defined windows, event-defined windows and sliding windows [68].

We adopted a sliding window approach to capture the temporal dependency between samples. In this case, we divided all data into fixed-length time windows for each activity. Our implementation uses 1-second widows with 0.5 seconds of overlap due to the results reported in [24]. In that work, we experimented with 1-second, 2-second and 3-second windows with 50% overlap by applying classic machine learning methods; the best performance was achieved with the 1-second window. In this work, we employ this performance evaluation as the baseline of our multiple-camera-based proposal. To this end, we are able to analyze the fall on each window in a simple way. The result of this step was multiple 1-second window length series of images to be processed in the next step.

4.3. Feature Extraction

Feature extraction is a general method in which one tries to transform the input space into a low-dimensional subspace that preserves most of the relevant information to improve data analysis [69]. In fall detection systems, there are multiple techniques used to extract relevant information depending on the data type; in vision-based approaches, the optical flow algorithm provides very rich information about the apparent movements in images, and these approaches have been used in multiple studies that combine CNN and optical flow to extract features from images [70, 37].

For feature extraction, each window is preprocessed to obtain sufficient information for describing the activity. In this case study, the optical flow algorithm [24, 26] was used for the visual features extracted from each camera. This algorithm helps us to obtain the apparent displacements between two windows and, in doing so, to distinguish movements and directions without considering the static features in the image. The obtained features

are the horizontal and vertical relative movements of the pixels in the images, i.e., U and V, respectively [26]. The resultant combination, D, of these values corresponds to the magnitude of the relative movement, as shown in (1), where the resultant matrix images are the same size of the original window images, i.e., 640×480 pixels in our case study.

$$D_{i,j} = \sqrt{U_{i,j}^2 + V_{i,j}^2} \tag{1}$$

4.4. Learning and Inference

In reference [6], there are multiple ways to achieve this phase; two of these methods are machine learning and deep learning, which have recently been used. This step searches, trains and tests the output from feature engineering to classify or predict the fall with inputs from environment sensors, wearable sensors, and, in this work, from the multicamera vision-based approach.

In deep learning, CNNs have revolutionized the way computer vision problems are addressed due to the automatic discovery of structure representations in large datasets. This method has dramatically improved the state-of-the-art methods in image processing [25].

However, finding a suitable CNN architecture is difficult [25]. To address this difficulty, the literature has reported multiple network architectures, depending on the problem to be solved. For example, in recent years, many network structures for image recognition and classification problems have been reported, such as AlexNet [27], ClarifaiNet [28], GoogLeNet [29] and VGGNet [30]. All these networks have proved to be efficient in their own problem domains, and they can also be used as pretrained models so that users can reduce the amount of time needed to retrain them for another task. However, these architectures are complex and can possibly be improved.

In this work, we design a CNN with three convolutional layers and three 2D max-pooling layers for feature extraction and three fully connected layers for fall detection. To this end, we fixed all images to 38 × 51 pixels; three fully connected layers were fixed and chosen because, for fully connected layers, the fixed-size constraint comes from only the fully connected layers, which exist at a deeper stage of the network [71]. The CNN receives the magnitudes D calculated from U and V, which were converted to grayscale images with 38×51 pixels, representing the optical flow features extracted. Then, these images go to the input layer, which consists of 128 convolution filters with a kernel size of 3×3. The second convolutional layer has 128 filters and the same kernel size, and the third layer has 64 convolutional filters and the same kernel. This convolutional layer architecture was selected by cross-validation and using the F1-score metric, as shown in Table 2. After each convolutional layer, 2D max-pooling layers are employed to synthetize output convolutions. Then, these results are input to three fully connected layers, i.e., 64 rectified linear units (ReLU) in the first layer, 128 ReLU in the second, 254 ReLU in the third and, finally, a 2D SoftMax layer with one output. The latter is employed to perform fall detection using fall (1) and no-fall (0) classes. Figure 3 shows the representation of the proposed CNN.

As described above, the UP-Fall Detection dataset is integrated by information from 17 subjects performing 11 different activities/falls, with three trials for each activity. To train the CNN, we divide data collection into trials 1 and 2 for each activity, and we use the subject as the training set (67%) and trial 3 for each activity and subject as the testing set (33%). The training dataset included 42,000 grayscale images that were 38×51 in size, and the optical flow was used for preprocessing; the testing dataset included 21,000 grayscale images with the same preprocessing flow. We trained during 50 epochs using the Adam optimizer and binary cross-entropy loss function, as defined in (2), where p is the prediction of the network, and t is the ground truth.

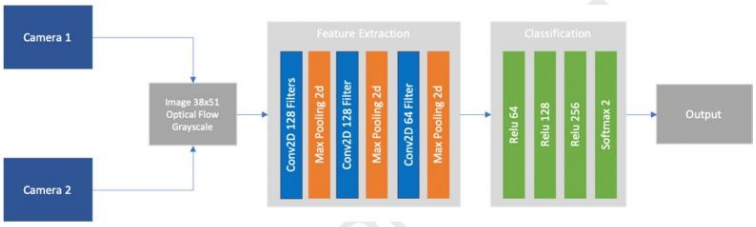


Figure 3: Our proposal of the CNN architecture for the multicamera vision-based fall detection system.

Table 2: Cross-validation for the convolutional architecture layers.

CNN Architecture	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
64 64 64	95.40	86.28	83.76	97.54	95.40
64 64 128	95.27	88.26	80.34	98.03	84.11
64 64 256	94.90	83.67	83.57	96.99	83.62
64 128 64	94.62	82.36	83.27	96.71	82.81
64 128 128	94.66	86.49	77.83	97.76	81.94
64 128 256	95.15	85.74	82.35	97.46	84.14
64 256 64	94.32	85.86	76.00	97.69	80.63
64 256 128	94.92	86.45	79.91	97.69	83.05
64 256 256	94.90	91.18	74.48	98.67	81.98
128 64 64	95.17	86.21	82.11	97.58	84.11
128 64 128	94.80	96.02	97.89	78.02	96.95
128 64 256	94.79	97.07	96.74	84.18	96.91
128 128 64	95.64	96.91	97.95	83.08	97.43
128 128 128	95.44	96.19	98.49	78.87	97.33
128 128 256	95.05	97.88	96.22	88.70	97.04
128 256 64	94.28	96.32	96.92	79.91	96.62
128 256 128	94.51	97.00	96.47	83.82	96.74
128 256 256	95.19	96.84	97.48	82.78	97.16
256 64 64	94.81	96.16	97.76	78.81	96.95
256 64 128	94.26	96.25	96.97	79.54	96.61
256 64 256	94.38	96.34	97.03	80.03	96.68
256 128 64	94.75	96.19	97.64	79.05	96.91
256 128 128	94.72	97.63	96.08	87.36	96.85
256 128 256	94.40	96.66	96.71	81.86	96.68

256 256 64	94.10	96.31	96.71	79.91	96.51
256 256 128	94.57	96.36	97.24	80.09	96.80
256 256 256	94.09	96.19	96.83	79.18	96.51

$$loss(p,t) = -(t \cdot \log p + (1-t) \cdot \log(1-p)) \quad (2)$$

5. Experimentation

To analyze our proposal, the following experiments were carried out: (i) experiments to test our CNN model and to compare it with classic machine learning methods, such as SVM, random forest (RF), MPL and KNN; (ii) experiments to compare monocular with multicamera vision-based fall detection system approaches; and (iii) tests of our proposal not only for detection but also for the classification of activities and falls using the multicamera vision-based approach.

In these experiments, we used training and testing datasets with information provided from two cameras. We used the information of one camera per model and then the information from both lateral and front viewpoint cameras at the same time [8]. For windowing, 1-second windows with 0.5-second overlaps were employed. The images were treated as grayscale, and the optical flow implementation was treated as feature extraction. We resized the images to 38×51 pixels, and we performed a benchmark comparison between the classic machine learning methods (i.e., SVM, MLP, RF and KNN) and the CNN depicted in Figure 3.

These experiments aim to explore and compare the performance of a monocular vision-based approach with multicamera vision-based fall detection systems and to create a benchmark comparison of classic machine learning methods and CNN for fall detection using the latter approach.

To evaluate the performance of our work, we use the following five metrics: accuracy, sensitivity, specificity, precision, and F1-score, as given by (3)–(7), where *TP* refers to true positives, *TN* to true negatives, *FP* to false positives, and *FN* to false negatives [32].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$precision = \frac{TP}{TN + FP} \quad (2)$$

$$sensitivity = \frac{TP}{TP + FP} \quad (5)$$

$$specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F_1 - score = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity} \tag{7}$$

All experiments were conducted in Python 3.7.3 using the sklearn3 framework for classic machine learning techniques and the keras4 framework for CNN, taking advantage of its GPUs management capabilities [31].

5.1. Results and Discussion

The experimental results are described in this section. Then, a discussion based on the analysis is presented.

5.1.1. Fall Detection Using Conventional Machine Learning Models

First, we conducted an experiment using the optical flow-based features from both cameras at the same time (Cam1 and Cam2). We trained four conventional machine learning models, namely, SVM, RF, MLP and KNN, as described above. Table 3 shows the meta-parameter settings for these models. For this experiment, we built the models using 67% for training and 33% for testing data. Table 4 summarizes the performance results using the visual features extracted in 1-second windows with 0.5 seconds of overlap.

From Table 4, it can be observed that the conventional machine learning models cannot predict human falls well in terms of accuracy, precision, sensitivity, specificity and F1-score. Currently, KNN seems to have the best performance based on the F1-score metric (15.27%). In terms of accuracy, SVM performs the best, with an accuracy of 32.40%. These machine learning models achieved an averaged accuracy of 29.77%. From the results, we might assume that the conventional machine learning methods using windowing and the sklearn library, as explained above, are not robust enough. To improve the performance, we implemented CNN, as described below.

Table 3: Parameter settings used to train the classification models.

Classifier	Parameters
SVM	kernel = “radial basis function” kernel coefficient =1 c=1 shrinking = 1 tolerance = 0.001

RF	minimum samples split = 2 minimum samples leaf = 1 estimators = 2 bootstrap = 1
MLP	activation function = "reLU" hidden layers = 100 penalty parameter = 0.0001 batch size = min(200, num_samples) shuffle = 1 initial learning rate = 0.001 tolerance = 0.0001 exponential decay(first moment) = 0.9 exponential decay(second moment) = 0.999 regularization coefficient = 0.000000001 solver = "stochastic gradient" maximum epochs = 10
KNN	neighbors = 5 leaf size = 30 distance metric = "euclidean"

Table 4: Performance obtained by the classic ML models.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
SVM	32.40	14.03	14.10	90.03	14.06
RF	29.30	14.45	14.30	91.26	14.37
MPL	30.08	9.05	11.03	93.65	9.94
KNN	27.30	16.32	14.35	90.96	15.27

Table 5: Performance of the CNN models using the lateral view, front view and both views.

Data	Method	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
(Cam1) Lateral view	Proposed CNN	95.24	95.24	97.72	81.58	97.20
(Cam2) Frontal view	Proposed CNN	94.78	96.30	97.57	79.67	96.93
(Cam1 & Cam 2)	Proposed CNN	95.64	96.91	97.95	83.08	97.43
(Cam1 & Cam 2)	VGG-16 CNN	84.44	84.44	100	0	91.56

5.1.2. Fall Detection Using CNN

In this experiment, we trained the following three CNN models: (i) a CNN model using visual features from the lateral view (Cam1), (ii) a CNN model using visual features from

the front view (Cam2), and (iii) a CNN model using visual features from both cameras at the same time.

A summary of the results is shown in Table 5. As can be observed, the performances are very similar. The lateral view (Cam1) is slightly better than the frontal view, as expected [7]. However, Cam2 shows less specificity (79.67%) than Cam1 (81.58%), which could lead to misclassification. Furthermore, the combination of both views maintains the output performance of the lateral view, which is important because, if some of the cameras are occluded, it will remain feasible to detect the fall with just one camera, as supported in reference [7]. Furthermore, we performed an experiment using the VGG-16 CNN architecture with images from 2 cameras (frontal and lateral); the results shown in Table 5 indicate that our proposal has significantly better performance than the VGG-16 CNN architecture using UP-Fall. Thus, we conclude that our multicamera vision-based fall detection system has acceptable performance, in contrast with the conventional machine learning models and using the VGG-16 CNN architecture; additionally, our system avoids the occlusion problem as long as we do not lose sight of the subject.

We also compared our proposed method to other multiple-camera vision-based fall detection systems reported in the literature [15], [35], [37], considering that the latter were implemented using conventional machine learning methods.

For this comparison, we used the multicamera vision-based database, called the Multicam dataset [34]. This dataset includes 24 performances; 22 trials have at least one human fall, and the remaining two contain confounding events. Each performance was recorded from 8 different views. The same stage is used for all videos, with some furniture reallocation [34]. To train our proposal method, we selected two viewpoints (lateral and frontal views) from this dataset, and divide the data into training (67%) and testing (33%) sets. Table 6 summarizes the performance results in terms of sensitivity and specificity, as reported in the literature [33, 35, 37].

Table 6: Comparison between our proposal and state-of-the-art multicamera vision-based fall detection systems reported in the literature, using the Multicam dataset.

Proposal	Method	Sensitivity (%)	Specificity (%)	Cameras
Wang et al. [33]	SVM	89.20	90.30	8
Wang et al. [35]	SVM	93.70	92.00	8
Núñez et al. [37]	VGG-16 CNN	99.00	96.00	8
Ours (Combined)	CNN	97.95	83.08	2

As shown in Table 6, our proposed method is competitive with state-of-the-art approaches, mainly in terms of sensitivity. In addition, our method can handle fall detection using two cameras, in contrast to the eight cameras utilized in the other approaches. Moreover, the network architecture of our proposal (Figure 3) is very simple compared to those of other

works. For example, Núñez-Marcos in [37] used a VGG-16 architecture modified to receive inputs, the authors in [33] used PCA to extract features and SVM for classification, and in [35], the authors presented a multivariate exponentially weighted moving average (MEWMA) and SVM with 2 steps for classification (see Table 6). In that sense, our system has good performance, considering that it requires much less time for training and the simplicity of its architecture.

5.1.3. Daily Activities and Fall Classification Using CNN

Finally, we conducted an experiment for daily activity and fall classification using our proposed method. In this case, each activity and type of fall recorded in the UP-Fall Detection dataset was considered, and so the CNN was converted into a multiclass classifier, as shown in Table 1.

We applied our proposed method using both cameras (Cam1 and Cam2) and the results, compared to the performance obtained in [24] using the same dataset, are depicted in Table 7. As shown, our proposed method is slightly inferior to the multimodal-based approach presented by Martínez-Villaseñor et al. [24], which is an expected result, since a multimodal approach (i.e., wearable sensors, EEG helmet and cameras) is better than a single modality approach such as ours. It is also important to note that the F1-scores of both approaches are similar, i.e., 72.94% for our proposal and 72.80% for the multimodal approach. From the results presented in Table 7, the performance obtained by our proposal can be considered competitive (e.g., similar F1-score), easier to implement (i.e., due to the number of sensors) and less obtrusive (i.e., wearable sensors) than the multimodal-based approach reported in [24].

Table 7: Comparison between our proposal and the multimodal approach reported using the UP-Fall dataset.

Data	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Ours	82.26	74.25	71.67	77.48	72.94
Martínez-Villaseñor, et al. [24]	95.00	77.70	69.90	99.50	72.80

6. Discussion

The proposed multicamera vision-based fall detection and classification system is comparable to state-of-the-art methods. The results suggest the predictive power of our proposed fall detection system (97.43% of F1-score), which outperforms conventional machine learning methods (SVM, RF, MLP and KNN) by using optical flow-based features and fewer cameras and achieves similar performance to the state-of-the-art methods reported in the literature (97.00% sensitivity and 80.00% specificity) and to a multimodal approach (72.80% F1-score).

The advantages of our proposal are as follows. Multicamera approaches offer robust solutions for fall recognition, even when occlusion occurs from one viewpoint, as long as one camera remains focused on the subject. This result can be observed in Table 5, which reports similar performance when using one camera or the other (lateral or frontal view) or both. Additionally, our proposal offers a simple CNN architecture (Figure 3) and a low computational cost. Due to the vision-based nature of our approach, privacy must be discussed due to the nature of constant video surveillance. Our work avoids privacy concerns by analyzing only the relevant information about the fall using the optical flow information calculated from the video sequence. Therefore, the privacy of the person is not affected because the data used to recognize a fall do not contain personal information.

It is also important to consider some limitations of our proposed method. A vision-based approach always depends on the quality of the captured image, the position of the cameras, and the presence of the subject. In addition, privacy issues should be addressed before the implementation. If privacy issues are a limitation, as mentioned before, then the original images captured by the cameras should not be stored; they can be used only for extracting the optical flow features. However, pervasiveness remains an important drawback because cameras are continually acquiring videos of the subjects. In addition, the computational complexity in terms of memory and processing time should be emphasized, as this complexity hinders the scalability of a real-time fall detection system [8].

Regarding the fall and activity samples in the UP-Fall dataset, 42,958 training samples and 21,038 testing samples arranged in 1-second windows were employed in our experiments. The results were competitive with those of the state-of-the-art methods for both detection (Table 5 and Table 6) and classification (Table 7) tasks. Furthermore, it is important to discuss the age of the subjects that performed the falls and activities when building the UP-Fall Detection dataset used in this work. This dataset was made using the information of 17 healthy subjects without impairments (9 males and 8 females) ranging from 18–24 years old. Nevertheless, in [73], it is shown that testing with a dataset built using young people does not significantly deviate from that with a dataset built using elderly people. Thus, we believe that our approach can be applied in real situations, which should be considered in future work.

In vision-based problems, the positioning of the cameras is difficult to determine in terms of the angle, height and distance between the cameras and the subject that performs the activities or falls. The UP-Fall dataset was made by recording falls and activities with fixed distances and angles and recording falls in the same direction. These aspects of the dataset could change under realistic conditions, making our approach difficult to replicate in the real world. However, in our work, we address this problem by using only the apparent movement in the image by applying the optical flow as the feature extraction method and then calculating the Euclidean distance as the apparent movement in each 1-second window. This approach helps us to precisely detect a fall, even when the subject is not positioned at the center of the images. Thus, this proposal might also be considered if the distances and angles between the cameras and the subject are different from the specifications of this dataset. As our CNN architecture receives only apparent movements as input, the angle, height and distance condition results are irrelevant to an inference of the detection only if these values change slightly. Otherwise, retraining is required. In future

work, transfer learning will be applied using our proposed CNN model to analyze changes in camera positions and conditions.

The experimental results showed that our proposal is competitive with state-of-the-art multicamera vision-based approaches for detection systems and for classification (Table 7), even compared to a multimodal approach, such as that reported in [24].

7. Conclusions

In this paper, we presented a multicamera vision-based fall detection and classification system that takes advantage of CNN. In addition, we combined the CNN models with visual features extracted from sequences of images using the optical flow method. In this work, we used the UP-Fall Detection dataset as a case study. We conducted experiments to compare our proposal with conventional machine learning models, analyzed the performance of our proposal for vision-based approaches with single and multiple cameras, and extended our model for fall classification.

From the experimental results, we conclude that our proposed multicamera vision-based fall detection and classification system outperforms conventional machine learning methods, reduces computation time due to the simple CNN architecture, and is competitive with state-of-the-art and multimodal-based approaches.

Future works should implement this approach in a real-world assisted living system and analyze and propose improvements to issues related to privacy, pervasiveness, changes in environmental conditions and occlusion. In addition, we will consider testing our system in a real situation by including the transfer learning approach and different camera positions.

Acknowledgments

This research was funded by Universidad Panamericana through the grant "Fomento a la Investigación UP 2018", under project code UP-CI- 2018-ING-MX-04.

Conflict of Interest

The authors have nothing to declare.

References

[1] Department of Health and Human Services. Fatalities and injuries from falls among older adults - United States, 1993-2003 and 2001- 2005. pages 12211224, November 2006. Morbidity and Mortality Weekly Re- port.
[2] Schneider, M. (2011). Introduction to public health. Sudbury, MA: Jones and Bartlett.
[3] Lord, S. R., Sherrington, C., Menz, H. B., & Close, J. C. (n.d.). Strategies for prevention. Falls in Older People,173-176. doi:10.1017/cbo9780511722233.011
[4] Oneill, T. W., Varlow, J., Silman, A. J., Reeve, J., Reid, D. M., Todd, C., & Woolf, A. D. (1994). Age and sex influences on fall characteristics. Annals of the Rheumatic Diseases,53(11), 773-775. doi:10.1136/ard.53.11.773

[5] Bourke, A., & Lyons, G. (2008). A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. Medical Engineering & Physics,30(1), 84-90. doi:10.1016/j.medengphy.2006.12.001

[6] Noury, N., Fleury, A., Rumeau, P., Bourke, A., Laighin, G. O., Ri- alle, V., & Lundy, J. (2007). Fall detection – Principles and Methods. 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Doi:10.1109/iembs.2007.4352627

[7] Rougier, C., Meunier, J., St-Arnaud, A., & Rousseau, J. (2011). Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. IEEE Transactions on Circuits and Systems for Video Technol- ogy,21(5), 611-622. doi:10.1109/tcsvt.2011.2129370

[8] Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. 2013, 15, 11921209.

[9] Yin, J., Yang, Q., & Pan, J. (2008). Sensor-Based Abnormal Human- Activity Detection. IEEE Transactions on Knowledge and Data Engineering,20(8), 1082-1090. doi:10.1109/tkde.2007.1042

[10] Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., & Qiu, Y. (2013). Exploring Techniques for Vision Based Human Activity Recognition: Methods, Systems, and Evaluation. Sensors, 13(2), 1635-1650. doi:10.3390/s130201635

[11] Dungkaew, T., Suksawatchon, J., & Suksawatchon, U. (2017). Impersonal smartphone-based activity recognition using the accelerometer sensory data. 2017 2nd International Conference on Information Technology (INCIT). doi:10.1109/incit.2017.8257856

[12] Bharti, P. (2017). Complex activity recognition with multi-modal multi- positional body sensing. Journal of Biometrics & Biostatistics,08(05). doi:10.4172/2155-6180-c1-005

[13] Chetty, G., White, M., Singh, M., & Mishra, A. (2014). Multimodal activity recognition based on automatic feature discovery. 2014 Inter- national Conference on Computing for Sustainable Global Development (INDIACom). doi:10.1109/indiacom.2014.6828039

[14] Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. 2008, 18, 14731488.

[15] Raty, T.D. Survey on contemporary remote surveillance systems for public safety. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 2010, 40, 493515.

[16] Albanese, M.; Chellappa, R.; Moscato, V.; Picariello, A.; Subrahmanian, V.S.; Turaga, P.; Udrea, O.A constrained probabilistic petri net framework for human activity detection in video. IEEE Trans. Multimed. 2008, 10, 14291443.

[17] Zerrouki, N., & Houacine, A. (2017). Combined curvelets and hidden Markov models for human fall detection. Multimedia Tools and Appli- cations, 77(5), 6405-6424. doi:10.1007/s11042-017-4549-5

[18] E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier, Fall detection with multiple cameras: An occlusion resistant method based on 3-D silhouette vertical distribution, IEEE Transactions on In- formation Technology in Biomedicine, vol. 15, no. 2, pp. 290300, 2011.

[19] Nez-Marcos A, Azkune G, Arganda-Carreras I (2017) Vision-based fall detection with convolutional neural networks. Wireless communications and mobile computing 2017: https://doi.org/10.1155/2017/9474806

[20] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 770778, July 2016.

[21] B. Kwolek and M. Kepski, Human fall detection on embedded platform using depth maps and wireless accelerometer, Computer Methods and Programs in Biomedicine, vol. 117, no. 3, pp.489501, 2014.

[22] Thome, N., Miguet, S., & Ambellouis, S. (2008). A Real-Time, Multi- view Fall Detection System: A LHMM-Based Approach. IEEE Transactions on Circuits and Systems for Video Technology,18(11), 1522-1532. doi:10.1109/tcsvt.2008.2005606

[23] Anderson, D., Luke, R. H., Keller, J. M., Skubic, M., Rantz, M., & Aud, M. (2009). Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Computer Vision and Image Under- standing,113(1), 80-89. doi:10.1016/j.cviu.2008.07.006

[24] Martnez-Villaseor, L., Ponce, H., Brieva, J., Moya-Albor, E., Nez- Martnez, J., & Peafort-Asturiano, C. (2019). UP-Fall Detection Dataset: A Multimodal Approach. Sensors,19(9), 1988. doi:10.3390/s19091988

[25] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature, vol. 521, no. 7553, pp. 436444, 2015.

[26] S. S. Beauchemin and J. L. Barron, The Computation of Optical Flow, ACM Computing Surveys, vol. 27, no. 3, pp.433466, 1995.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS 12), pp. 10971105, Lake Tahoe, Nev, USA, December 2012.
[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In ECCV, pages 818833, 2014.
[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 1, 2, 3
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014. 1, 2
[31] C. Francois and et al., Keras, 2015, https://github.com/fchollet/keras.
[32] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Man- agement,45(4), 427-437. doi:10.1016/j.ipm.2009.03.002
[33] Wang, S., Chen, L., Zhou, Z., Sun, X., & Dong, J. (2015). Human fall detection in surveillance video based on PCANet. Multimedia Tools and Applications,75(19), 11603-11613. doi:10.1007/s11042-015-2698-y
[34] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. DIRO-Universit de Montral, Tech. Rep, 1350, 2010.
[35] Charfi, I.; Miteran, J.; Dubois, J.; Atri, M.; Tourki, R. Definition and Performance Evaluation of a Robust SVM Based Fall Detection Solution. SITIS 2012, 12, 218224.
[36] Kozina, S., Gjoreski, H., Gams, M., & Lutrek, M. (2013). Efficient Activity Recognition and Fall Detection Using Accelerometers. Communications in Computer and Information Science Evaluating AAL Systems Through Competitive Benchmarking,13-23. doi:10.1007/978-3-642- 41043-7 2
[37] K.Wang, G. Cao, D. Meng,W. Chen, andW. Cao, Automatic fall detection of human in video using combination of features, in Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, pp. 12281233, China, December 2016.
[38] Blanc-Talon, J. (2006). Advanced concepts for intelligent vision systems: 8th International Conference, ACIVS 2006: Antwerp, Belgium, September 18-21, 2006: Proceedings. Berlin: Springer.
[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, Training computationally efficient smartphone-based human activity recognition models, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8131 LNCS, pp. 426433, 2013.
[40] Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R.,& Dürucheb, R. (2017). CNN-based sensor fusion techniques for multimodal human activity recognition. Proceedings of the 2017 ACM International Symposium on Wearable Computers - ISWC 17. doi:10.1145/3123021.3123046
[41] L. C. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, Context-aware mobile health monitoring: Evaluation of different pat- tern recognition methods for classification of physical activity, in 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 52505253, 2008.
[42] Jalal, A., Kamal, S., & Kim, D. (2014). A Depth Video Sensor- Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. Sensors,14(7), 11735-11759. doi:10.3390/s140711735
[43] Torres-Huitzil, C., & Nuno-Maganda, M. (2015). Robust smartphone- based human activity recognition using a tri-axial accelerometer. 2015 IEEE 6th Latin American Symposium on Circuits & Systems (LAS- CAS). doi:10.1109/lascas.2015.7250435
[44] Vilarinho, T.; Farshchian, B.; Bajer, D.G.; Dahl, O.H.; Egge, I.; Hegdal, S.S.; Lnes, A.; Slettevold, J.N.; Weggersen, S.M. A combined smart- phone and smartwatch fall detection system. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Auto- nomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 2628 October 2015; pp. 14431448.
[45] Vavoulas, G., Pediaditis, M., Chatzaki, C., Spanakis, E. G., & Tsiknakis, M. (n.d.). The MobiFall Dataset. Artificial Intelligence,1218-1231. doi:10.4018/978-1-5225-1759-7.ch048
[46] Kerdjidj, O., Ramzan, N., Ghanem, K., Amira, A., & Chouireb, F. (2019). Fall detection and human activity classification using wearable sensors and compressed sensing. Journal of Ambient Intelligence and Humanized Computing. doi:10.1007/s12652-019-01214-4
[47] Khojasteh, S., Villar, J., Chira, C., Gonzlez, V., & Cal, E. D. (2018). Improving Fall Detection Using an On-Wrist Wearable Accelerometer. Sensors,18(5), 1350. doi:10.3390/s18051350

[48] Bortnikov, M., Khan, A., Khattak, A. M., & Ahmad, M. (2019). Accident Recognition via 3D CNNs for Automated Traffic Monitoring in Smart Cities. Advances in Intelligent Systems and Computing Advances in Computer Vision, 256-264. doi:10.1007/978-3-030-17798-0 22
[49] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. & Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks, Nature, 542 (7639), 115-118, doi:10.1038/nature21056
[50] A. H. Fakhrulddin, X. Fei, and H. Li. Convolutional neural networks (cnn) based human fall detection on body sensor networks (bsn) sensor data. In 2017 4th ICSAI, Nov 2017.
[51] Nait Aicha, A., Englebienne, G., van Schooten, K. S., Pijnappels, M.,& Krse, B. (2018). Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry. Sensors (Basel, Switzerland) , 18 (5), 114. https://doi.org/10.3390/s18051654
[52] Lu, N., Wu, Y., Feng, L., & Song, J. (2019). Deep Learning for Fall Detection: Three-Dimensional CNN Combined With LSTM on Video Kinematic Data. IEEE Journal of Biomedical and Health Informatics,23(1), 314-323. doi:10.1109/jbhi.2018.2808281
[53] Casilari, E., Santoyo-Ramn, J., & Cano-Garca, J. (2017). Analysis of Public Datasets for Wearable Fall Detection Systems. Sensors, 17(7), 1513. doi:10.3390/s17071513
[54] Shieh, W., & Huang, J. (2012). Falling-incident detection and throughput enhancement in a multi-camera video- surveillance system. Medical Engineering & Physics,34(7), 954-963. doi:10.1016/j.medengphy.2011.10.016
[55] Mousse, M. A., Motamed, C., & Ezin, E. C. (2016). Percentage of human-occupied areas for fall detection from two views. The Visual Computer,33(12), 1529-1540. doi:10.1007/s00371-016-1296-y
[56] Zhang, S., Li, Z., Wei, Z., & Wang, S. (2016). An automatic human fall detection approach using RGBD cameras. 2016 5th International Conference on Computer Science and Network Technology (ICCSNT). doi:10.1109/iccsnt.2016.8070265
[57] Hekmat, M., Mousavi, Z., & Aghajan, H. (2016). Multi-view Feature Fusion for Activity Classification. Proceedings of the 10th International Conference on Distributed Smart Camera - ICDSC 16. doi:10.1145/2967413.2967434
[58] Su, S., Wu, S., Chen, S., Duh, D., & Li, S. (2015). Multi-view fall detection based on spatio-temporal interest points. Multimedia Tools and Applications,75(14), 8469-8492. doi:10.1007/s11042-015-2766-3
[59] Kong, Y., Huang, J., Huang, S., Wei, Z., & Wang, S. (2019). Learning spatiotemporal representations for human fall detection in surveillance video. Journal of Visual Communication and Image Representation,59, 215-230. doi:10.1016/j.jvcir.2019.01.024
[60] Koshmak, G., Loutfi, A., & Linden, M. (2016). Challenges and Issues in Multisensor Fusion Approach for Fall Detection: Review Paper. Journal of Sensors,2016, 1-12. doi:10.1155/2016/6931789
[61] Wu, Y., Su, Y., Hu, Y., Yu, N., & Feng, R. (2019). A Multi-sensor Fall Detection System Based on Multivariate Statistical Process Analysis. Journal of Medical and Biological Engineering , 39 (3), 336351. https://doi.org/10.1007/s40846-018-0404-z
[62] Wu, Y., Su, Y., Hu, Y., Yu, N., & Feng, R. (2019). A Multi-sensor Fall Detection System Based on Multivariate Statistical Process Analysis. Journal of Medical and Biological Engineering , 39 (3), 336351. https://doi.org/10.1007/s40846-018-0404-z
[63] Mubashir, M., Shao, L., & Seed, L. (2013). A survey on fall detection: Principles and approaches. Neurocomputing , 100 , 144152. https://doi.org/10.1016/j.neucom.2011.09.037
[64] Dong, Z., Li, F., Ying, J., & Pahlavan, K. (2018). Indoor motion detection using Wi-Fi channel state information in flat floor environments versus in staircase environments. Sensors (Switzerland) , 18 (7). https://doi.org/10.3390/s18072177
[65] Mao, A., Ma, X., He, Y., & Luo, J. (2017). Highly portable, sensor- based system for human fall monitoring. Sensors (Switzerland), 17(9). https://doi.org/10.3390/s170920 96
[66] Zhang, Z., Conly, C., & Athitsos, V. (2014). Evaluating Depth-Based Computer Vision Methods for Fall Detection under Occlusions . 196207. https://doi.org/10.1007/978-3-319-14364-4 19
[67] Zhang, Z., Conly, C., & Athitsos, V. (2015). A survey on vision-based fall detection. Proceedings of the 8th ACM international conference on Pervasive technologies related to assistive environments. ACM, 2015. http://dx.doi.org/10.1145/2769493.2769540
[68] Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window Size Impact in Human Activity Recognition. Sensors, 14(4), 64746499. doi: 10.3390/s140406474

[69] Khalid, S.; Khalil, T.; Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In Proceedings of the Science and Information Conference (SAI), London, UK, 2729 August 2014.
[70] Hsieh, Y.-Z., & Jeng, Y.-L. (2018). Development of Home Intelligent Fall Detection IoT System Based on Feedback Optical Flow Convolutional Neural Network. IEEE Access, 6, 60486057. doi: 10.1109/access.2017.2771389
[71] K. He, X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in Proc. 13th Eur. Conf. Comput. Vis., 2014, pp. 346361.
[72] Akula, N. V. A., Shah, A. K., & Ghosh, R. (2018). A spatio-temporal deep learning approach for human action recognition in infrared videos. Optics and Photonics for Information Processing XII. doi: 10.1117/12.2502993
[73] Sucerquia, A., Lpez, J. D., & Vargas-Bonilla, F. (2018). Real- Life/Real-Time Elderly Fall Detection with a Triaxial Accelerometer. doi: 10.20944/preprints201711.0087.v3

HIGHLIGHTS

- A human fall detection system based on multiple cameras and CNN is proposed.
- This fall detection system achieves an accuracy of 95.64% using only two cameras.
- This fall detection system competes with the state-of-the-art methods using images.

Manuscript Title:
A Vision-Based Approach for Fall Detection using Multiple Cameras and Convolutional Neural Networks: A Case Study using the UP-Fall Detection Data Set

Authors:
Ricardo Espinosa, Hiram Ponce, Sebastián Gutiérrez, Lourdes Martínez-Villaseñor, Jorge Brieva, Ernesto Moya-Albor

Conflict of Interest Statement:
None Declared

13. Prueba de aceptación.

---------- Forwarded message ---------
De: **Computers in Biology and Medicine** <eesserver@eesmail.elsevier.com>
Date: vie., 25 oct. 2019 a las 12:38
Subject: Computers in Biology and Medicine - Your Submission CBM-D-19-01416R4
To: <hiram.eredin@gmail.com>, <hponce@up.edu.mx>
Cc: <edwardciaccio@gmail.com>

With Reference To:
Journal: Computers in Biology and Medicine
Manuscript Number: CBM-D-19-01416R4
Article Type: Full Length Article
Title: A Vision-Based Approach for Fall Detection Using Multiple Cameras and Convolutional Neural Networks: A Case Study Using the UP-Fall Detection Dataset
Authors: Ricardo Espinosa, M.Sc.; Hiram Eredin Ponce, Ph.D.; Sebastián Gutiérrez, Ph.D.; Lourdes Martínez-Villaseñor, Ph.D.; Jorge Brieva, Ph.D.; Ernesto Moya-Albor, Ph.D.

Dear Dr. Hiram Eredin Ponce,

We are pleased to accept your manuscript for publication in our journal.

Editor-In-Chief:
and studies on modality approaches for fall detection and classification are required.
maybe change in the galleys to -
and more studies on modality approaches for fall detection and classification are needed.

Thank you for your contribution. The Publisher will contact you shortly with further details about completing the publication process.

No action is required from you at this time.

14. Prueba de publicación.

Computers in Biology and Medicine

Available online 30 October 2019, 103520

In Press, Journal Pre-proof

A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset

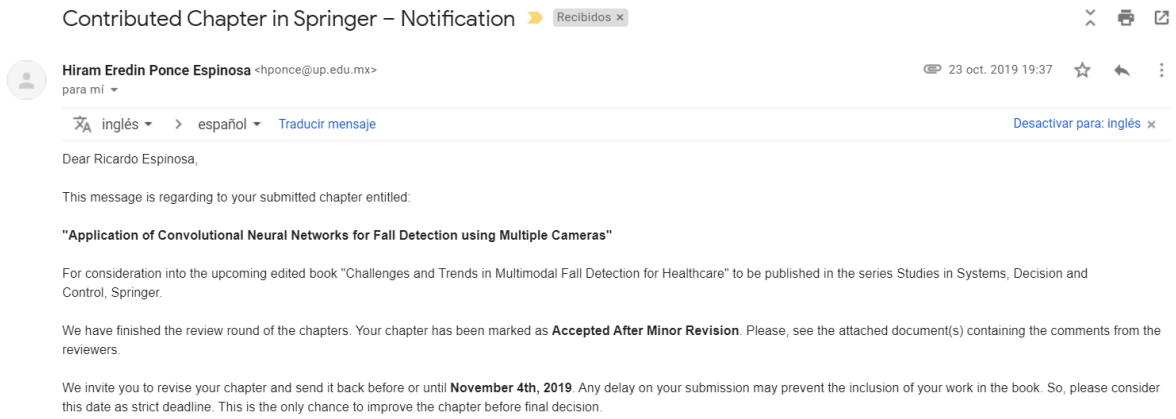
Ricardo Espinosa [a] ✉, Hiram Ponce [b] ✉, Sebastián Gutiérrez [a] ✉, Lourdes Martínez-Villaseñor [b] ✉, Jorge Brieva [b] ✉, Ernesto Moya-Albor [b] ✉

⊞ Show more

https://doi.org/10.1016/j.compbiomed.2019.103520

Get rights and content

15. Trabajos relacionados.



Contributed Chapter in Springer – Notification Recibidos ×

Hiram Eredin Ponce Espinosa <hponce@up.edu.mx> 23 oct. 2019 19:37
para mí

inglés español Traducir mensaje Desactivar para: inglés

Dear Ricardo Espinosa,

This message is regarding to your submitted chapter entitled:

"Application of Convolutional Neural Networks for Fall Detection using Multiple Cameras"

For consideration into the upcoming edited book "Challenges and Trends in Multimodal Fall Detection for Healthcare" to be published in the series Studies in Systems, Decision and Control, Springer.

We have finished the review round of the chapters. Your chapter has been marked as **Accepted After Minor Revision**. Please, see the attached document(s) containing the comments from the reviewers.

We invite you to revise your chapter and send it back before or until **November 4th, 2019**. Any delay on your submission may prevent the inclusion of your work in the book. So, please consider this date as strict deadline. This is the only chance to improve the chapter before final decision.