



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

**Centro Nacional de Investigación
y Desarrollo Tecnológico**

Tesis de Maestría

**Descripción automática de acciones en secuencias
de imágenes**

presentada por

Ing. Diego Alejandro Reyes Solorio

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis

Dr. Raúl Pinto Elías

Codirector de tesis

Dr. Noé Alejandro Castro Sánchez

Comité tutorial

Dra. Andrea Magadán Salazar

Dr. Jonathan Tavira Villanueva

Cuernavaca, Morelos, México. Octubre de 2025.



Cuernavaca, Mor., 06/Octubre/2025

OFICIO No. DCC/265/2025

Asunto: Aceptación de documento de tesis

CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **Diego Alejandro Reyes Solorio** con número de control **M23CE076**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Descripción automática de acciones en secuencias de imágenes"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al Servicio de México"



Dr. Raúl Pinto Elías
Director de tesis



Dr. Noé Alejandro Castro Sánchez
Codirector de tesis



Dr. Andrea Magadán Salazar
Revisor 1



Dr. Jonathan Villanueva Tavira
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante



2025
Año de
La Mujer
Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx





Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 10/octubre/2025

Oficio No. SAC/238/2025

Asunto: Autorización de impresión de tesis

DIEGO ALEJANDRO REYES SOLORIO
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Descripción automática de acciones en secuencias de imágenes"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

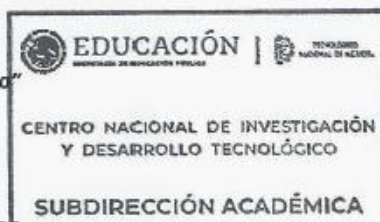
Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al Servicio de México"

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



2025
Año de
La Mujer
Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4104,
e-mail: acad_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico



DEDICATORIA

A mi madre, por su fortaleza, sacrificios y por ser siempre el pilar de mi vida y mostrarme que no hay límites cuando hay motivación. Su ejemplo, su amor y su dedicación han sido mi mayor inspiración. Este logro es por ella y para ella.

A mi esposa, por ser mi compañera de vida, por su amor incondicional, su apoyo constante y su fe en mí incluso cuando yo dudaba.

A mis hermanos, por su apoyo, su compañía y por compartir conmigo cada paso de este camino. Gracias por motivarme a seguir adelante.

A mis compañeros, con quienes compartí no solo aprendizajes académicos, sino también momentos que hicieron más significativa esta etapa.

AGRADECIMIENTOS

Principalmente me gustaría externar mi más profundo agradecimiento a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), cuyo respaldo económico fue un pilar fundamental para la realización de mis estudios de posgrado. Su apoyo no solo alivió la carga financiera, sino que también representó un voto de confianza en mi formación académica y profesional.

Al Tecnológico Nacional de México / Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), gracias por abrirme las puertas a un entorno académico de excelencia. Agradezco sinceramente la oportunidad, así como las instalaciones, herramientas y un ambiente propicio para el desarrollo de este trabajo. Este logro no habría sido posible sin el respaldo institucional que me ofrecieron en cada etapa del proceso.

A mi director de tesis, el Dr. Raúl Pinto Elías y a los miembros del comité tutorial conformado por la Dr. Andrea Magadán Salazar y el Dr. Jonathan Tavera Villanueva les extiendo mi más sincera gratitud por su orientación constante, sus observaciones y sus recomendaciones, ya que no solo enriqueció este proyecto, sino que también fue clave en mi formación personal.

A todos los docentes que contribuyeron a mi formación a lo largo del programa, les agradezco profundamente por compartir sus conocimientos, su experiencia y por fomentar en mí una actitud reflexiva. Cada una de sus enseñanzas ha dejado una huella duradera en mi camino profesional.

RESUMEN

El reconocimiento automático de acciones humanas plantea diversos desafíos técnicos relacionados con la variabilidad del movimiento corporal, la similitud entre ciertas acciones y las limitaciones inherentes a los enfoques basados en imágenes RGB. Factores como la complejidad física del sujeto, la velocidad del movimiento, el entorno o el mobiliario introducen variaciones incluso dentro de una misma clase de acción. Además, existen acciones con trayectorias de movimiento similares, especialmente en sus fases iniciales o finales, lo cual dificulta su distinción por parte de los modelos de aprendizaje profundo. Estas condiciones motivan la necesidad de sistemas que utilicen representaciones más estructuradas del cuerpo humano, como los esqueletos, para mejorar la precisión y la generalización.

Este proyecto presenta el desarrollo e implementación de un sistema de reconocimiento automático de acciones humanas, utilizando representaciones esqueléticas y redes neuronales convolucionales en grafos con modelado espaciotemporal (ST-GCN). El objetivo principal fue diseñar una arquitectura que, a partir de la estimación de poses corporales en videos, pudiera identificar un conjunto definido.

Los resultados obtenidos muestran que el modelo alcanzó una precisión promedio del 99.37% en el conjunto de prueba, con valores F1-Score superiores al 98.9% en todas las clases. Las acciones con diferencias muy pronunciadas, fueron clasificadas de manera correcta, mientras que, en acciones con posturas parecidas entre sí, como “de pie” y “dolor de pecho”, se presentaron algunas confusiones que no comprometen el rendimiento general del modelo. Además, se validó su funcionamiento mediante transmisiones de video, confirmando su capacidad para realizar inferencias continuas con baja latencia y alta estabilidad en escenarios controlados.

Este trabajo muestra que el uso de grafos esqueléticos y arquitecturas ST-GCN permite identificar acciones humanas específicas, representando una solución para aplicaciones de monitoreo en ambientes controlados.

ABSTRACT

Automatic human action recognition presents several technical challenges related to the variability of body movement, the similarity between certain actions, and the inherent limitations of RGB image-based approaches. Factors such as the subject's physical build, movement speed, environment, or surrounding objects introduce variations even within the same action class. Additionally, some actions share similar motion trajectories, particularly in their initial or final stages, making them difficult to distinguish for deep learning models. These conditions highlight the need for systems that utilize more structured representations of the human body, such as skeletons, to improve accuracy and generalization.

This project presents the development and implementation of an automatic human action recognition system using skeletal representations and spatio-temporal graph convolutional networks. The main objective was to design an architecture capable of identifying a predefined set of actions based on body pose estimation from image sequences.

The results show that the model achieved an average accuracy of 99.37% on the test set, with F1-Score values above 98.9% across all classes. Actions with clearly distinct kinematic patterns were classified correctly, while actions with similar body postures exhibited minor misclassifications that did not compromise the overall model performance. Additionally, the model was validated through RTSP video streams, confirming its ability to perform continuous inference with low latency and high stability in controlled environments.

ÍNDICE

1	Capítulo 1 Introducción	1
1.1	Descripción del problema.....	1
1.1.1	Delimitación del problema.....	2
1.1.2	Complejidad del problema	2
1.2	Contexto teórico/práctico	3
1.2.1	Marco conceptual.....	4
1.2.2	Antecedentes del proyecto	11
1.2.3	Exploración del estado del arte.....	12
1.2.4	Resumen del análisis al estado del arte	23
1.2.5	Propuesta de solución	29
1.3	Discusión	29
1.4	Estructura del documento	30
2	Capítulo 2 Propuesta de solución	31
2.1	Planteamiento de solución	31
2.1.1	Objetivo general.....	31
2.1.2	Objetivos específicos.....	32
2.1.3	Alcances	32
2.1.4	Limitaciones.....	32
2.1.5	Justificación	33
2.1.6	Beneficios	34
2.2	Modelo conceptual de la solución	34
2.3	Descripción de los elementos del modelo conceptual.....	36
2.3.1	Redes neuronales gráficas espaciales y temporales	36
2.3.2	Estructura general del modelo ST-GCN	36
2.3.3	Representación del cuerpo como grafo y matriz de adyacencia.....	37
2.3.4	Alternativas recientes a ST-GCN y comparación.....	38
2.3.5	Integración conceptual de secuencias de video.....	39
2.3.6	Estructura temporal de entrada y <i>buffers</i> de cuadros	39
2.3.7	Formato de entrada y conversión a tensores.....	40
2.3.8	Implementación de inferencia con RTSP	41
2.3.9	Técnicas de estabilización mediante votación	42
2.4	Variables y criterios de control	42

2.5	Discusión y motivación del enfoque adoptado	43
3	Capítulo 3 Diseño e implementación de la propuesta de solución.....	45
3.1	Diseño de la solución	45
3.1.1	Modelo conceptual ST-GCN para la detección de acciones.....	47
3.1.2	Representación esquelética y grafo espaciotemporal de poses	47
3.2	Implementación de la propuesta de solución	48
3.2.1	Adquisición de datos esqueléticos (MediaPipe).....	48
3.2.2	Preprocesamiento de datos esqueléticos (filtrado de ruido y normalización)	50
3.2.3	Preparación del conjunto de datos (división en conjuntos de entrenamiento, validación y prueba).....	52
3.2.4	Generación del grafo espaciotemporal de poses.....	53
3.2.5	Implementación del modelo ST-GCN	54
3.2.6	Entrenamiento del modelo (configuración de hiperparámetros).....	57
3.2.7	Evaluación del modelo entrenado.....	60
3.3	Integración del sistema	65
3.3.1	Flujo de video en tiempo real vía RTSP.....	65
3.3.2	Interfaz de usuario para visualización de resultados en tiempo real	66
3.3.3	Generación de evidencias, registro y almacenamiento de eventos .	66
3.3.4	Despliegue del sistema.....	67
4	Capítulo 4 Experimentación y resultados.....	70
4.1	Resumen de la experimentación.....	70
4.2	Herramientas y tecnologías	72
4.2.1	Hardware utilizado	72
4.2.2	Entorno de desarrollo.....	72
4.2.3	Bibliotecas principales	73
4.2.4	Gestión del sistema	74
4.3	Balance y partición de datos	78
4.3.1	Acciones contempladas y justificación de selección	79
4.3.2	Configuración y condiciones de captura	80
4.3.3	Proceso de extracción de datos.....	80
4.4	Resultados experimentales.....	81
4.4.1	Experimento 1: Entrenamiento base con conjunto de datos balanceado	82

4.4.2	Experimento 2: Evaluación del impacto de la longitud de la secuencia temporal	84
4.4.3	Experimento 3: Evaluación del modelo propuesto con normalización espacial de esqueletos	85
4.4.4	Experimento 4: Evaluación del modelo propuesto con aumento de datos	86
4.4.5	Experimento 5: Evaluación del tamaño de lote	87
4.4.6	Experimento 6: Evaluación del número de épocas	88
4.4.7	Experimento 7: Comparación de tasa de aprendizaje	89
4.4.8	Experimento 8: Evaluación del desempeño por clase	90
4.4.9	Experimento 9: Inferencia mediante voto por mayoría en ventana deslizante.....	91
4.4.10	Experimento 10: Sistema de inferencia	92
4.5	Análisis de los Resultados Obtenidos	95
4.6	Interpretación sobre la Aplicabilidad Práctica	95
5	Capítulo 5 Conclusiones	97
5.1	Conclusiones generales.....	97
5.2	Objetivos alcanzados	97
5.3	Productos generados	100
5.4	Trabajo futuro.....	100
	Referencias	102
	Anexo A.....	115

ÍNDICE DE FIGURAS

Figura 1.1 Flujo general de un sistema HAR (elaboración propia)	3
Figura 1.2 Flujo general del reconocimiento de acciones en video tomado de [21]	4
Figura 1.3 Ejemplo de estimación de pose y representación esquelética tomado de [24]	5
Figura 1.4 Topología de esqueleto 2D y numeración de articulaciones tomado de [5]	6
Figura 1.5 Estimación de pose multipersona OpenPose 2D, enlaces para cuerpo, manos y rasgos faciales tomado de [34]	7
Figura 1.6 Modelo SMPL con 22 huesos, izq. Pose neutra der. Aplicación de rotaciones tomado de [42]	8
Figura 1.7 Diagrama de pasos fundamentales en el procesamiento de imágenes digitales traducido de [46]	10
Figura 2.1 Modelo conceptual de la propuesta de solución entrenamiento (elaboración propia)	35
Figura 2.2 Modelo conceptual de la propuesta de solución validación (elaboración propia)	35
Figura 3.1 Modelo esquelético de 25 articulaciones del NTU RGB+D tomado de [6]	49
Figura 3.2 Flujo de adquisición y preparación de datos para el modelo ST-GCN (elaboración propia).	51
Figura 3.3 Ejemplo de grafo utilizado como entrada al modelo	54
Figura 3.4 Representación del grafo espaciotemporal de poses utilizad, tomado de [99]	55
Figura 3.5 Flujo general del modelo ST-GCN, tomado de [99].	56
Figura 3.6 Diagrama de flujo del proceso de entrenamiento (elaboración propia)	62
Figura 3.7 Organización y almacenamiento de evidencias	67
Figura 3.8 Interfaz desarrollada para gestionar los modos de inferencia del sistema	68
Figura 3.9 Flujo de video recibido desde una cámara IP durante el proceso de inferencia.	69
Figura 4.1 Modelo esquelético de articulaciones utilizado, tomado de [6]	79
Figura 4.2 Comportamiento métricas durante el entrenamiento del modelo, por época.	83
Figura 4.3 Comparativa de precisión, recall y F1 por clase.	83
Figura 4.4 Matriz de confusión generada a partir de las predicciones del modelo propuesto sobre el conjunto	90
Figura 4.5 Reconocimiento de la acción “Caerse” con alerta visual	94

ÍNDICE DE TABLAS

Tabla 1.1 Resumen de análisis del estado del arte	24
Tabla 3.1 Distribución de las secuencias por clase en los subconjuntos de entrenamiento, validación y prueba.....	52
Tabla 3.2 Resumen de bloques del modelo ST-GCN	57
Tabla 3.3 Hiper parámetros utilizados.....	60
Tabla 4.1 Objetivos realizados	70
Tabla 5.1 Síntesis de objetivos y alcances definidos para el desarrollo y evaluación del modelo.....	98

ACRÓNIMOS

Acrónimo	Significado en inglés	Significado en español
3Mformer	<i>Multi-Order Multi-Mode Transformer</i>	Transformador Multi-Orden y Multi-Modalidad
AGCLSTM	<i>Attention Enhanced Graph Convolutional LSTM Network</i>	Red LSTM con Convolución Gráfica mejorada con atención
API	<i>Application Programming Interface</i>	Interfaz de Programación de Aplicaciones
BoW	<i>Bag of Words</i>	Bolsa de Palabras
CNN	<i>Convolutional Neural Network</i>	Red Neuronal Convolutacional
Conv-RBM	<i>Convolutional Restricted Boltzmann Machine</i>	Máquina de Boltzmann Restringida Convolutacional
CPU	<i>Central Processing Unit</i>	Unidad Central de Procesamiento
FPS	<i>Frames per second</i>	Cuadros por segundo
GCN	<i>Graph Convolutional Network</i>	Red Convolutacional de Grafos
GMM	<i>Gaussian Mixture Model</i>	Modelo de Mezcla Gaussiana
GPU	<i>Graphics Processing Unit</i>	Unidad de Procesamiento Gráfico
GRU	<i>Gated Recurrent Unit</i>	Unidad Recurrente con Puertas
HAR	<i>Human Action Recognition</i>	Reconocimiento de Acciones Humanas
HDD	<i>Hard Disk Drive</i>	Disco Duro
HMDB51	<i>Human Motion Database 51</i>	Base de Datos de Movimiento Humano 51
HMM	<i>Hidden Markov Model</i>	Modelo Oculto de Markov

ACRÓNIMOS (Continuación)

Acrónimos	Significado en inglés	Significado en español
HOF	<i>Histogram of Optical Flow</i>	Histograma de Flujo Óptico
HOG3D	<i>Histogram of Oriented 3D Gradients</i>	Histograma de Gradientes 3D Orientados
iDT	<i>Improved Dense Trajectories</i>	Trayectorias Densas Mejoradas
IIP	<i>Intra-Inter Part Transformer</i>	Transformador de Partes Intra-Inter
ISTs	<i>Interactive Spatiotemporal Tokens</i>	Tokens Espaciotemporales Interactivos
IXMAS	<i>INRIA Xmas Motion Acquisition Sequences</i>	Secuencias de Adquisición de Movimiento INRIA Xmas
JHMDB	<i>Joint Human Motion Database</i>	Base de Datos de Movimiento Humano Conjunto
KLHA3D102	<i>KLHA3D102 Dataset</i>	Conjunto de Datos KLHA3D102
KTH	<i>KTH Dataset</i>	Conjunto de Datos KTH
LGST	<i>Local and Global Spatial–Temporal Transformer</i>	Transformador Espaciotemporal Local y Global
LoDVP	<i>Local Daily Video Pose Dataset</i>	Conjunto de Datos de Poses de Video Diario Local
LSTM	<i>Long Short-Term Memory</i>	Memoria a Largo Corto Plazo
MBH	<i>Motion Boundary Histogram</i>	Histograma de Bordos de Movimiento
MOD20	<i>Multimodal Office Dataset 20</i>	Conjunto de Datos Multimodal de Oficina 20

ACRÓNIMOS (Continuación)

Acrónimo	Significado en inglés	Significado en español
MSAAGCN	<i>Multi-Stream Adaptive Attention-enhanced Graph Convolutional Network</i>	Red de Convolución de Grafos Multi-Stream con Atención Adaptativa Mejorada
NTU RGB+D	<i>Nanyang Technological University RGB+D Dataset</i>	Conjunto de Datos RGB+D de la Universidad Tecnológica de Nanyang
NW-UCLA	<i>Northwestern University - University of California Los Angeles Dataset</i>	Conjunto de Datos de la Universidad Northwestern
PP	<i>Percentage points</i>	Puntos porcentuales
RAM	<i>Random Access Memory</i>	Memoria de Acceso Aleatorio
RBM	<i>Restricted Boltzmann Machine</i>	Máquina de Boltzmann Restringida
RGB	<i>Red, Green, Blue (color model)</i>	Rojo, Verde, Azul (modelo de color)
RNN	<i>Recurrent Neural Network</i>	Red Neuronal Recurrente
RTSP	<i>Real Time Streaming Protocol</i>	Protocolo de Transmisión en Tiempo Real
SBU	<i>Stony Brook University Dataset</i>	Conjunto de Datos de la Universidad de Stony Brook
SkateFormer	<i>Skeletal Temporal Transformer</i>	Transformador Temporal Esquelético
SMPL	<i>Skinnd Multi-Person Linear Model</i>	Modelo Lineal Multi-Persona con Piel
SMPL-X	<i>SMPL eXtended</i>	SMPL Extendido
SpStGCN	<i>Spatial-Structural Graph Convolutional Network</i>	Red Convolutiva de Grafos Espacial-Estructural

ACRÓNIMOS (Continuación)

Acrónimo	Significado en inglés	Significado en español
SSD	<i>Solid State Drive</i>	Unidad de Estado Sólido
ST-GCN	<i>Spatio-Temporal Graph Convolutional Network</i>	Red Convolutiva de Grafos Espaciotemporales
STAN	<i>SpatioTemporal Analysis Network</i>	Red de Análisis Espaciotemporal
STAR	<i>SpatioTemporal Cross Attention Transformer</i>	Transformador Espaciotemporal de Atención Cruzada
STEP CAT	<i>SpatioTemporal Effective BodyPart Cross Attention Transformer</i>	Transformador Espaciotemporal Efectivo de Atención Cruzada por Partes del Cuerpo
STIP	<i>Space-Time Interest Points</i>	Puntos de Interés Espacio-Temporales
STMT	<i>SpatioTemporal Mesh Transformer</i>	Transformador de Malla Espaciotemporal
SVM	<i>Support Vector Machine</i>	Máquina de Vectores de Soporte
UCF101	<i>University of Central Florida 101 Dataset</i>	Conjunto de Datos de la Universidad de Florida Central 101
ViViT	<i>Video Vision Transformer</i>	Transformador de Visión para Video

GLOSARIO

Término	Definición
<i>accuracy</i>	Métrica de evaluación que mide el porcentaje de predicciones correctas.
<i>augmentation</i>	Técnicas de modificación de datos para aumentar la diversidad del <i>dataset</i> .
<i>batch</i>	Subconjunto de datos procesados en una sola actualización de parámetros.
<i>dataset</i>	Conjunto de datos estructurados usado para entrenamiento o evaluación.
<i>descriptor</i>	Representación numérica de características de una imagen o secuencia.
<i>embedding</i>	Representación vectorial de características en un espacio de menor dimensión.
<i>epoch</i>	Iteración completa sobre todo el <i>dataset</i> durante el entrenamiento.
<i>F1-Score</i>	Medida de equilibrio entre <i>precision</i> y <i>recall</i> , expresada como su media armónica.
<i>feature</i>	Característica extraída de los datos que representa información relevante.
<i>frame</i>	Imagen individual dentro de una secuencia de video.
<i>ground truth</i>	Etiqueta o valor real de referencia en un <i>dataset</i> .
<i>hyperparameter</i>	Parámetro ajustado manualmente que controla el comportamiento de un modelo.
<i>inference</i>	Proceso mediante el cual un modelo entrenado genera predicciones.
<i>joint</i>	Punto de articulación en un modelo de esqueleto humano.
<i>kernel</i>	Matriz utilizada en operaciones de convolución para detección de patrones.

GLOSARIO (Continuación)

Término	Definición
<i>loss function</i>	Función que mide el error entre predicciones y valores reales.
<i>noise</i>	Información irrelevante o distorsión presente en los datos originales.
<i>normalization</i>	Proceso de ajuste de datos para mantener escalas consistentes.
<i>optimization</i>	Proceso de ajuste de parámetros para minimizar la función de pérdida.
<i>overfitting</i>	Fenómeno en el cual un modelo aprende patrones específicos de entrenamiento y pierde capacidad de generalización.
<i>pipeline</i>	Secuencia de etapas de procesamiento que sigue un flujo de datos.
<i>precision</i>	Métrica que mide la proporción de verdaderos positivos entre positivos predichos.
<i>recall</i>	Métrica que mide la proporción de verdaderos positivos respecto al total de positivos reales.
<i>robustness</i>	Capacidad de un modelo para mantener rendimiento frente a variaciones o ruido.
<i>sequence</i>	Conjunto ordenado de <i>frames</i> que representan un movimiento o acción.
<i>skeleton</i>	Representación del cuerpo humano mediante puntos articulares.
<i>stream</i>	Flujo de video o datos transmitidos en tiempo real.
<i>testing</i>	Evaluación de un modelo en datos no vistos para medir su rendimiento.
<i>threshold</i>	Valor de referencia usado para tomar decisiones en la clasificación.

GLOSARIO (Continuación)

Término	Definición
<i>topology</i>	Estructura de conexiones entre nodos en un grafo, aplicada al esqueleto humano.
<i>training</i>	Proceso de ajuste de parámetros de un modelo con datos de entrenamiento.
<i>validation</i>	Conjunto de datos usado para ajustar hiperparámetros y prevenir <i>overfitting</i> .

Capítulo 1

Introducción

El reconocimiento de acciones humanas consiste en identificar qué actividad realiza una persona a partir del análisis de secuencias de video [1]. Este trabajo busca que un sistema computacional interprete los patrones de movimiento del cuerpo para clasificar la acción correspondiente.

Una estrategia mencionada dentro del estado del arte permite representar el cuerpo humano mediante un modelo esquelético [2], la estimación de pose usa las coordenadas de las articulaciones del cuerpo. Esta representación abstrae la forma corporal y reduce la influencia de factores como el fondo o la iluminación, a diferencia de los métodos basados en imágenes RGB completas [3].

El modelo ST-GCN interpreta cada cuadro de video como un grafo de articulaciones conectado en el espacio y el tiempo. Esta arquitectura captura tanto la postura corporal instantánea como su evolución dinámica en la secuencia, aprendiendo directamente de los datos esqueléticos sin necesidad de extracción manual de características [4].

La precisión del sistema depende directamente de la calidad de la estimación de pose. Para este propósito, se emplean herramientas como MediaPipe, desarrollada por Google, que permite estimar coordenadas articulares a partir de secuencias de imágenes [5].

En este trabajo se desarrolla un sistema que integra la estimación de pose y una red neuronal ST-GCN entrenada con el conjunto de datos NTU RGB+D [6], orientado al reconocimiento de cinco acciones: sentarse, estar de pie, saltar, caerse y dolor de pecho.

Esto permite reducir ambigüedades que se presentan cuando se utilizan imágenes completas y mejora la discriminación entre clases de acciones, incluso cuando las posturas involucradas comparten características estructurales.

1.1 Descripción del problema

Identificar de manera automática un conjunto definido de acciones, proporcionando a la entrada del sistema información en secuencias de imágenes (videos), mediante la estimación de poses, utilizando una red neuronal de tipo ST-GCN, entrenada previamente con secuencias esqueléticas [7].

En las últimas décadas, el HAR ha evolucionado desde el análisis de gestos simples hasta el monitoreo de actividades. No obstante, la adopción de este tipo de soluciones en escenarios reales aún es limitada, debido a que no se cuenta con sistemas suficientemente confiables para distinguir con precisión una acción humana en el momento en que ocurre [8], [9].

A pesar de los avances, persiste un desfase entre las capacidades de los modelos desarrollados y los requerimientos de las aplicaciones donde se espera su implementación [10]. En particular, la detección de acciones humanas sigue presentando complicaciones que reducen su utilidad práctica en contextos que exigen inmediatez, estabilidad y precisión.

Esto ha generado la necesidad de replantear cómo se define y resuelve el problema de reconocer una acción a partir de secuencias visuales. Es indispensable contar con sistemas que puedan adaptarse a la diversidad de situaciones, y emitir respuestas confiables, lo que convierte al reconocimiento de acciones humanas en un problema abierto que requiere soluciones más robustas, específicas y evaluables, [11].

1.1.1 Delimitación del problema

La investigación se limita al estudio del reconocimiento de acciones humanas en entornos controlados, considerando únicamente representaciones de esqueletos humanos como fuente de información. No se abordan escenarios con múltiples personas, variaciones ambientales ni acciones fuera del conjunto predefinido de interés. Esta delimitación establece el marco en el cual se evalúa la propuesta, diferenciando el alcance del trabajo frente a aplicaciones más generales o no estructuradas.

1.1.2 Complejidad del problema

Distintas actividades pueden compartir fases intermedias similares. Acciones como “sentarse” y “levantarse”, o “caminar” y “correr”, presentan trayectorias posturales parcialmente coincidentes que inducen confusiones tanto en algoritmos como en observadores humanos, por ello la ambigüedad entre clases de acción complica la discriminación final [12], [13].

A ello se suma la variabilidad inter e intra sujeto. Factores como edad, complexión, velocidad de ejecución y estilo individual hacen que una misma acción adopte múltiples manifestaciones, ampliando la dispersión estadística intra-clase y desafiando la capacidad de generalización de los modelos de reconocimiento [14].

Por último, varios enfoques requieren observar la acción completa antes de decidir, lo que resulta poco compatible con escenarios que exigen respuesta inmediata o continua [15]. Esta condición impone restricciones técnicas por ello es necesario equilibrar precisión, estabilidad y eficiencia bajo ruido, oclusiones parciales o posturas atípicas [16].

De este modo, el problema demanda estrategias que abstraigan elementos irrelevantes del entorno, preserven la evolución temporal del movimiento y permitan comparar ejecuciones distintas sin perder discriminación.

1.2 Contexto teórico/práctico

Entre sus áreas de aplicación destacan la seguridad y la videovigilancia, utilizadas para el monitoreo de espacios y la detección de comportamientos anómalos, en la salud, donde se desarrollan sistemas para la detección de caídas y el seguimiento de la movilidad en pacientes y adultos mayores [17], en los deportes, donde se analizan patrones de movimiento para evaluar el rendimiento de atletas y prevenir lesiones y la interacción humano-computadora, en la que los gestos corporales permiten habilitar interfaces naturales en entornos de realidad virtual, aumentada o domótica [18].

En el ámbito industrial, el HAR se aplica a la ergonomía, la supervisión de procesos y la prevención de riesgos laborales.

La Figura 1.1 muestra de manera esquemática el flujo general de un sistema HAR, desde las modalidades de adquisición de datos hasta la clasificación de acciones o la detección de eventos.

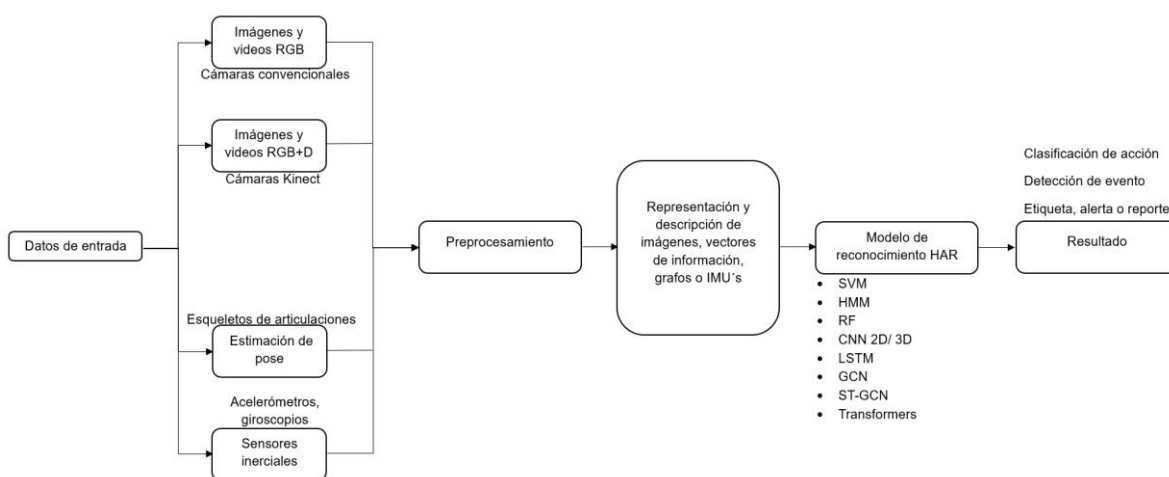


Figura 1.1 Flujo general de un sistema HAR (elaboración propia)

1.2.1 Marco conceptual

Reconocimiento de acciones humanas

El reconocimiento de acciones humanas es una disciplina que combina visión por computadora, aprendizaje automático y modelado del movimiento humano para identificar actividades a partir de secuencias de imágenes o video. Su propósito es permitir que los sistemas computacionales interpreten comportamientos humanos [19].

En términos computacionales, una acción puede definirse como una secuencia organizada de posturas del cuerpo humano (ver Figura 1.2) que, al observarse a lo largo del tiempo, forman un patrón reconocible [20]. Esta secuencia suele involucrar múltiples articulaciones que se coordinan en el espacio y el tiempo, como sucede en acciones como correr, agacharse o saludar.

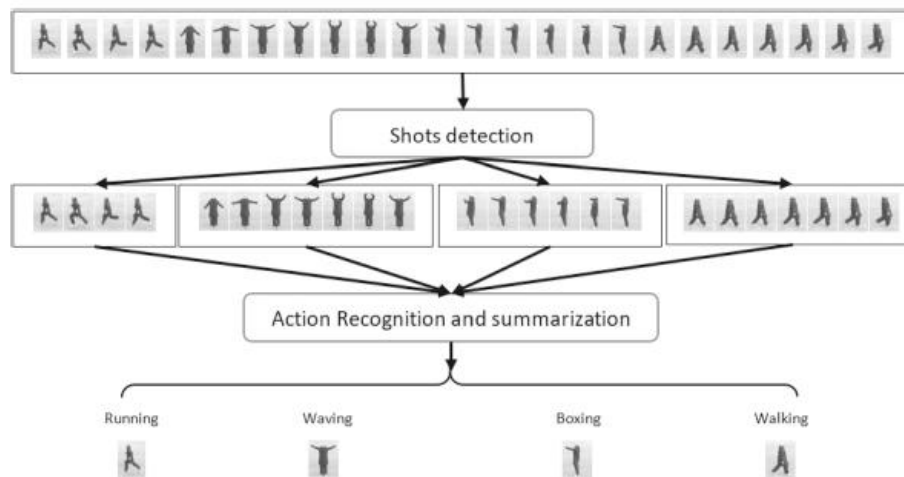


Figura 1.2 Flujo general del reconocimiento de acciones en video tomado de [21]

Cada acción puede variar ligeramente dependiendo del individuo que la realiza, su estilo de movimiento, velocidad y contexto, lo que introduce un grado de dificultad mayor para los sistemas computacionales.

Para superar estas limitaciones, surgieron representaciones centradas en la geometría del cuerpo humano como, se ilustra en la Figura 1.3, en particular a través de la estimación de pose. Esta técnica permite identificar las coordenadas espaciales de puntos clave del cuerpo [22] a lo largo de una secuencia, formando una estructura esquelética simplificada. Al abstraer la forma y apariencia del individuo, estas representaciones ofrecen menos variaciones ante factores externos y reducen la dimensionalidad de los datos de entrada [23].

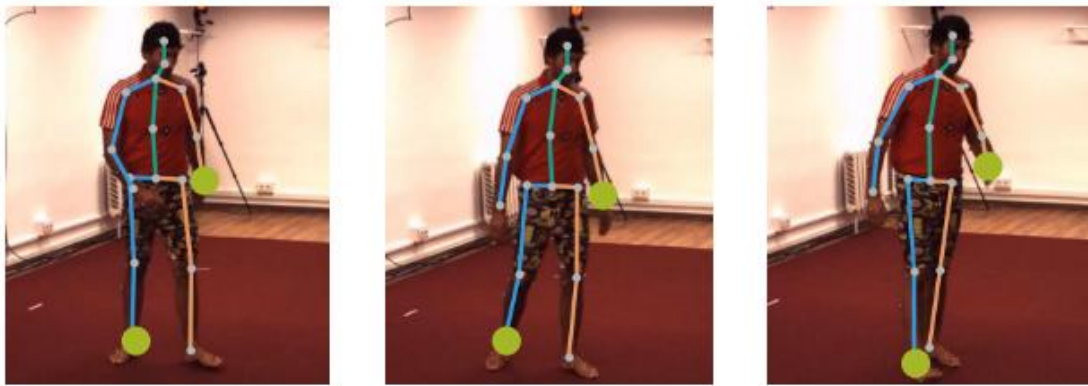


Figura 1.3 Ejemplo de estimación de pose y representación esquelética tomado de [24]

No obstante, la simple detección de posturas estáticas no basta para reconocer acciones, por esta razón, los sistemas HAR modernos requieren modelos que puedan capturar la evolución temporal del movimiento y las relaciones entre las articulaciones a lo largo del tiempo. Esta necesidad ha dado lugar a arquitecturas que integran componentes espaciales y temporales de manera conjunta para representar dinámicamente la acción en desarrollo [25].

Desde esta perspectiva, el reconocimiento de acciones humanas no solo implica identificar posiciones del cuerpo, sino también interpretar su progresión en una secuencia, analizando cómo interactúan las partes del cuerpo para generar movimiento. Esta capacidad permite a los sistemas inteligentes tener una comprensión contextual aproximada del comportamiento humano.

Estimación de pose y modelos esqueléticos

La estimación de pose humana se refiere a la tarea de detectar y localizar las partes del cuerpo de una persona (articulaciones) en datos visuales, para luego representarlas mediante una estructura articulada (esqueleto) [26]. En otras palabras, la estimación de pose identifica la posición de un conjunto de puntos articulares que representan todo el cuerpo humano a partir de imágenes o secuencias de video [27].

Esta representación suele modelar el cuerpo como un esqueleto de segmentos rígidos unidos por articulaciones (ver Figura 1.4), capturando así la postura o configuración corporal de la persona en la escena [28].

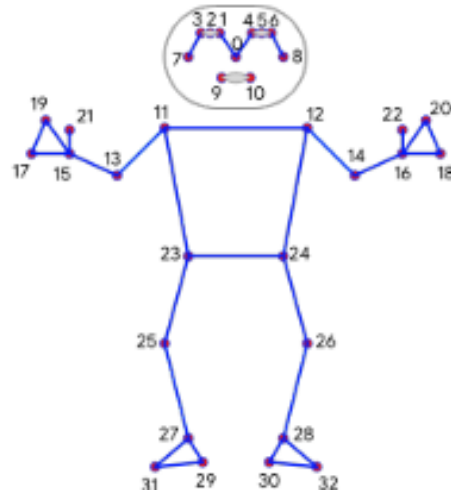


Figura 1.4 Topología de esqueleto 2D y numeración de articulaciones tomado de [5]

Formalmente, el resultado de la estimación de pose es un conjunto de coordenadas para cada articulación de interés (cabeza, hombros, codos, rodillas, etc.), ya sea en 2D (coordenadas en la imagen) o en 3D (coordenadas espaciales) según el tipo de pose estimada [29].

Si bien la estimación 2D es más sencilla y eficiente, la pose 3D ofrece una descripción más fiel del movimiento, aunque enfrenta desafíos adicionales como la estimación de profundidad y la necesidad de datos más complejos para su entrenamiento [30].

El uso de representaciones esqueléticas presenta ventajas al abstraer detalles visuales logrado mejorar la representación corporal lo que disminuye el error sobre variaciones del entorno y de los sujetos [31]. Además, los datos esqueléticos son más compactos, reduciendo el volumen de información a procesar y permitiendo que los algoritmos se centren en los patrones espaciales y temporales del movimiento.

La evolución de la estimación de pose ha seguido un recorrido similar al de otras áreas de la visión artificial. Los métodos tradicionales se basaban en modelos geométricos y características manuales, con resultados limitados en entornos reales, en contraste con los enfoques modernos, impulsados por el aprendizaje profundo.

Arquitecturas basadas en redes convolucionales como Open Pose o HRNet [32] permiten identificar y detectar puntos clave en cuerpo como manos y rostro, tal y como se observa en la Figura 1.5 [33]. Asimismo, el uso de modelos gráficos ha facilitado la incorporación de restricciones anatómicas y de coherencia estructural en la estimación.



Figura 1.5 Estimación de pose multipersona OpenPose 2D, enlaces para cuerpo, manos y rasgos faciales tomado de [34]

Sin embargo, persisten retos abiertos. Las oclusiones parciales, las posturas inusuales, los movimientos rápidos y las variaciones entre individuos continúan afectando la precisión de los modelos [35]. Además, en escenarios no controlados, las predicciones pueden incluir articulaciones mal localizadas, lo que compromete el análisis posterior de acciones. Estos problemas impulsan las investigaciones en técnicas de completado de pose y fusión de datos multimodales para mejorar la estabilidad de los sistemas [36].

Gracias a esto se puede decir que la representación esquelética reduce la dependencia del contexto visual y concentra la información en la configuración y dinámica corporal. Con esta base, en la sección 2.3.1 se describen los modelos gráficos espaciales-temporales que utilizan dicha representación.

Representación estructural del cuerpo humano

El análisis de acciones a partir de datos esqueléticos requiere definir cómo se organiza la información obtenida de las articulaciones detectadas. Para ello, se utilizan modelos que describen las relaciones espaciales entre puntos clave del cuerpo mediante estructuras jerárquicas que conectan articulaciones según su relación anatómica [37]. Esta organización permite representar de forma compacta la configuración corporal sin necesidad de procesar información visual adicional.

Topologías de grafo y representación jerárquica

Los esqueletos pueden definirse mediante topologías fijas o adaptativas. Las arquitecturas convencionales emplean una estructura predeterminada basada en conexiones anatómicas directas. En contraste, propuestas recientes introducen grafos con relaciones jerárquicas y escalas múltiples.

Entre estos enfoques se encuentran los grafos dirigidos condicionales, que ajustan las conexiones mediante aristas dirigidas según la postura analizada, así como las redes de grafos jerárquicos que incorporan enlaces entre articulaciones no adyacentes, ampliando el contexto espacial y temporal en el análisis [38].

Estos enfoques permiten capturar dependencias entre partes corporales que no están conectadas físicamente pero que influyen en el patrón de movimiento.

Tipos de modelos de esqueleto

Existen dos aproximaciones principales para representar el cuerpo humano a partir de datos articulares. El modelo cinemático organiza las articulaciones como puntos conectados mediante segmentos rígidos, lo que permite describir el movimiento sin necesidad de detalles adicionales sobre la forma corporal. Este modelo se utiliza en conjuntos de datos como Human3.6M y NTU RGB+D debido a su bajo costo computacional [39].

Por otra parte, el modelo volumétrico o de malla (del inglés *mesh*) describe el cuerpo mediante mallas tridimensionales paramétricas, como las implementadas en SMPL o SMPL-X (ver Figura 1.6), que permiten una representación más detallada de la forma corporal. Este tipo de modelo se emplea en aplicaciones que requieren mayor precisión geométrica, como animación digital o realidad virtual, donde es importante capturar variaciones anatómicas y deformaciones del cuerpo [40], [41].



Figura 1.6 Modelo SMPL con 22 huesos, izq. Pose neutra der. Aplicación de rotaciones tomado de [42]

Aunque ambos enfoques son útiles, en el reconocimiento de acciones se utiliza con mayor frecuencia el modelo cinemático, ya que presenta menor complejidad y se adapta mejor a las arquitecturas basadas en grafos empleadas en sistemas de análisis de movimiento.

Relevancia para el reconocimiento de acciones

La estructura esquelética facilita la extracción de patrones espaciales y su variación temporal, lo que se utiliza para clasificar acciones en diferentes contextos. Estudios recientes señalan que esta representación es menos sensible a factores externos como cambios de fondo, iluminación o apariencia, lo que contribuye a obtener resultados más consistentes en entornos no controlados [43].

Además, la organización jerárquica del esqueleto favorece el uso de métodos basados en grafos, como GCN o modelos híbridos, que procesan la información espacial y temporal de manera coherente. Estas técnicas aprovechan la matriz de adyacencia del grafo para transmitir información entre articulaciones de forma estructurada.

La estructura del esqueleto sirve como base para aplicar técnicas de análisis más profundas. A partir de esta organización, las siguientes secciones abordan el uso de redes neuronales gráficas y otros modelos que explotan esta representación para el reconocimiento automático de acciones.

Fundamentos de visión artificial en el análisis de secuencias

El análisis de secuencias por visión artificial se centra en el procesamiento consecutivo de imágenes o cuadros para identificar patrones espaciales y temporales que caractericen el comportamiento humano en vídeo. Este enfoque es útil para el reconocimiento de actividades porque permite capturar tanto poses como movimientos a través del tiempo, ofreciendo evidencia útil para clasificar acciones complejas [44].

En fases iniciales, los sistemas trataban cada cuadro de forma independiente, aplicando detectores de bordes, contornos o descriptores visuales clásicos (ver Figura 1.7). No obstante, estos enfoques no permiten modelar la dinámica del movimiento. Más adelante, el uso del flujo óptico describió desplazamientos aparentes entre cuadros, aportando dirección y velocidad del movimiento [45].

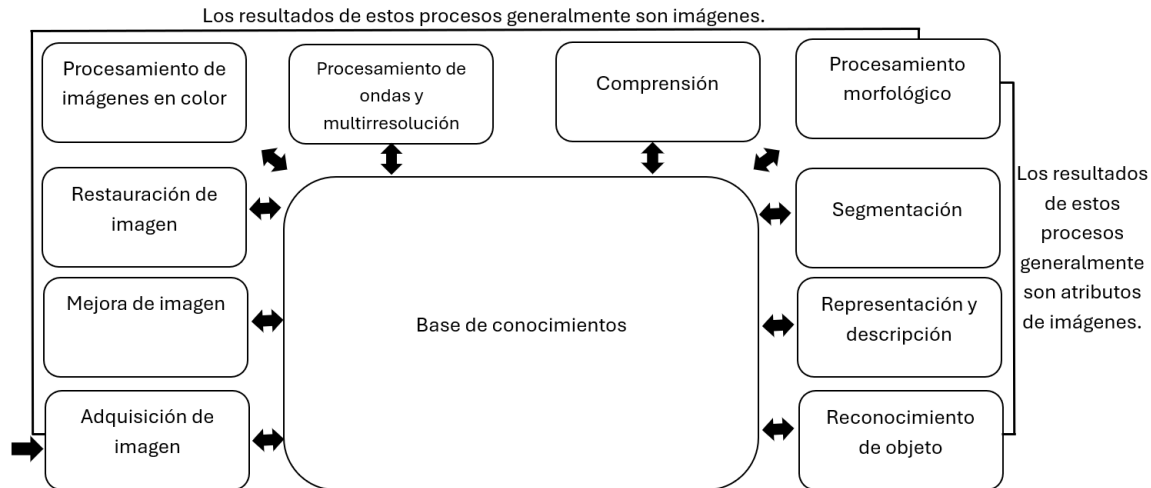


Figura 1.7 Diagrama de pasos fundamentales en el procesamiento de imágenes digitales traducido de [46]

Posteriormente, la evolución tecnológica dio paso a modelos más sofisticados como las redes convolucionales 3D que extienden la convolución en el eje temporal para capturar características espacio-temporales, y los modelos híbridos CNN+RNN para dependencias temporales prolongadas [47].

Más recientemente, los transformadores han permitido modelar dependencias globales entre cuadros mediante atención, sin recurrir a estructuras recurrentes. Asimismo, se han estudiado esquemas de fusión que integran información visual y esquelética para mejorar la robustez del reconocimiento [48]. Finalmente, también se exploran arquitecturas enfocadas en eficiencia computacional que mantienen buen rendimiento con menor complejidad.

También existen modelos que integran directamente atención cruzada entre modalidades visuales y esqueléticas. Por ejemplo, el *STAR-Transformer* presenta un mecanismo de atención que fusiona la información de vídeos RGB y la información de pose humana, mejorando su capacidad para distinguir acciones similares.

Reconocimiento de patrones en datos espaciotemporales

El estudio de relaciones que evolucionan en el espacio y el tiempo permite comprender cómo se desarrollan las acciones humanas en secuencias de datos. Este enfoque analiza no solo la configuración corporal en instantes aislados, sino también la continuidad y transición entre diferentes posturas, lo que aporta una descripción más completa de la actividad [49].

En los métodos clásicos, esta tarea se abordaba mediante el diseño manual de características, incluyendo distancias entre articulaciones, trayectorias de puntos clave o descriptores de movimiento obtenidos a partir de flujo óptico. Los

clasificadores empleados en esta etapa, como las máquinas de soporte vectorial [50], modelos de mezclas gaussianas y redes ocultas de Markov [51], lograban resultados adecuados únicamente en entornos controlados y mostraban dificultades al generalizar a escenarios con mayor variabilidad [52].

La introducción del aprendizaje profundo permitió superar estas limitaciones mediante arquitecturas capaces de aprender representaciones directamente de los datos. Las redes convolucionales tridimensionales y los modelos híbridos que integran convolución y recurrencia mostraron aptitudes al capturar tanto la información espacial de los fotogramas como las relaciones temporales entre ellos [53].

Posteriormente, las redes neuronales gráficas introdujeron un enfoque que modela los datos esqueléticos como grafos, propagando información entre articulaciones conectadas. Las variantes espaciotemporales ampliaron esta idea al incorporar explícitamente la dimensión temporal en la propagación de características [54].

Más recientemente, los modelos basados en mecanismos de atención, como los *transformers*, han mostrado ventajas en el análisis de secuencias complejas al capturar dependencias globales sin recurrir a estructuras recurrentes.

De forma complementaria, las estrategias multimodales que combinan información visual, esquelética y señales inerciales han dado lugar a sistemas más robustos frente a condiciones reales y entornos no estructurados [55].

Los avances recientes en el estudio de patrones espaciotemporales han ampliado las capacidades de los sistemas de reconocimiento de acciones, ofreciendo soluciones que se adaptan a diferentes escenarios y son menos sensibles a datos incompletos. Con estas bases, la sección siguiente aborda el análisis de soluciones previas y enfoques actuales presentes en el estado del arte.

1.2.2 Antecedentes del proyecto

En este apartado se exponen los antecedentes vinculados a la institución en la que se desarrolló la investigación, incorporando además una revisión de estudios previos afines que aborden al fenómeno en estudio o que propongan enfoques y herramientas aplicables al análisis del reconocimiento de acciones humanas.

Los trabajos desarrollados en el CENIDET han abordado diversas aproximaciones para el reconocimiento y análisis de actividades humanas, aportando soluciones basadas en visión por computadora y fusión de sensores.

Arriaga Santos (2024) [56] implementó un modelo para la estimación y clasificación de poses humanas a partir de puntos clave obtenidos de secuencias de imágenes,

comparando cuatro modelos de estimación de la pose (OpenPose, HRNet, YoloPose y BlazePose) sobre los conjuntos de datos CrowdPose y OcHuman. Para la etapa de clasificación, evaluó los algoritmos *random forest*, *logistic regression*, *ridge classifier* y *gradient boosting classifier* con métricas de *accuracy*, *precision* y *recall*, incluyendo experimentos con oclusiones parciales y auto-oclusiones.

García García [57] (2024) desarrolló un sistema de identificación automática de actividades humanas de interés, como estar de pie, caminar, sentarse, caer y “sin interés”, mediante un enfoque multimodal que integra una cámara RGB y sensores inerciales para la adquisición de datos. El modelo combina modelos de *machine learning* con reglas definidas, lo que permite operar con entradas provenientes de uno o varios sensores de manera individual o combinada. La validación del sistema evidenció su capacidad para adaptarse a distintos escenarios de monitoreo, manteniendo un desempeño estable en la clasificación de actividades en tiempo real y orientándose a aplicaciones en salud y seguridad.

Urquiza Flores [58] (2022) diseñó un modelo para la detección de anomalías en el comportamiento de una persona en entornos domésticos, utilizando cámaras RGB, acelerómetros y giroscopios como fuentes de datos. El modelo registra y modela la rutina diaria de una persona para generar un patrón de referencia y, posteriormente, identificar desviaciones que indiquen comportamientos anormales o inhabituales. La propuesta abarca desde la adquisición y procesamiento de datos, hasta la estimación de la pose y el modelado de secuencias de actividades mediante grafos dirigidos y máquinas de estado finito. Los resultados experimentales arrojaron que el modelo es capaz de detectar cambios en el patrón de comportamiento, clasificando eventos como anomalías por periodo, por día o por actividad específica.

1.2.3 Exploración del estado del arte

En esta sección se presentan estudios que abordan el reconocimiento de acciones humanas, centrándose en la identificación y análisis de patrones de movimiento del cuerpo. Se examinan diversos enfoques utilizados para detectar y clasificar acciones, desde movimientos cotidianos hasta actividades complejas, empleando técnicas basadas en visión por computadora, aprendizaje automático y modelado esquelético.

El objetivo es identificar herramientas, metodologías e instrumentos que puedan ser aplicados o adaptados al análisis automático del comportamiento humano, contribuyendo así al avance en la comprensión y detección de actividades a partir de secuencias de movimiento.

1.2.3.1 Métodos de aprendizaje tradicional para el reconocimiento de acciones

Action recognition using space-time interest points (STIP)

STIP detecta puntos salientes espaciotemporales (extensión de detectores clásicos al dominio 3D) y describe su vecindad con histogramas (HOG/HOF), que luego se cuantifican en un diccionario visual y se clasifican con SVM. Este enfoque y sus variantes fueron ampliamente evaluados en conjuntos clásicos (KTH, Weizmann y posteriores como UCF101 [59] y HMDB [60]) y constituyen la base de múltiples sistemas previos a *deep learning* [61].

Bag of visual words for action categorization

Este enfoque utiliza puntos de interés para detectar regiones relevantes en el video y luego extrae descriptores locales en cada región. Las características se cuantifican en un vocabulario visual mediante *clustering (k-means)*, formando un histograma de “palabras visuales” que representa cada secuencia, y finalmente se clasifican con un SVM.

Evaluado en los conjuntos KTH y Hollywood2, logró alrededor de 83% de *accuracy* en KTH, aunque en Hollywood2 el rendimiento fue menor debido a la alta variabilidad de escenas y ruido de fondo. Este método mostro su eficacia en acciones simples y reveló los desafíos presentes en entornos no controlados [62].

Hidden Markov models for action recognition [63]

Este enfoque modela cada acción como una secuencia de estados ocultos que representan las fases temporales del movimiento. A partir de la secuencia de video, se extraen características por cuadro (por ejemplo, siluetas) que sirven como observaciones para entrenar modelos HMM, cada acción se reconoce calculando la probabilidad de la secuencia bajo cada HMM y eligiendo la mayor.

Los autores evaluaron su método en un conjunto de datos propio de videos deportivos (golf, tenis, natación) de baja resolución, mostrando capacidad para distinguir acciones sencillas en secuencias bien segmentadas con variabilidad controlada. Este trabajo introdujo el modelado probabilístico temporal en HAR, aunque luego fue superado por técnicas más recientes.

Improved dense trajectories (iDT) [64]

Los autores desarrollaron iDT, describiendo acciones mediante trayectorias densas extraídas del video y calculando en ellas descriptores locales HOG, HOF y MBH que capturan información detallada de apariencia y movimiento. Estas características locales se agregan en representaciones globales mediante

diccionarios visuales y codificación estadística. Un aporte importante de iDT es la compensación del movimiento de cámara mediante transformaciones homográficas, aislando el movimiento humano del fondo. Validado en conjuntos de videos realistas como Hollywood2, UCF101 y HMDB51, iDT obtuvo resultados superiores a métodos contemporáneos gracias a la compensación de cámara, que aumentó la precisión en escenarios complejos.

Optical flow and motion boundary histograms [65]

El estudio propuso describir el movimiento en video combinando histogramas de flujo óptico con *Motion Boundary Histograms*, calculados como gradientes del flujo óptico. Mientras HOF captura la información de movimiento global, MBH elimina el componente de cámara y resalta variaciones internas debidas al movimiento de objetos o extremidades, aislando patrones de movimiento relevantes de la acción.

En pruebas con conjuntos de referencia de acciones simples en entornos controlados (KTH, Weizmann), la combinación HOF+MBH mejoró la precisión de clasificación respecto a usar solo descriptores de gradientes de imagen. Por ejemplo, en el conjunto de datos HMDB51 el descriptor MBH por sí solo superó el estado del arte en 16%, y combinándolo con otros descriptores espacio-temporales logró 5% adicional en *accuracy*.

Gaussian mixture models for action segmentation [66]

La investigación se centra en un método para segmentar y reconocer acciones modelando cada acción con un modelo de mezcla de Gaussiana a partir de características de bajo nivel (flujo óptico o gradientes).

Cada acción se representa como una combinación probabilística de componentes gaussianos, y la segmentación temporal se realiza clasificando ventanas temporales deslizantes usando el modelo de cada acción, evitando la necesidad de algoritmos complejos de alineamiento temporal.

En una versión modificada del conjunto de datos KTH, este método alcanzó 78.3% de *accuracy*, superando a un enfoque basado en HMM (71.2%) por 7.1 puntos porcentuales.

Silhouette-based shape context descriptors for action recognition [67]

En este artículo se representó cada acción como una secuencia de siluetas extraídas de los fotogramas del video, sobre las cuales calcularon descriptores *Shape Context*. Estos descriptores codifican la distribución espacial de los puntos del contorno en coordenadas polares logarítmicas, siendo invariantes a escala y

traslación, y robustos ante ligeras deformaciones corporales y ruido en la segmentación de siluetas.

En experimentos con secuencias sintéticas (variando el tipo de cuerpo, la orientación y el nivel de ruido), se comparó Shape Context con momentos de Hu y descriptores de Fourier usando clasificación *k*-NN. Bajo variaciones moderadas de vista y ruido, Shape Context obtuvo un error angular promedio de 150° en la estimación conjunta de articulaciones, frente a 152° con Fourier y 165° con Hu (suma de errores en nueve articulaciones), mostrando mejoras de 2° a 15° sobre los descriptores clásicos, evidenciando que los Shape Context proporcionan una representación aceptable de acciones basadas en siluetas.

Histogram of oriented 3D gradients (HOG3D) for motion analysis [68]

El método propuesto por los autores calcula gradientes tridimensionales (componentes espaciales + temporales) del video, proyecta sus orientaciones sobre un icosaedro para un conteo de forma discreta, y acumula esos gradientes en histogramas dentro de celdas 3D. Luego normaliza y concatena las celdas para formar el vector descriptor final, que captura conjuntamente la apariencia y la dinámica del movimiento.

Usando un modelo *Bag of Words* y SVM, se evaluó en los conjuntos Weizmann y KTH, comparándolo con HOG 2D, HOF y la combinación HOG-HOF. En Weizmann (diccionario de 4000 palabras), HOG3D alcanzó 90.7% de *accuracy*, superando a HOF (88.8%), HOG-HOF (85.6%) y a HOG 2D (75.3%). En KTH logró 89.0%, ligeramente por debajo de HOF (92.1%) y HOG-HOF (91.8%), pero por encima del HOG tradicional.

Spatio-temporal pyramids for human action classification [69]

Los autores presentaron una extensión tridimensional del descriptor SIFT aplicada a videos. Detectaron puntos de interés en el espacio y el tiempo mediante una pirámide de diferencias de Gaussianas, y calcularon descriptores SIFT 3D alrededor de esos puntos para capturar información espacial y temporal conjunta. Luego, cuantificaron esos descriptores en un diccionario visual (*bag of spatio-temporal words*) y clasificaron las secuencias con SVM, evaluando el desempeño en conjuntos como KTH y UCF Sports en comparación con descriptores 2D y métodos previos.

El estudio indicó que el descriptor SIFT 3D superó a los métodos basados en gradientes 2D y a enfoques anteriores en la clasificación de acciones, especialmente en escenarios donde las acciones presentan variaciones temporales significativas y se requiere invariancia espacio-temporal.

Trajectory-aligned features with Bag-of-Words encoding [70]

En el estudio se propuso caracterizar acciones mediante trayectorias densas alineadas con múltiples descriptores locales. A lo largo de cada trayectoria en el video se extraen descriptores HOG, HOF y MBH, los cuales luego se codifican mediante un modelo *Bag of Words* (*clustering k-means*) y se clasifican con una SVM de núcleo *chi-square*. La alineación de los descriptores a lo largo de las trayectorias asegura una captura coherente tanto de la apariencia como del movimiento a través del tiempo [71].

En experimentos con UCF Sports y Hollywood2, esta estrategia mejoró el desempeño frente a descriptores individuales y métodos previos: en UCF Sports, la combinación alineada HOG+HOF+MBH logró 88.0% de *accuracy*, superando a usar solo MBH (84.2%) y a las características de trayectoria sin alinear (75.4%); en Hollywood2 obtuvo 58.2% de mAP, por encima de HOF (50.3%), MBH (55.1%) y aproximaciones anteriores (53.3%).

1.2.3.2 Enfoques basados en aprendizaje profundo

Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network [72]

Los autores propusieron una arquitectura de doble flujo para reconocimiento de acciones basada en esqueletos humanos. El primer flujo es un módulo LSTM temporal que procesa la secuencia de coordenadas articulares 3D, capturando las dependencias temporales del movimiento, el segundo flujo (denominado *joint-motion module*) emplea convoluciones separables en profundidad para explotar patrones de movimiento entre articulaciones. Como entrada, se utilizan coordenadas 3D extraídas con MediaPipe y coordenadas 2D obtenidas mediante MoveNet, alimentando respectivamente la LSTM y la red convolucional.

La fusión de ambos flujos permite superar limitaciones de las RNN tradicionales (como la atenuación del gradiente y la dificultad para modelar dinámicas espaciotemporales complejas). El modelo alcanzó aproximadamente 73.3%, 97.7%, 95.2% y 94.0% de *accuracy*, en los conjuntos de datos JHMDB, Florence 3D, SBU Interaction y Penn Action respectivamente, superando los resultados reportados por otros métodos en cada caso.

Novel YOLO–LSTM approach for enhanced human action recognition in video sequences [73]

La investigación presenta un enfoque híbrido que emplea YOLO para la extracción de características espaciales y un modelo LSTM para capturar las dependencias temporales en secuencias de video. En este esquema, YOLO detecta objetos clave

y regiones relevantes en cada fotograma, produciendo representaciones espaciales que luego sirven de entrada a un LSTM optimizado (con *dropout*, normalización por capas y mecanismos inspirados en atención) para modelar la dinámica del movimiento humano a lo largo del tiempo.

El modelo fue evaluado en cuatro conjuntos de datos públicos (UCF101, KTH, Weizmann, IXMAS) y logró puntajes de 96% en las métricas *accuracy* y *precision*, 97% en *recall* y 96% en *F1-Score* para el conjunto de datos UCF101 obtuvo 99% en todos estos indicadores para KTH 100% de precisión en Weizmann y 98% en IXMAS.

A New Deep-Learning Method for Human Activity Recognition [74]

En el presente estudio propusieron un modelo profundo que combina convoluciones tridimensionales (3D-CNN) con una LSTM convolucional (*ConvLSTM*). En tres conjuntos de pruebas el modelo alcanzó 89.1% de precisión (LoDVP), 83.9% (UCF50mini) y 87.8% (MOD20), superando a modelos que utilizan solo 3D-CNN o solo LSTM por separado.

Lightweight human action recognition system based on LSTM and deformable convolution [75]

En la investigación desarrollaron un sistema ligero que integra un módulo LSTM para el procesamiento temporal con convoluciones deformables para la dimensión espacial. El LSTM extrae características globales desde la secuencia temporal de la acción, mientras que la red convolucional deformable se ajusta dinámicamente a variaciones espaciales en la posición y postura del sujeto. Gracias a la combinación de ambos módulos, la arquitectura logra capturar patrones dinámicos y espaciales, reduciendo el número de parámetros en comparación con modelos convencionales.

En evaluaciones comparativas, este modelo mostró un rendimiento igual o superior al de arquitecturas más complejas, destacando por su rapidez de inferencia y su capacidad de adaptación a variaciones en los datos, sin comprometer la precisión.

A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection [76]

Se diseñó una técnica híbrida que combina una máquina de Boltzmann restringida convolucional 2D (*Conv-RBM*) con una red LSTM, junto con un mecanismo optimizado de selección de fotogramas para aprender características espaciales (como bordes, texturas o patrones locales de movimiento) preservando las relaciones espaciales mediante la compartición de pesos. Estas características alimentan una LSTM que modela dependencias temporales tanto de corto como de largo plazo entre fotogramas consecutivos. Este método logró 97.3%, 94.8% y

81.5% de *accuracy* en los conjuntos de datos KTH, UCF Sports y HMDB51 respectivamente.

Improved semantic-guided network for skeleton-based action recognition [77]

La investigación propuso una versión mejorada del modelo *Semantic-Guided Neural Network*, proponiendo una arquitectura más sencilla a partir de datos esqueléticos. Esta propuesta se fundamenta en GCN's adaptativos que capturan automáticamente la estructura compleja de las articulaciones sin depender de topologías predefinidas. Luego, se incorpora un mecanismo de atención para reforzar la representatividad informativa de los nodos, tanto a nivel espacial como temporal.

Este marco alcanzó valores de *accuracy* de 89.7% en la configuración *Cross-Subject* y 95.0% en *Cross-View* en NTU-RGB+D 60. Para NTU-RGB+D 120, los valores fueron 84.6% en *Cross-Subject* y 85.8% en *Cross-Setup*.

Deep learning for human activity recognition on 3D human skeleton: survey and comparative study [78]

El artículo presenta una revisión de métodos de aprendizaje profundo con datos esqueléticos 3D, abarcando arquitecturas RNN, CNN, GCN/GNN y modelos híbridos. Se realiza un análisis comparativo en múltiples conjuntos de datos, entre los que se incluyen NTU RGB+D, NTU RGB+D 120, Kinetics Skeleton, SBU, entre otros, reportando más de 250 resultados de *accuracy*.

En el caso particular del conjunto KLHA3D102, los modelos DDNet y PA-ResGCN alcanzaron únicamente 1.96% y 8.56% de *accuracy* (porcentajes mejores al 10%), respectivamente. Estos valores bajos no corresponden a fallos de los modelos, sino a la complejidad del conjunto de datos, que contiene 102 clases con alta similitud entre acciones y pocas muestras por categoría, lo cual dificulta la generalización. En otros conjuntos de datos más extensos y balanceados, las GCN/GNN mantienen su desempeño.

Skeleton-based human action recognition via convolutional neural networks (CNN) [79]

La investigación se centra en un enfoque basado en esqueletos utilizando exclusivamente CNNs. Las secuencias de coordenadas articulares se transforman en representaciones tipo imagen que son procesadas por la red convolucional, evitando el uso de grafos. El método incluye técnicas de optimización de datos para compensar la falta de estructuras explícitas como en las GCNs.

En las pruebas con NTU-RGB+D 60, el modelo alcanzó 95% de *accuracy*, rendimiento comparable al de modelos basados en GCN, concluyendo que, con un

diseño adecuado, las CNN pueden ser una alternativa eficiente y precisa para el reconocimiento de acciones humanas basado en esqueletos.

Skeleton-based action recognition with multi-stream adaptive graph convolutional networks [80]

El estudio presenta un modelo denominado MS-AAGCN (del inglés *Multi-Stream Adaptive Attention-enhanced Graph Convolutional Network*), que refina los enfoques GCN de esqueletos. La innovación clave reside en que la topología del grafo es adaptativa, aprendida automáticamente durante el entrenamiento, en lugar de estar predefinida o fija, lo que otorga mayor flexibilidad para adaptarse a diferentes articulaciones y acciones. Además, el modelo incorpora un módulo de atención espaciotemporal (del inglés *spatial-temporal-channel attention*) que prioriza explícitamente las articulaciones, cuadros de video y características más relevantes durante la inferencia.

Este enfoque trabaja a través de múltiples flujos, su combinación permite capturar información de primer y segundo orden (coordenadas y movimiento óseo) para representar la dinámica humana. En NTU-RGB+D 60 (configuración Cross-Subject), el uso combinado de flujos de articulaciones y huesos ("*joint + bone + motion*") alcanzó 90.9% de *accuracy*, superior a las combinaciones parciales ("*joint + bone*" y "*bone + motion*") obtuvo 89.7% y 88.9% respectivamente.

An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition [81]

La investigación propone un modelo AGC-LSTM que combina GCN's y LSTM con un mecanismo de atención espaciotemporal, permitiendo resaltar articulaciones clave y mejorar la representación dinámica de las acciones. Esta arquitectura jerárquica integra *pooling* temporal para ampliar el rango de dependencias sin incrementar el costo computacional.

Evaluable en NTU RGB+D, alcanzó 89.2% de *accuracy* (*Cross-Subject*) y 95.0% (*Cross-View*). En Northwestern-UCLA obtuvo 93.3%, superando modelos previos como ST-GCN y PB-GCN.

Skeleton-Based Action Recognition with Spatial-Structural Graph Convolution [82]

El artículo presenta una arquitectura SpSt-GCN, un modelo de dos ramas diseñado para mejorar el reconocimiento de acciones. La rama espacial utiliza conexiones topológicas fijas del cuerpo humano, mientras que la rama estructural incorpora conexiones dinámicas aprendidas entre nodos periféricos, mitigando el problema de sobre suavizado común en las GCN.

El modelo fue evaluado en los conjuntos NTU RGB+D y NTU RGB+D 120. En NTU RGB+D, alcanzó 91.6% en *Cross-Subject* y 95.8% en *Cross-View*, superando métodos como EfficientGCN-B0 y CTR-GCN. En NTU RGB+D 120 obtuvo 87.8% en *Cross-Subject* y 88.8% en *Cross-View*, destacando también frente a modelos previos. Además, se reporta alta eficiencia computacional, con solo 0.48 M de parámetros y 4.90×10^9 FLOPs, permitiendo su implementación en escenarios con recursos limitados.

1.2.3.3 Modelos de atención y transformadores aplicados a HAR

SkateFormer: Skeletal-temporal transformer for human action recognition [83]

El artículo presenta un modelo de transformadores que introduce el mecanismo de atención *Skate-MSA*, que segmenta las relaciones esquelético-temporales en cuatro particiones (articulaciones físicas cercanas y lejanas, cuadros locales y globales) para modelar dependencias relevantes sin recurrir a *tokenización* global. Además, incorpora *Skate-Embedding*, que combina *embeddings* esqueléticos de entrenamiento con índices temporales fijos mediante un producto externo, fortaleciendo la codificación posicional y la representación de las secuencias.

Evaluado con los conjuntos de datos NTU RGB+D, NTU RGB+D 120 y NW-UCLA, y haciendo uso de tres configuraciones de ensamblado de modalidades (E1: solo articulaciones, E2: articulaciones + huesos, E4: cuatro modalidades combinadas), SkateFormer alcanzó 92.6% (X-Sub) y 93.0% (X-View) con modalidad E1; 93.5%, 97.0%, y 97.4% en combinación E2; y 97.8% (E4) en NTU RGB+D. En NTU RGB+D 120 obtuvo 89.4% (X-Sub) y 89.8% (X-Set) con E1, y 91.0%, 91.4%, 98.3% con E2 y E4 respectivamente. En NW-UCLA, logró 98.3% de *accuracy* en la modalidad E4. A nivel de eficiencia computacional, SkateFormer usa solo 2.03M de parámetros y aproximadamente 3.62G FLOPs, con tiempo de inferencia de 11.46 ms

3Mformer: multi-order multi-mode transformer for skeletal action recognition [84]

El estudio introduce un modelo a base de transformadores que captura relaciones entre articulaciones mediante hipergráficas. El modelo divide las secuencias de acción en bloques temporales y aplica el módulo *Higher-Order Transformer* (HoT) para generar *embeddings* basados en articulaciones individuales, pares (conexiones) y estructuras de orden superior. Estos *embeddings* se combinan con atención en modos acoplados (*channel-temporal block*, *channel-body joint*, *channel-hyper-edge*), usando dos módulos esenciales *Multi-order Pooling* (MP) y *Temporal block Pooling* (TP), que mejoran la representación espacial y temporal de los datos.

Las evaluaciones experimentales en los conjuntos de datos NTU RGB+D 60, NTU RGB+D 120, Kinetics-Skeleton y Northwestern-UCLA logran 94.8% y 98.7% de *accuracy* en *Cross-Subject* y *Cross-View* respectivamente para NTU RGB+D 60, superando al *baseline* que obtuvo 89.8% y 91.4%, respectivamente. En NTU RGB+D 120 alcanza 92.0% en *Cross-Subject* y 93.8% en *Cross-Setup*, mientras que en Kinetics-Skeleton obtiene un Top-1 *accuracy* de 48.3%, superando al *baseline* con 38.6%.

STEP CATFormer: spatial-temporal effective body-part cross attention transformer for skeleton-based action recognition [85]

Una arquitectura híbrida fue diseñada, que combina refinamiento topológico canal-wise (CTR-GCN) con múltiples módulos de atención cruzada enfocados en partes del cuerpo (del inglés *Single y Multi Body-part Cross Attention*), además de un transformador temporal dinámico (del inglés Super Dynamic Temporal Attention) para capturar patrones esqueléticos en secuencia.

En NTU RGB+D 60, STEP CATFormer alcanza una mejora de 0.8 puntos porcentuales sobre CTRGCN en *Cross-Subject* y de 0.2 puntos porcentuales en *Cross-View*. En NTU RGB+D 120, supera a CTRGCN por 1.1 puntos porcentuales en *Cross-Subject* y 0.6 puntos porcentuales en *Cross-Setup*. Incluso frente a InfoGCN mantiene una ligera ventaja de 0.1 puntos porcentuales en diversas estrategias de ensemble.

STMT: A spatial-temporal mesh transformer for MoCap-based action recognition [86]

Los autores proponen STMT, un modelo basado en *mesh* con secuencias de Motion Capture (MoCap), que modela directamente las mallas 3D sin convertirlas a esqueletos. La arquitectura incluye un transformador jerárquico con atención intra-frame e inter-frame, y se entrena usando dos tareas de aprendizaje auto-supervisado: *masked vertex modeling* y *future frame prediction*, para capturar contexto espacial y temporal a gran escala.

En las evaluaciones derivadas del conjunto NTU RGB+D a partir de estimación de poses en malla, STMT alcanzó un *top-1 accuracy* de 64.04%, superando en 1.79 puntos porcentuales al mejor modelo basado en esqueletos (CTRGCN) y en 3.44 puntos porcentuales al mejor modelo basado en *point cloud* (PSTNet).

Local and global spatial–temporal transformer for skeleton-based action recognition

El artículo presenta una arquitectura de tipo transformador denominada LG-ST (del inglés *Local and Global Spatial–Temporal Transformer*), que hace uso de datos esqueleto humano. El modelo incorpora bloques de atención espacial y temporal que trabajan de forma conjunta de manera local (capturando relaciones cercanas entre articulaciones o cuadros de video) y global (considerando dependencias a largo plazo o entre partes del cuerpo alejadas). Con esta estrategia se logra combinar la información local detallada con contexto global relevante, algo que los transformadores convencionales o las GCN por sí solas no logran con eficiencia computacional moderada.

STAR-Transformer: A spatio-temporal cross attention transformer for human action recognition [87]

Los autores proponen un modelo que integra información de video y datos de articulaciones humanas mediante un mecanismo de atención cruzada espaciotemporal. El enfoque utiliza *tokens* globales de imagen y *tokens* de articulaciones para modelar simultáneamente dependencias espaciales y temporales. Con la incorporación de dos módulos de atención, *Full Spatio-Temporal Attention (FAttn)*, procesa espacio y tiempo de manera conjunta, y *Zigzag Spatio-Temporal Attention (ZAttn)*, que mejora la captura de fases clave de las acciones.

El modelo se evaluó en los conjuntos Penn-Action, NTU RGB+D 60 y NTU RGB+D 120 y los resultados muestran un *accuracy* de 98.7% en Penn-Action, 92.0% (*Cross-Subject*) y 96.5% (*Cross-View*) en NTU RGB+D 60, y 90.3% (*Cross-Subject*) y 92.7% (*Cross-Setup*) en NTU RGB+D 120.

Frequency guidance matters: Skeletal action recognition by frequency - aware mixed transformer [88]

El artículo revisa avances en el reconocimiento de acciones humanas a partir de datos de articulaciones, enfocado en mejorar la discriminación de acciones con patrones similares. El modelo incorpora un módulo de atención sensible a la frecuencia que aplica la transformada discreta del coseno (DCT) para descomponer los movimientos articulares en componentes de alta y baja frecuencia, reforzando aquellos detalles que permiten diferenciar acciones complejas. Estas características en el dominio de frecuencia se fusionan con descriptores espaciales mediante bloques de atención mixta, mientras que un transformador temporal capta las dependencias globales entre cuadros de la secuencia.

La arquitectura se evaluó en los conjuntos NTU RGB+D 60, NTU RGB+D 120 y NW-UCLA, considerando además subgrupos de acciones clasificadas por nivel de dificultad (*Hard, Medium y Easy*). En NTU RGB+D 60 alcanzó valores de *accuracy* entre 93.0% y 96.0%, mejorando el rendimiento de modelos como SkeMixFormer, CTR-GCN e InfoGCN. En acciones difíciles elevó la precisión de 71.9% a 73.9%, en acciones de dificultad media pasó de 84.6% a 86.1%, y en acciones fáciles alcanzó 95.2%.

Interactive spatiotemporal token attention network [89]

El enfoque introduce los *Interactive Spatiotemporal Tokens (ISTs)*, una representación unificada que captura información de todas las entidades involucradas a lo largo del tiempo, evitando la necesidad de fusionar modelos o conocer de antemano la identidad de cada participante. La arquitectura fue evaluada en cuatro conjuntos de datos orientados a acciones interactivas bajo configuraciones generales donde los sujetos no mantienen un rol fijo, logrando mayor capacidad para modelar relaciones entre diferentes sujetos.

IIP-Transformer: Intra-inter-part transformer for skeleton-based action recognition [90]

Los autores exploran las relaciones jerárquicas principalmente en dos niveles, entre articulaciones dentro de una misma región del cuerpo o intra-parte y entre diferentes regiones o inter-parte. Esta organización busca reducir el ruido en los datos y obtener representaciones más discriminativas con menor complejidad computacional.

El modelo codifica las articulaciones en grupos como cabeza, torso y extremidades, aplicando un módulo de atención que captura dependencias locales y globales, complementado con técnicas de aumento de datos específicas para partes del cuerpo. Evaluado en los conjuntos NTU RGB+D 60 y NTU RGB+D 120, bajo configuraciones *Cross-Subject* y *Cross-View/Cross-Setup*, alcanzó rendimientos competitivos superando a arquitecturas como DSTA-Net con más de ocho veces menos complejidad.

1.2.4 Resumen del análisis al estado del arte

La Tabla 1.1 sintetiza los trabajos revisados, ofreciendo una visión comparativa de las principales técnicas identificadas. En ella se contrastan los enfoques metodológicos, los conjuntos de datos utilizados, las métricas de evaluación reportadas y las observaciones más relevantes de cada estudio.

Tabla 1.1 Resumen de análisis del estado del arte

Cita	Modelo	Conjunto de datos	Resultados	Notas
Laptev et al., 2005	STIP (<i>Space-Time Interest Points</i>)	KTH Weizmann	Mayor al 90% <i>accuracy</i>	Detector de esquinas 3D, <i>Bag of Words</i> + SVM
Niebles et al., 2008	<i>Bag of Visual Words</i>	KTH Hollywood2	Aprox. 83% <i>accuracy</i> (KTH)	Histograma de "palabras visuales"; eficacia en acciones simples, menor en entornos no controlados
Yamato et al., 1992	HMM (<i>Hidden Markov Model</i>)	Datos propios (golf, tenis, natación)	Reconoce acciones deportivas simples (baja resolución)	Primer modelado probabilístico temporal en HAR.
Wang et al., 2013	<i>Improved Dense Trajectories</i> (iDT)	Hollywood2 UCF101 HMDB51	Estado del arte en su época (mejor precisión con compensación de cámara)	Compensación de movimiento de cámara mediante homografías.
Dalal et al., 2006	HOF + MBH	KTH Weizmann HMDB51	MBH supera estado del arte en HMDB51 por 16 pp. HOF+MBH mejora 5 pp.	MBH elimina movimiento de cámara, resaltando movimiento de extremidades.
Liu et al., 2010	GMM (<i>Gaussian Mixture Model</i>)	KTH (modificado)	78.3% <i>accuracy</i> (71.2% con HMM)	Segmentación temporal por clasificación en ventana deslizante.
Blank et al., 2005	<i>Shape Context</i> en siluetas	Secuencias sintéticas (variando vista, ruido)	Mejor precisión que momentos de Hu y Fourier (error angular 150° vs 165°)	Representación de acciones con base en siluetas.

Tabla 1.1 Resumen de análisis del estado del arte (Continuación)

Cita	Modelo	Conjunto de datos	Resultados	Notas
Kläser et al., 2008	HOG3D (<i>Histogram of Oriented 3D Gradients</i>)	KTH <i>Weizmann</i>	Weizmann: 90.7% <i>accuracy</i> ; KTH: 89.0%	Captura apariencia y dinámica.
Scovanner et al., 2007	3D SIFT + pirámide espaciotemporal	KTH <i>UCF Sports</i>	3D SIFT supera métodos 2D en dichas acciones	Extensión 3D de SIFT para la captura de variaciones temporales mejor que descriptores 2D.
Sun et al., 2011	Trayectorias alineadas + BoW	<i>UCF Sports</i> <i>Hollywood2</i>	UCF: 88.0% <i>accuracy</i> (vs 84.2% MBH solo) Hollywood2: 58.2% mAP	Alineación mejora coherencia temporal de descriptores.
Le et al., 2025	Dual LSTM+CNN (esqueletos 3D/2D)	JHMDB Florence 3D SBU PennAction	JHMDB: 73.3%, Florence: 97.7%, SBU: 95.2%, Penn: 94.0% <i>accuracy</i>	Arquitectura de dos flujos fusionados.
Elnady et al., 2025	YOLO + LSTM híbrido	UCF101 KTH <i>Weizmann</i> IXMAS	UCF101: aprox. 96% acc, KTH: aprox 99% acc, Weizmann: 100% acc, IXMAS: 98% acc	YOLO extrae regiones clave y LSTM modela la dinámica temporal.
Yin et al., 2025	LSTM + CNN deformable (ligero)	Comparaciones en múltiples sets	Precisión comparable o superior a arquitecturas más pesadas	LSTM captura dinámica global y CNN deformable se adapta a variaciones espaciales.

Tabla 1.1 Resumen de análisis del estado del arte (Continuación)

Cita	Modelo	Conjunto de datos	Resultados	Notas
Joudaki et al., 2025	Conv-RBM + LSTM (selección de <i>frames</i>)	KTH UCF Sports HMDB51	KTH: 97.3% Acc UCF Sports: 94.8% Acc HMDB51: 81.5% accuracy	Combina extracción de características con modelado secuencial.
Mansouri et al., 2024	<i>Semantic-Guided GCN</i> (mejorado)	NTU RGB+D 60 NTU RGB+D 120	NTU60: 89.7% (XSub), 95.0% (XView); NTU120: 84.6% (XSub), 85.8% (XSetup)	Posee un mecanismo de atención que resalta nodos importantes.
Nguyen et al., 2023	<i>Revisión Deep Learning (3D skeleton)</i>	NTU RGB+D 60 NTU RGB+D 120 Kinetics SBU	250+ resultados comparados, GCN/GNN lideran accuracy	Analiza más de 250 métricas de accuracy y destaca alta precisión de GCN con problemas de eficiencia y generalización.
Ali et al., 2023	CNN pura (esqueletos 2D/3D)	NTU RGB+D 60	95% accuracy (NTU60)	Demuestra que una CNN bien diseñada puede igualar GCN en esqueletos.
Shi et al., 2019	MS-AAGCN (<i>Multi-Stream Adaptive GCN</i>)	NTU RGB+D 60	90.9% accuracy (XSub, <i>joint+bone+motion</i>)	Módulo de atención resalta articulaciones por cuadros clave.
Si et al., 2019	AGC-LSTM (<i>Atención+GCN+LSTM</i>)	NTU RGB+D 60 Northwestern-UCLA	NTU60: 89.2% (XSub), 95.0% (XView); NW-UCLA: 93.3%	Arquitectura jerárquica con <i>pooling</i> temporal.

Tabla 1.1 Resumen de análisis del estado del arte (Continuación)

Cita	Modelo	Conjunto de datos	Resultados	Notas
Wang et al., 2024	SpStGCN (<i>Spatial-Structural GCN</i>)	NTU RGB+D 60 NTU RGB+D 120	NTU60: 91.6% (XSub), 95.8% (XView) NTU120: 87.8% (XSub), 88.8% (XView)	Mitiga sobre suavizado con conexiones dinámicas entre nodos periféricos.
Do et al., 2024	SkateFormer	NTU RGB+D 60 NTU RGB+D 120 NW-UCLA	NTU60: aprox. 97.8% <i>acc</i> NTU120: aprox. 91.4% <i>acc</i> NW-UCLA: 98.3% <i>acc</i>	Segmenta relaciones esqueleto-tiempo para atención focalizada.
Wang et al., 2023	3MFormer (<i>Multi-Order Multi-Mode TF</i>)	NTU RGB+D 60 NTU RGB+D 120 Kinetics NW-UCLA	NTU60: 94.8% (XSub), 98.7% (XView) NTU120: 92.0% (XSub), 93.8% (XSetup) Kinetics: 48.3% top-1	Emplea <i>transformers</i> para capturar patrones conjuntos, pares y grupos.
Nguyen et al., 2023	STEP-CATFormer	NTU RGB+D 60 NTU RGB+D 120	(XSub) +0.8 pp y +1.1 pp (XView) +0.2 pp y +0.6 pp sobre CTRGCN en NTU60 y 120 respectivamente	Integración de atención en partes del cuerpo (<i>single/multi body-part</i>) con refinamiento por canal (CTRGCN).
Zhu et al., 2023	STMT (<i>Mesh transformer</i>)	(NTU RGB+D derivado)	NTU (malla): 64.0% top-1 (mejora +1.79% sobre CTRGCN esqueleto)	Modela directamente la malla 3D sin convertir a esqueleto.
Liu et al., 2025	LGST	NTU RGB+D 60 NTU RGB+D 120	Resultados similares a SOTA	Combina detalle local y contexto global en un solo modelo.

Tabla 1.1 Resumen de análisis del estado del arte (Continuación)

Cita	Modelo	Conjunto de datos	Resultados	Notas
Ahn et al., 2023	STARTransformer	NTU RGB+D 60 NTU RGB+D 120	PennAction: 98.7% NTU60: 92.0% (XSub), 96.5% (XView) NTU120: 90.3% (XSub), 92.7% (XSetup)	Fusiona información visual y esquelética mediante atención espaciotemporal conjunta.
Wu et al., 2024	FA-MixFormer (<i>Frequency-Aware TF</i>)	NTU RGB+D 60 NTU RGB+D 120 NW-UCLA	NTU60: 93–96% Acc (según partición) +2 pp en acciones difíciles (71.9 → 73.9%)	Destaca detalles discriminativos en dominio frecuencia para distinguir acciones similares.
Wen et al., 2023	ISTNet (<i>Interactive ST Tokens</i>)	Conjuntos de interacción multi persona	Mejora la modelación de relaciones multi-sujeto (rinde mejor sin rol fijo)	Representación unificada de entidades evita arquitecturas separadas.
Wang et al., 2023	IIP-Transformer	NTU RGB+D 60 NTU RGB+D 120	(supera a DSTANet) con complejidad ocho veces menor	Organización por partes reduce ruido y captura dependencias entre e intra-parte con atención especializada.

1.2.5 Propuesta de solución

Está integrada por un flujo de procesamiento que inicia con la estimación de poses humanas a partir de secuencias de video, generando representaciones esqueléticas compactas y estructuradas. A partir de estas, se construye un grafo espaciotemporal que modela la dinámica del movimiento corporal, lo que permite capturar tanto la postura instantánea como su evolución en el tiempo.

Sobre esta representación se entrena y despliega una red ST-GCN, diseñada para identificar un conjunto predefinido de acciones. El modelo se complementa con un proceso de normalización y filtrado que mejora la robustez frente a variaciones individuales, además de una estrategia de votación en ventanas temporales para estabilizar la inferencia en secuencias prolongadas.

La arquitectura planteada busca reducir la ambigüedad entre acciones con trayectorias similares y mejorar la generalización en escenarios controlados. De igual forma, se contempla la integración del modelo en un entorno de inferencia, con capacidad de procesar transmisiones de video mediante RTSP y registrar evidencias de las detecciones realizadas.

Con este enfoque, la solución articula estimación de pose, modelado esquelético y aprendizaje profundo en un esquema unificado, orientado a garantizar precisión, estabilidad y eficiencia en el reconocimiento automático de acciones humanas.

1.3 Discusión

El análisis inicial del problema permitió identificar una brecha clara entre las capacidades de los enfoques tradicionales y los requerimientos de las aplicaciones que exigen reconocimiento inmediato y confiable de acciones humanas.

Estos retos muestran que no basta con incrementar la complejidad de los algoritmos o utilizar descriptores visuales más sofisticados, lo que se necesita es replantear la forma en que se representa y procesa la información del cuerpo humano. De ahí surge la importancia de enfoques basados en representaciones esqueléticas.

De este modo, se establece una base conceptual para el desarrollo de la propuesta planteada, la cual se examina con mayor detalle en los capítulos posteriores, partiendo de la revisión de fundamentos teóricos y soluciones previas que han buscado abordar el mismo desafío desde diferentes perspectivas.

1.4 Estructura del documento

El presente trabajo de tesis se divide en cinco capítulos con la distribución siguiente:

El Capítulo 1 establece el marco de referencia de la investigación. También se incluyen antecedentes y un panorama del estado del arte que fundamentan el tema.

El Capítulo 2 presenta la solución propuesta. Se introduce el modelo conceptual y se detallan sus elementos principales que guían el desarrollo de la propuesta.

El Capítulo 3 se centra en el diseño e implementación del sistema, abarcando desde la adquisición de los datos hasta el despliegue del modelo en sistemas de video.

El Capítulo 4 expone los experimentos realizados y los resultados alcanzados, describiendo las condiciones de validación, el desempeño del modelo bajo distintos escenarios.

Finalmente, el Capítulo 5 reúne las conclusiones del estudio, también se presentan los productos derivados del proyecto y se plantean posibles líneas de trabajo futuro.

Capítulo 2

Propuesta de solución

Este apartado presenta la estrategia pensada para dar respuesta al problema planteado, articulando los elementos conceptuales y técnicos que sustentan la propuesta. Se describen los criterios de diseño que guiaron la construcción del modelo, así como los componentes que lo integran y su interacción dentro de un flujo coherente de procesamiento.

Asimismo, se detallan los métodos empleados para la adquisición, preprocesamiento y representación de los datos, junto con la justificación de las decisiones adoptadas en cada etapa. Finalmente, se expone la arquitectura implementada, enfatizando cómo esta contribuye a resolver las limitaciones identificadas en enfoques previos y a alcanzar los objetivos de la investigación.

2.1 Planteamiento de solución

Las limitaciones observadas en los enfoques tradicionales evidencian la necesidad de un sistema capaz de abstraer factores externos y centrarse en la dinámica corporal. Con este objetivo, la solución propuesta integra la estimación de pose a partir de secuencias de video y su representación esquelética como punto de partida para el análisis.

Sobre esta base se implementa una red ST-GCN, diseñada para capturar tanto la configuración espacial de las articulaciones como la evolución temporal de sus movimientos. Esta estrategia permite diferenciar acciones con similitudes posturales y mejorar la discriminación en escenarios controlados, superando métodos basados únicamente en imágenes RGB.

2.1.1 Objetivo general

Desarrollar un modelo de detección e identificación de patrones de movimiento corporal asociados a acciones humanas en secuencias de imágenes, considerando atributos como la trayectoria, duración y transición de las posturas, utilizando estimación de pose.

2.1.2 Objetivos específicos

- Estudiar diferentes enfoques a partir del análisis del estado del arte y las principales tendencias reportadas en la literatura científica reciente sobre el reconocimiento de acciones humanas.
- Analizar herramientas, arquitecturas y técnicas de representación mayormente utilizadas, considerando su aplicabilidad en distintos contextos y entornos.
- Adaptar un conjunto de datos personalizado basado en la topología de NTU RGB+D, asegurando diversidad y consistencia estructural en las secuencias.
- Implementar un modelo computacional capaz de identificar acciones humanas, a partir de los enfoques seleccionados en la fase de análisis.
- Evaluar el rendimiento del modelo desarrollado para determinar su eficacia y confiabilidad.

2.1.3 Alcances

- El estudio se centra en el reconocimiento de cinco acciones humanas específicas, seleccionadas para contextos de análisis de comportamiento en entornos controlados.
- Las acciones son identificadas a partir de secuencias de imágenes procesadas mediante estimación de pose, utilizando exclusivamente información esquelética del cuerpo humano.
- Se emplea una arquitectura ST-GCN adaptada y entrenada desde cero, que modela la estructura espaciotemporal del cuerpo humano en forma de grafo.
- La evaluación del desempeño se realiza empleando métricas clásicas del reconocimiento de patrones.
- El sistema es probado en condiciones controladas, lo cual permite una validación estructurada del funcionamiento del modelo con secuencias de datos segmentadas

2.1.4 Limitaciones

- El análisis se restringe al uso de representaciones esqueléticas del cuerpo humano; no se consideran características visuales adicionales como texturas, expresiones faciales o fondo.
- El reconocimiento está limitado a cinco clases de acción previamente definidas; no se contempla la detección de acciones fuera de este conjunto.
- Las pruebas se realizan en entornos controlados, por lo que el modelo no ha sido validado en escenarios no estructurados o con condiciones ambientales cambiantes.

- La capacidad de detección depende de la calidad de la estimación de pose; errores en la localización de articulaciones pueden afectar el desempeño del modelo.
- El sistema está diseñado para una sola persona en escena, sin considerar interacciones entre múltiples personas.

2.1.5 Justificación

El monitoreo de actividades humanas representa un reto en diferentes ámbitos. Es necesario detectar situaciones de riesgo como caídas o pérdida de movilidad de manera temprana y confiable.

Los sistemas tradicionales basados en sensores portátiles presentan limitaciones, ya que requieren colocación por parte del usuario, resultan intrusivos y pueden perderse o dejar de funcionar adecuadamente [91]. En contraste, un sistema visual fundamentado en la estimación de pose esquelética ofrece una alternativa no invasiva.

Esta modalidad preserva la privacidad al abstraer características personales y permite la observación continua sin interacción directa con la persona. Bajo estas condiciones, se favorece la implementación de esquemas de asistencia remota y monitoreo cotidiano en entornos estructurados y de bajo riesgo, lo que constituye un aporte para la prevención y la respuesta oportuna[92].

Desde la perspectiva técnica, la variabilidad en posturas y movimientos, así como la presencia de oclusiones en ambientes interiores, representan desafíos para los sistemas de reconocimiento de acciones.

Para enfrentarlo, se han propuesto arquitecturas con atención temporal adaptativa que priorizan los cuadros más relevantes para la decisión [93]. Adicionalmente, los modelos de convolución en grafos espaciotemporales han mostrado eficacia al capturar la cinemática del movimiento y mantener su desempeño en entornos controlados [94], [95].

En este sentido, el uso de representaciones esqueléticas se justifica por su capacidad de abstraer factores externos como iluminación o vestimenta. Estos modelos se enfocan en las relaciones dinámicas entre articulaciones, lo que contribuye a la estabilidad del sistema y a la precisión en condiciones experimentales.

2.1.6 Beneficios

La propuesta planteada ofrece beneficios tanto en el ámbito metodológico como en el operativo. En primer lugar, el uso de representaciones esqueléticas disminuye la dependencia de factores externos, permitiendo que la clasificación se concentre completamente en la ubicación de los puntos del cuerpo. Esta abstracción contribuye a mejorar la precisión y la generalización del sistema.

De manera complementaria, la incorporación de una red ST-GCN posibilita diferenciar acciones con trayectorias posturales similares, reduciendo la ambigüedad entre clases que en métodos convencionales suelen generar confusión. Esto se traduce en un rendimiento superior en métricas de evaluación, con mejores valores de exactitud y *F1-Score*.

El modelo desarrollado se orienta específicamente al reconocimiento de cinco acciones definidas, alcanzando resultados satisfactorios en precisión y estabilidad. En caso de requerir la identificación de un conjunto distinto de actividades, basta con seleccionar otras cinco acciones de interés y realizar el entrenamiento correspondiente, de modo que el sistema pueda adaptarse a nuevos escenarios manteniendo la misma estructura metodológica.

Finalmente, el sistema cuenta con la capacidad de operar en tiempo real mediante transmisión de video, lo que garantiza baja latencia y estabilidad en escenarios controlados.

2.2 Modelo conceptual de la solución

El modelo conceptual de la solución se visualiza a través de dos fases que se mencionan a continuación.

En la primera fase se representa el proceso seguido para entrenar el modelo ST-GCN. A partir de un conjunto de datos previamente etiquetado, se realiza la preparación de las secuencias de video, la extracción de coordenadas esqueléticas y la construcción de grafos que representan la dinámica de las articulaciones humanas. Estos grafos constituyen la entrada al modelo, el cual, mediante sus bloques de convoluciones espaciales y temporales, abstrae patrones espaciotemporales del movimiento humano. El resultado de esta fase es un modelo entrenado que reconoce distintas acciones de un grupo seleccionado a partir de representaciones esqueléticas (ver Figura 2.1).

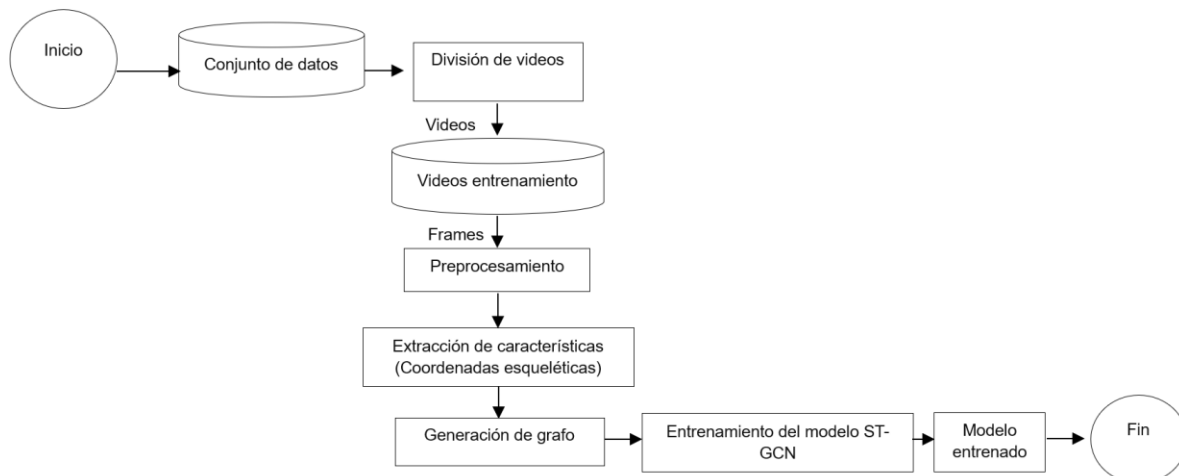


Figura 2.1 Modelo conceptual del entrenamiento de la propuesta de solución (elaboración propia)

En la segunda fase se muestra la aplicación del modelo ya entrenado sobre datos nuevos obtenidos desde una cámara. Cada *frame* del video es procesado para extraer las coordenadas esqueléticas y generar el grafo correspondiente, el cual es contrastado con los patrones previamente aprendidos por el modelo. La red aplica la función *softmax* para obtener probabilidades de clase y, finalmente, se selecciona la acción más probable. De esta manera, el sistema es capaz de emitir una etiqueta de acción y mostrar los resultados en la interfaz de usuario, permitiendo el reconocimiento (ver Figura 2.2).

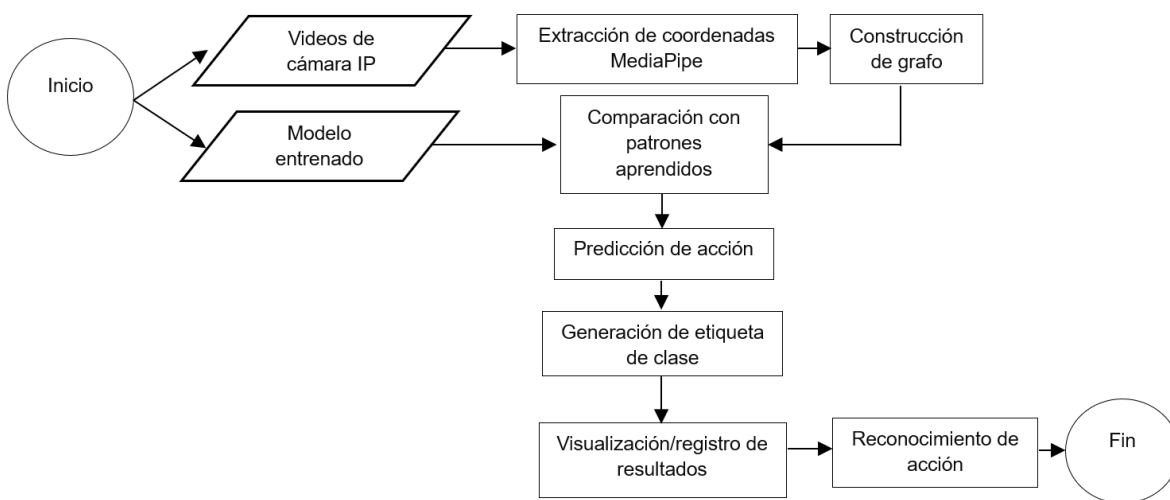


Figura 2.2 Modelo conceptual de la validación de la propuesta de solución (elaboración propia)

Esta combinación permite aprovechar la información espacial (la estructura del cuerpo en cada instante) y temporal (la evolución del movimiento a lo largo del tiempo) dentro de un mismo marco de trabajo.

En términos generales, la arquitectura transforma datos de video en una predicción de acción logrando que el modelo entrenado pueda inferir la acción observada en nuevos videos.

2.3 Descripción de los elementos del modelo conceptual

A continuación, se describen los componentes principales del modelo conceptual ilustrado en la Sección 2.2, destacando su propósito y funcionamiento dentro de la arquitectura propuesta.

2.3.1 Redes neuronales gráficas espaciales y temporales

Las redes neuronales gráficas espaciotemporales han emergido como una solución basada en esqueletos para el reconocimiento de acciones humanas. En lugar de tratar las coordenadas de las articulaciones como vectores independientes en el tiempo o imágenes pseudo esqueléticas, estos modelos explotan directamente la estructura de grafo del cuerpo, representando articulaciones y sus conexiones y aprendiendo dependencias espaciales y temporales de manera conjunta [96].

Un grafo permite representar cada articulación como un nodo y cada hueso como una arista, conservando explícitamente la topología del esqueleto. Al extender el grafo en la dimensión temporal (conexiones entre la misma articulación en cuadros sucesivos) se obtiene un grafo espacio-temporal adecuado para extraer características dinámicas del movimiento.

En esencia, las GCN espacio-temporales aplican “convoluciones” sobre ese grafo 3D para capturar, a la vez, relaciones entre articulaciones y su evolución en el tiempo, variantes recientes integran componentes adicionales o híbridos con CNN para mejorar la discriminación. Como alternativa basada en pose, existen representaciones que codifican explícitamente la dinámica de la pose sin operar directamente con GCN, lo cual ilustra otras formas de explotar la información esquelética [97].

2.3.2 Estructura general del modelo ST-GCN

Las redes convolucionales en grafos espaciales-temporales extienden las GCN para modelar secuencias dinámicas en datos esqueléticos. En este enfoque, una acción se representa como un grafo donde los nodos corresponden a las articulaciones del cuerpo en cada instante temporal y las aristas integran tanto las conexiones anatómicas (intra-frame) como las dependencias temporales entre articulaciones homólogas en cuadros consecutivos (inter-frame) [98].

$$G = (V, E)$$

Donde los nodos V corresponden a las articulaciones del cuerpo humano en cada instante temporal, y las aristas E integran tanto las conexiones anatómicas (*intra-frame*) como las dependencias temporales entre articulaciones homólogas en *frames* consecutivos (*inter-frame*) [99].

La convolución en grafos se define propagando información desde cada nodo hacia su vecindario. Para el caso espacial, el vecindario se construye a partir de la topología corporal (conexiones hombro–codo–muñeca), mientras que en el caso temporal cada nodo se enlaza con su representación en instantes previos y posteriores, lo que permite capturar la continuidad del movimiento [100]. Estas operaciones posibilitan extraer patrones locales y globales de co-movimiento, reflejando de manera conjunta la estructura corporal y su dinámica [101].

A diferencia de arquitecturas previas, como las CNN aplicadas sobre mapas de calor o las RNN aplicadas sobre secuencias vectoriales, las ST-GCN integran en un solo marco el espacio y el tiempo, evitando la pérdida de información estructural. El modelo aprende filtros que reconocen configuraciones y transiciones específicas de articulaciones, lo que resulta especialmente adecuado para el reconocimiento de acciones humanas a partir de representaciones esqueléticas [102].

El principio central es que las relaciones espaciales y temporales no se fijan de antemano, sino que se parametrizan y optimizan durante el entrenamiento, permitiendo que el modelo se adapte a distintos estilos de movimiento y variaciones individuales. Esta capacidad lo convierte en un enfoque robusto y eficiente para tareas de reconocimiento de acciones en secuencias de video [103].

En el capítulo 3 se detallan las adaptaciones específicas de este modelo a la propuesta desarrollada en este trabajo, incluyendo la definición de parámetros, normalización y estrategias de entrenamiento.

2.3.3 Representación del cuerpo como grafo y matriz de adyacencia

Un paso importante en el uso de redes ST-GCN es la representación del cuerpo humano como un grafo. Para construir el grafo del esqueleto, se toman las V articulaciones detectadas del cuerpo (por ejemplo, 25 articulaciones) como nodos, y se conectan con aristas aquellos nodos que corresponden a huesos físicamente conectados (por ejemplo, codo unido a hombro, rodilla unida a cadera, etc.). Esta topología inicial suele basarse en conocimiento anatómico. La conectividad del grafo se representa mediante una matriz de adyacencia A de tamaño $V \times V$ que codifica la conectividad del grafo [104].

$$A \in R^{V \times V}$$

Donde $A_{ij} = 1$ si la articulación i está conectada con la articulación j y $A_{ij} = 0$ en caso contrario [105].

En implementaciones basadas en grafos para reconocimiento de acciones, la convolución en grafo se define propagando información desde cada nodo hacia su vecindario. En el plano espacial, el vecindario se induce por la topología corporal (hombro–codo–muñeca), en el plano temporal, cada nodo se enlaza con su representación en instantes adyacentes, lo que permite capturar la continuidad del movimiento y extraer patrones locales y globales de movimiento

Dado que una matriz de adyacencia fija (solo anatómica) puede no ser la mejor al ignorar relaciones no locales (coordinación entre ambas manos), trabajos posteriores introducen grafos adaptativos en los que parte de la conectividad se aprende a partir de los datos, permitiendo incorporar aristas “no explícitas” cuando mejoran la predicción y estableciendo relaciones globales entre articulaciones distantes [106].

En suma, la representación del cuerpo como grafo puede iniciarse con la estructura esquelética tradicional (matriz de adyacencia fija) y refinarse mediante términos aprendidos que capturan correlaciones relevantes más allá de los huesos directos; la matriz efectiva usada por el modelo puede entenderse como una combinación de la estructura fija con componentes aprendidos durante el entrenamiento para maximizar el rendimiento en la clasificación.

2.3.4 Alternativas recientes a ST-GCN y comparación

Una primera tendencia es la introducción de grafos adaptativos, en los que la matriz de adyacencia no se limita a la conectividad anatómica fija, sino que incorpora componentes aprendidos que permiten capturar relaciones globales y dependencias específicas de cada muestra. Esta adaptabilidad mejora la representación de patrones complejos en acciones con interacciones poco evidentes entre articulaciones distantes [107].

Otra dirección relevante explora topologías estructuradas (multi-flujo/jerárquicas) que diferencian cómo se propaga la información entre regiones del cuerpo, útil cuando la dirección de la acción y la organización funcional del movimiento incide en la decisión [108].

En tercer lugar, se han propuesto estructuras dinámicas por secuencia y alternativas basadas en atención que actualizan conectividades de forma *data-driven* (o

directamente prescinden del grafo fijo mediante *transformers*), lo que favorece dependencias de largo alcance [109].

Un cuarto grupo de propuestas se centra en la optimización computacional de las operaciones de grafo. Técnicas como el desplazamiento de características entre nodos, en lugar de multiplicaciones matriciales completas, han permitido reducir el número de operaciones sin comprometer la precisión, lo que las hace adecuadas para aplicaciones en tiempo real o en dispositivos con recursos limitados [110].

Finalmente, se han explorado enfoques de refinamiento topológico por canal y mecanismos de atención global. El refinamiento por canal asigna diferentes matrices de adyacencia a cada dimensión de característica, incrementando la expresividad del modelo, mientras que la integración de bloques de atención, inspirados en arquitecturas *transformer*, permite representar dependencias espaciales y temporales de largo alcance sin restringirse a una estructura de grafo fija [111].

En conjunto, estas alternativas muestran que la evolución de ST-GCN se ha orientado hacia un mayor grado de adaptabilidad, la incorporación de información contextual más rica y la optimización para distintos escenarios de despliegue.

2.3.5 Integración conceptual de secuencias de video

Al llevar estas redes a la práctica, es crucial entender cómo integrar y procesar secuencias de video reales para alimentar el modelo de reconocimiento de acciones. Una secuencia de video puede verse como una serie temporal de *frames* (imágenes estáticas) que capturan el movimiento.

En el caso de reconocimiento basado en esqueletos, cada *frame* primero debe convertirse en una representación de esqueleto (mediante algoritmos de estimación de pose) y luego ensamblarse en una secuencia estructurada para la red neuronal. Esta sección describe la estructura temporal de entrada mediante *buffers* de *frames*, el formato de los datos de entrada convertido a tensores, y consideraciones para la inferencia usando RTSP.

2.3.6 Estructura temporal de entrada y *buffers* de cuadros

Para que un modelo aprenda patrones temporales, necesita recibir varios *frames* consecutivos como entrada (por ejemplo, una acción como *saltar* no se reconoce en un solo *frame*, sino en la secuencia del movimiento). Una forma común de manejar esto es mediante una ventana deslizante de *frames*, implementada con un *buffer circular*. El concepto de *buffer de cuadros* consiste en almacenar los últimos

T frames capturados y actualizarlos conforme llega un nuevo *frame*, descartando el más antiguo cuando el *buffer* está lleno [112].

De este modo, en cada instante se dispone de un fragmento temporal de longitud fija (por ejemplo, $T = 30$ cuadros corresponde a 1 segundo de video a 30 FPS) para alimentar al modelo de inferencia en cualquier momento. Este esquema de ventana deslizante temporal permite procesar videos de duración arbitraria, es decir, el modelo clasifica acciones sobre la ventana actual y luego la ventana avanza un paso (un *frame*) para la siguiente predicción.

Diversos trabajos analizan el tamaño de ventana y su efecto en precisión y latencia, usando deslizamientos de un frame y ventanas cortas para reactividad, o más largas para mayor contexto, según el dominio y la arquitectura empleada [113]. En general, este enfoque *divide* la secuencia continua en segmentos temporales solapados que el modelo puede manejar, permitiendo detección continua de acciones.

En escenarios estrictamente online, puede añadirse un pequeño *buffer* futuro cercano para estabilizar la predicción sin penalizar en exceso la latencia, estrategia reportada en marcos de reconocimiento continuo [114].

Para sistemas basados en pose y *pipelines* ligeros, la eficiencia del preprocesamiento (estimación de pose optimizada) ayuda a mantener T y el *stride* de la ventana sin degradar tiempos de respuesta en despliegues reales.

2.3.7 Formato de entrada y conversión a tensores

Una vez definidos los segmentos temporales (ventanas de *frames*), es necesario convertir estos datos en un tensor numérico apropiado para la entrada de la red. En el caso de modelos basados en esqueletos como ST-GCN y sus variantes, la convención es estructurar los datos en un arreglo de cinco dimensiones (N, C, T, V, M) .

Aquí, N corresponde al número de secuencias en el *batch* (por ejemplo, cuántos clips o ventanas se procesan juntos), C es el número de *canales* de característica por cada articulación ($C = 3$ coordenadas (x, y, z) o (x, y) más alguna información adicional como confianza), T es la longitud temporal de la secuencia (número de *frames* en la ventana), V es el número de nodos o articulaciones en el esqueleto (por ejemplo $V = 25$) y M es el número de sujetos en la escena (por ejemplo $M = 1$ para acciones de una sola persona, o $M = 2$ si se consideran interacciones entre dos personas) [115].

En la práctica, cuando se obtiene un *frame* de video, primero se aplica un modelo de estimación de pose para extraer las coordenadas de las articulaciones del cuerpo. Esas coordenadas por cada articulación forman un vector de tamaño C . Al procesar los *frames* T sucesivos, se construye una matriz de tamaño CT para cada articulación. Finalmente, al considerar todas las articulaciones V , se organiza la información en un tensor de dimensión CTV .

Durante la etapa de pre-procesamiento, puede ser necesario normalizar las coordenadas (por ejemplo, trasladar el esqueleto al centro o escalar según la distancia de cámara) para garantizar coherencia espacial a través de diferentes secuencias [116]. También es común reordenar las articulaciones en un índice consistente con la definición del grafo esperado por el modelo.

2.3.8 Implementación de inferencia con RTSP

Real Time Streaming Protocol es un protocolo para transmitir video en tiempo real, utilizado por muchas cámaras IP y sistemas de CCTV. Integrar el modelo de reconocimiento de acciones con una fuente RTSP implica primero capturar y decodificar el flujo de video en *frames*. Para ello, es frecuente usar bibliotecas de visión como *OpenCV*, que provee mecanismos para leer *entradas* RTSP de forma similar a como se leería un archivo de video [117].

En una implementación sencilla, el sistema corre en un bucle continuo, donde la lectura del *frame* RTSP y la inferencia deben estar sincronizadas con la tasa de *frames* (FPS), tratando de mantener una latencia baja. Muchas veces, para lograr predicciones lo más cercano a una interpretación en tiempo real, se utilizan optimizaciones como procesamiento en paralelo (mientras la red está infiriendo sobre el *frame* t , ya se va decodificando el *frame* $t + 1$ en otro hilo).

Asimismo, es importante manejar la posible pérdida de *frames* o retrasos. Un búfer interno de pocos *frames* puede ayudar a absorber variaciones en el tiempo de cómputo, asegurando que siempre haya datos listos para el modelo sin congelar la reproducción.

Otra consideración es que, al operar en un flujo continuo, las predicciones del modelo pueden variar en cada *frame*. Por ello suele combinarse la inferencia con técnicas de estabilización (como se menciona en la sección 2.10) para evitar saltos erráticos en la clasificación de la acción.

Los resultados pueden mostrarse sobre el video (marcando la acción detectada) y/o enviarse a otro sistema. Es importante destacar que, dado que RTSP añade cierta

latencia de transmisión, el sistema en su conjunto suele operar con un ligero retraso respecto al evento (con un intervalo aproximado de 1-2 segundos), pero esto es generalmente aceptable en aplicaciones de monitoreo y se compensa con la *ventana temporal* que da contexto para decisiones más robustas [118].

2.3.9 Técnicas de estabilización mediante votación

Las predicciones realizadas cuadro a cuadro en el reconocimiento de acciones a partir de secuencias de video, suelen presentar fluctuaciones, generando resultados inconsistentes. Este problema se debe a que las redes neuronales, al procesar ventanas de entrada limitadas, pueden verse afectadas por ruido, posturas intermedias o movimientos ambiguos, lo que ocasiona cambios abruptos en la clasificación. Para mitigar esta inestabilidad, se emplean técnicas de estabilización basadas en votación, cuyo objetivo es reforzar la coherencia temporal de las predicciones [119].

El enfoque más simple corresponde al *majority voting*, donde se selecciona como etiqueta final aquella que aparece con mayor frecuencia dentro de una ventana temporal [120]. Una variante de este método es el *weighted voting*, en el que cada predicción se pondera por la confianza asignada por el clasificador, favoreciendo aquellas decisiones con mayor probabilidad [121].

Otra estrategia consiste en aplicar suavizado temporal mediante filtros o promedios móviles, de modo que la salida final no dependa de un único frame, sino de un consenso construido a lo largo de varios instantes [122].

Estas técnicas no modifican la arquitectura del modelo de reconocimiento, sino que actúan como un post-procesamiento que mejora la robustez de los resultados. Su principal ventaja es que incrementan la estabilidad de las etiquetas predichas sin incrementar significativamente el costo computacional, lo que resulta adecuado para aplicaciones en tiempo real.

En el capítulo 3 se describe cómo estas estrategias se integran en la propuesta de este trabajo, mientras que en el capítulo 4 se evalúa su impacto sobre la estabilidad y la precisión de las predicciones.

2.4 Variables y criterios de control

El planteamiento de la solución considera un conjunto de variables y criterios de control que aseguran la validez de los experimentos y la estabilidad del sistema en su fase de operación. Estos elementos permiten establecer condiciones homogéneas de comparación y delimitan el comportamiento esperado del modelo.

Entre los principales aspectos de control se incluyen:

- Hiperparámetros de entrenamiento, tales como número de épocas, tamaño de lote, tasa de aprendizaje y longitud de la secuencia temporal utilizada como entrada. Estos factores condicionan la convergencia del modelo y su capacidad de generalización (véase sección 3.2.6).
- Partición y balanceo del conjunto de datos, asegurando que las clases de acciones estén equilibradas y que exista una separación consistente entre entrenamiento, validación y prueba, lo que previene sesgos en la evaluación (véase sección 4.3).
- Normalización y preprocesamiento de poses, que regulan la consistencia espacial y temporal de los esqueletos, reduciendo el efecto de variaciones individuales y mejorando la comparabilidad entre secuencias (véase sección 3.2.2).
- Estrategias de inferencia, como la votación en ventanas deslizantes, que funcionan como criterio de estabilización frente a fluctuaciones en predicciones cuadro a cuadro.

En conjunto, estas variables de control permiten que la solución propuesta sea evaluable bajo condiciones definidas, al mismo tiempo que ofrecen un marco de referencia para futuros ajustes y extensiones del modelo.

2.5 Discusión y motivación del enfoque adoptado

La literatura muestra una transición desde técnicas tradicionales con descriptores manuales hacia modelos profundos que aprenden representaciones directamente de los datos. En escenarios controlados, los métodos clásicos (puntos de interés y gradientes espacio-temporales) ofrecieron avances relevantes en su época, sin embargo, su rendimiento disminuye significativamente ante variaciones de vista, iluminación o fondo, y su costo computacional aumenta al incorporar una dinámica más detallada del movimiento [123].

Posteriormente, aproximaciones como *Improved Dense Trajectories* incorporaron información del flujo óptico, alcanzando mejoras de *benchmarks*, aunque con un elevado costo computacional [124].

La implementación de las redes neuronales profundas permitió superar varias de estas limitaciones. Modelos basados en CNN bidimensionales capturaron patrones espaciales en mapas de calor, mientras que arquitecturas recurrentes como LSTM extendieron el análisis a dependencias temporales.

No obstante, estas aproximaciones tienden a perder la estructura anatómica subyacente del cuerpo humano, lo que reduce su capacidad de generalización en contextos con alta variabilidad [125].

En este marco, los enfoques basados en grafos, y particularmente las redes ST-GCN, han logrado ser efectivos al integrar en un mismo esquema las relaciones espaciales entre articulaciones y las dinámicas temporales de movimiento. Esta característica les permite representar de manera natural la secuencia esquelética y adaptarse a diferentes estilos de movimiento sin necesidad de diseñar manualmente descriptores [126].

La motivación para adoptar este enfoque en el presente trabajo radica en tres aspectos principales. Primero, la robustez de ST-GCN frente a variaciones en apariencia, ya que opera sobre representaciones esqueléticas. Segundo, su equilibrio entre precisión y eficiencia computacional. Tercero, la posibilidad de integrar mecanismos adicionales de estabilización mediante votación, con el fin de mitigar fluctuaciones en las predicciones y asegurar mayor coherencia temporal en la salida del sistema.

De esta manera, el análisis de las soluciones previas sustenta la elección de ST-GCN como núcleo metodológico de la propuesta, reforzado con técnicas de estabilización para garantizar un desempeño confiable en entornos prácticos.

Capítulo 3

Diseño e implementación de la propuesta de solución

El desarrollo de la implementación se orienta a garantizar que cada uno de los módulos implementados cumpla criterios de eficiencia, precisión y adaptabilidad, permitiendo que el sistema funcione frente a variaciones en el entorno y en el patrón de movimiento de los sujetos. Para lograrlo, se ha priorizado la integración de técnicas de preprocesamiento, filtrado de ruido, estructuración de grafos y optimización del entrenamiento, respaldadas por hallazgos recientes con buenos resultados en estos procedimientos en aplicaciones de vigilancia y asistencia [127].

El resultado esperado es un sistema capaz de realizar detección y clasificación de acciones con un balance óptimo entre precisión y tiempo de respuesta, adaptable a distintos escenarios y escalable a futuras mejoras tanto en los modelos como en las fuentes de adquisición de datos

3.1 Diseño de la solución

El sistema propuesto se desarrolló bajo un flujo modular que articula de manera secuencial los componentes necesarios para garantizar tanto la eficacia en la clasificación como la viabilidad de su uso. Este flujo se organiza en cuatro bloques principales, adquisición de video, estimación de pose, representación en grafo espaciotemporal y clasificación mediante una red ST-GCN.

La finalidad de esta estructura es asegurar que cada etapa aporte de manera específica a la reducción de la complejidad del problema y a la mejora de la capacidad de generalización del modelo, sin duplicar funciones ni generar redundancias.

En la adquisición de video, se opta por un mecanismo de transmisión basado en RTSP, lo que permite trabajar en entornos donde la supervisión remota o la asistencia requiere inmediatez en la respuesta. La elección de este protocolo responde al criterio de minimizar la latencia en el tránsito de datos entre la captura y la inferencia, importante en escenarios de monitoreo continuo y que condiciona la utilidad práctica del sistema.

La decisión de emplear RTSP no se fundamenta únicamente en su capacidad de transmisión, sino también en la compatibilidad con dispositivos de bajo costo y en su flexibilidad para integrarse.

El segundo bloque corresponde a la estimación de pose, donde se extraen coordenadas articulares que sintetizan la estructura del cuerpo humano en cada cuadro del video. A diferencia de métodos basados en píxeles o texturas, este enfoque elimina la dependencia de factores externos como variaciones de iluminación, color de la ropa o complejidad del fondo.

De este modo, se obtiene una representación que conserva la dinámica del movimiento y descarta información irrelevante para la tarea de reconocimiento.

La tercera etapa consiste en la construcción de un grafo espaciotemporal, en el cual las articulaciones detectadas se organizan como nodos conectados por aristas que reflejan relaciones espaciales y temporales.

Esta estructura actúa como unión entre la representación y el modelo de aprendizaje profundo, permitiendo que las relaciones anatómicas y las secuencias de movimiento se integren en una misma unidad de procesamiento.

Finalmente, en la cuarta etapa, la clasificación de las acciones se lleva a cabo mediante una ST-GCN. La justificación del uso de esta arquitectura reside en su capacidad para procesar estructuras de grafo y extraer dependencias a distintas escalas temporales, superando las limitaciones de enfoques estrictamente secuenciales.

La inclusión de una ST-GCN asegura que el sistema pueda aprender patrones de interacción complejos entre articulaciones y mantener un equilibrio entre precisión y eficiencia computacional.

El diseño global del sistema está guiado por cuatro criterios que regulan su funcionamiento:

1. Baja latencia.
2. Robustez frente a variaciones en la estimación de pose.
3. Estabilidad en las predicciones.
4. Implementación en módulos independientes.

La combinación de estas etapas y principios de diseño da como resultado una propuesta práctica y adaptable a diferentes contextos de aplicación. De esta forma, el sistema no solo responde a la necesidad de reconocer acciones, sino que también establece un marco técnico que prioriza la integración en entornos reales.

3.1.1 Modelo conceptual ST-GCN para la detección de acciones

En este modelo, los datos de entrada corresponden a coordenadas esqueléticas obtenidas previamente mediante un módulo de estimación de poses (véase sección 3.2.1), que son transformadas en una representación estructurada capaz de capturar relaciones espaciales y temporales.

El diseño de ST-GCN integra convoluciones definidas sobre grafos para modelar las dependencias espaciales entre articulaciones, junto con convoluciones temporales que procesan la evolución de dichas articulaciones a lo largo de la secuencia. Este enfoque es eficaz para tareas de reconocimiento de acciones al preservar simultáneamente la información de conectividad física y la dinámica del movimiento.

En la propuesta actual, la implementación permitió personalizar la estructura de capas, los parámetros de convolución y los mecanismos de agregación de características, optimizando el rendimiento para el conjunto de acciones objetivo. El uso de módulos independientes del modelo facilita incorporar futuras extensiones, como mecanismos adaptativos de conectividad o atención espaciotemporal, sin necesidad de alterar la lógica fundamental de procesamiento.

3.1.2 Representación esquelética y grafo espaciotemporal de poses

La representación esquelética empleada en este trabajo organiza las coordenadas de las articulaciones en una estructura de grafo, las conexiones espaciales reflejan la estructura física del cuerpo, mientras que las conexiones temporales enlazan cada articulación consigo misma a lo largo de fotogramas consecutivos, permitiendo modelar la continuidad y el flujo del movimiento como se presenta en el trabajo de Yan [99].

Para la construcción del grafo, se emplea una matriz de adyacencia que codifica las relaciones de conectividad entre nodos, enriquecida con enlaces adicionales para representar dependencias relevantes no directamente conectadas en la anatomía. Una descripción más detallada de la formulación y propiedades de la matriz de adyacencia puede consultarse en la sección 2.3.3.

Esta representación es importante para que la red ST-GCN (véase sección 3.1.1) pueda operar de manera eficiente, ya que transforma secuencias de coordenadas aisladas que conserva la coherencia espacial y temporal. El enfoque propuesto optimiza el uso de esta representación para capturar interacciones, manteniendo un equilibrio entre precisión y eficiencia computacional.

3.2 Implementación de la propuesta de solución

La implementación de la solución propuesta se estructura en un flujo de procesamiento que transforma secuencias de video en representaciones esqueléticas optimizadas para su análisis mediante redes neuronales sobre grafos espaciotemporales. Este desarrollo integra procesos de adquisición, preprocesamiento, modelado y evaluación, organizados de manera modular para facilitar su mantenimiento y escalabilidad. Cada etapa de este flujo, desde la obtención de los datos hasta la generación de predicciones, se detalla en las subsecciones siguientes (véanse secciones 3.2.1 a 3.2.7), preservando la coherencia con el diseño conceptual y garantizando que el sistema opere en entornos de tiempo real.

3.2.1 Adquisición de datos esqueléticos

La base de datos utilizada para el entrenamiento del modelo proviene del conjunto NTU RGB+D [128], el cual contiene grabaciones multivista y multimodales de 60 clases de acciones humanas, realizadas por 40 participantes en entornos controlados. El conjunto fue capturado mediante cámaras RGB, sensores de profundidad y sensores de esqueleto.

En esta investigación se seleccionaron únicamente cinco clases de acción para el problema planteado, cada una con 948 muestras en formato de video RGB. La decisión de limitar el número de clases responde a la necesidad de optimizar el entrenamiento y enfocar el sistema un número específico de acciones, manteniendo un balance de datos entre clases y focalizándolo hacia un grupo de acciones en específico.

Aunque el NTU RGB+D proporciona datos esqueléticos obtenidos originalmente mediante sensores Kinect, en este trabajo se optó por recalcular las coordenadas usando MediaPipe Pose. Esta decisión se fundamenta en la necesidad de unificar el método de estimación tanto en entrenamiento como en inferencia, evitando depender de hardware especializado y garantizando que el modelo procese información obtenida de la misma fuente en todas las etapas.

MediaPipe Pose detecta 33 puntos clave tridimensionales del cuerpo humano (x, y, z) junto con un valor de visibilidad, pero para mantener la compatibilidad con la estructura definida en NTU RGB+D y reducir dimensionalidad, se conservaron únicamente 25 puntos clave correspondientes a las articulaciones relevantes en la especificación original (ver Figura 3.1), descartando nodos adicionales como detalles faciales y de manos.

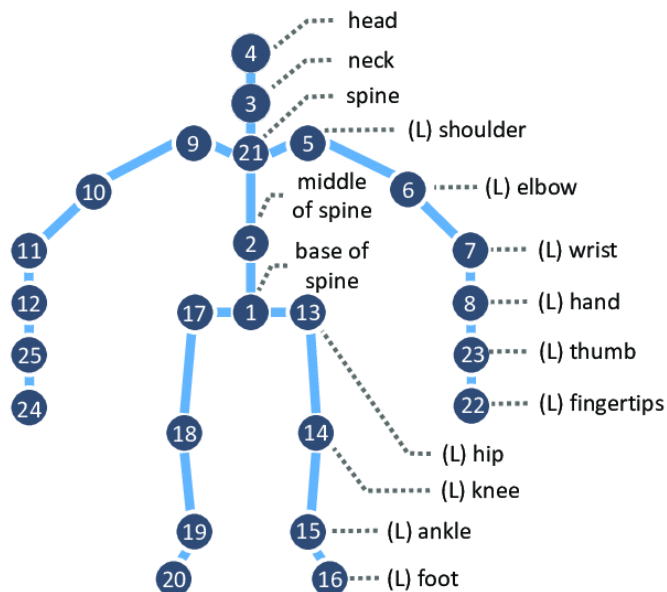


Figura 3.1 Modelo esquelético de 25 articulaciones del NTU RGB+D tomado de [6]

El flujo de adquisición y preparación de datos da inicio con la lectura de los videos seleccionados del conjunto NTU RGB+D, los cuales fueron decodificados cuadro por cuadro en formato RGB para su procesamiento. Posteriormente, en cada fotograma se aplicó el modelo de estimación de poses de MediaPipe Pose, obteniendo 33 puntos clave con sus coordenadas tridimensionales y un valor de visibilidad asociado.

De estos puntos solo se utilizaron los necesarios siguiendo el esquema propuesto por el conjunto de datos NTU RGB+D. Esta reducción garantiza la compatibilidad con el grafo espaciotemporal empleado en la etapa de modelado (véase sección 3.2.4) y elimina información redundante proveniente de puntos adicionales no relevantes para el análisis de acciones.

Una vez filtrados los puntos, las coordenadas normalizadas de las articulaciones se almacenaron en archivos CSV, donde cada fila representa un fotograma y cada columna corresponde a una coordenada específica de un punto clave (x, y, z). Este formato intermedio permite realizar inspecciones y depuraciones antes de alimentar al modelo.

Posteriormente, los archivos fueron convertidos a formato PKL, optimizando la lectura secuencial y reduciendo los tiempos de carga durante el entrenamiento. El archivo final contiene un arreglo tridimensional de forma (N, T, V, C) donde N es el número de muestras, T la longitud temporal en fotogramas, V el número de vértices del grafo y C el número de canales (tres coordenadas espaciales).

Finalmente, las coordenadas fueron sometidas a un proceso de normalización espacial, eliminando variaciones debidas a la distancia de la cámara o al encuadre, de modo que el esqueleto permanezca centrado y escalado uniformemente en todas las secuencias. Este procedimiento permite que el modelo procese datos homogéneos, sin información irrelevante, y alineados con la estructura de grafos utilizada en las fases posteriores del sistema.

3.2.2 Preprocesamiento de datos esqueléticos (filtrado de ruido y normalización)

El objetivo de esta etapa es transformar cada secuencia esquelética en un tensor uniforme, listo para el modelado espaciotemporal. Se operó exclusivamente sobre las coordenadas articulares y sus máscaras de visibilidad, aplicando controles de calidad, reparación de discontinuidades, suavizado temporal y normalizaciones geométricas y estadísticas, para obtener tensores con forma (N, C, T, V) coherentes a lo largo del conjunto.

Se marcaron como inválidos los fotogramas con cobertura insuficiente de articulaciones visibles y se identificaron *outliers* por articulación y canal, para reducir la sensibilidad a errores puntuales de la pose. Los valores atípicos se corrigieron con límites derivados de percentiles y, cuando fue necesario, se sustituyeron por valores interpolados localmente.

Las oclusiones breves o pérdidas de seguimiento se repararon por interpolación temporal (lineal o cúbica) sobre ventanas acotadas, manteniendo la continuidad cinemática sin introducir retardos. Las ausencias prolongadas se mantuvieron como máscaras para evitar introducir información artificial.

Se centró el esqueleto restando la articulación de referencia (pelvis) y se escaló por una medida estable del tamaño corporal para eliminar traslación y escala entre sujetos de esta manera lograr la normalización geométrica intra-secuencia. Adicionalmente, se normalizó la orientación mediante una alineación por cuadro (ejes hombro-hombro y pelvis-columna), garantizando comparabilidad angular entre ejecuciones sin alterar las relaciones relativas entre articulaciones.

Para facilitar la optimización del modelo, cada canal por articulación se estandarizó con media y desviación calculadas solo en el conjunto de entrenamiento, evitando *data leakage* hacia validación y prueba.

Finalmente, las secuencias se dispusieron en el formato requerido por la implementación (de (T, V, C) a (C, T, V)), manteniendo la máscara de visibilidad paralela al tensor principal para un posible uso en etapas posteriores.

Con este flujo, las series esqueléticas quedan libres de valores inválidos, suavizadas sin desfase y comparables en escala, orientación y duración, reduciendo la varianza irrelevante y preservando la cinemática discriminativa que el modelo explota en etapas posteriores. El proceso descrito en esta sección, junto con la etapa de adquisición inicial, se resume en la Figura 3.2, donde se ilustra el flujo completo desde la obtención de coordenadas articulares hasta la generación de tensores listos para el modelo ST-GCN.

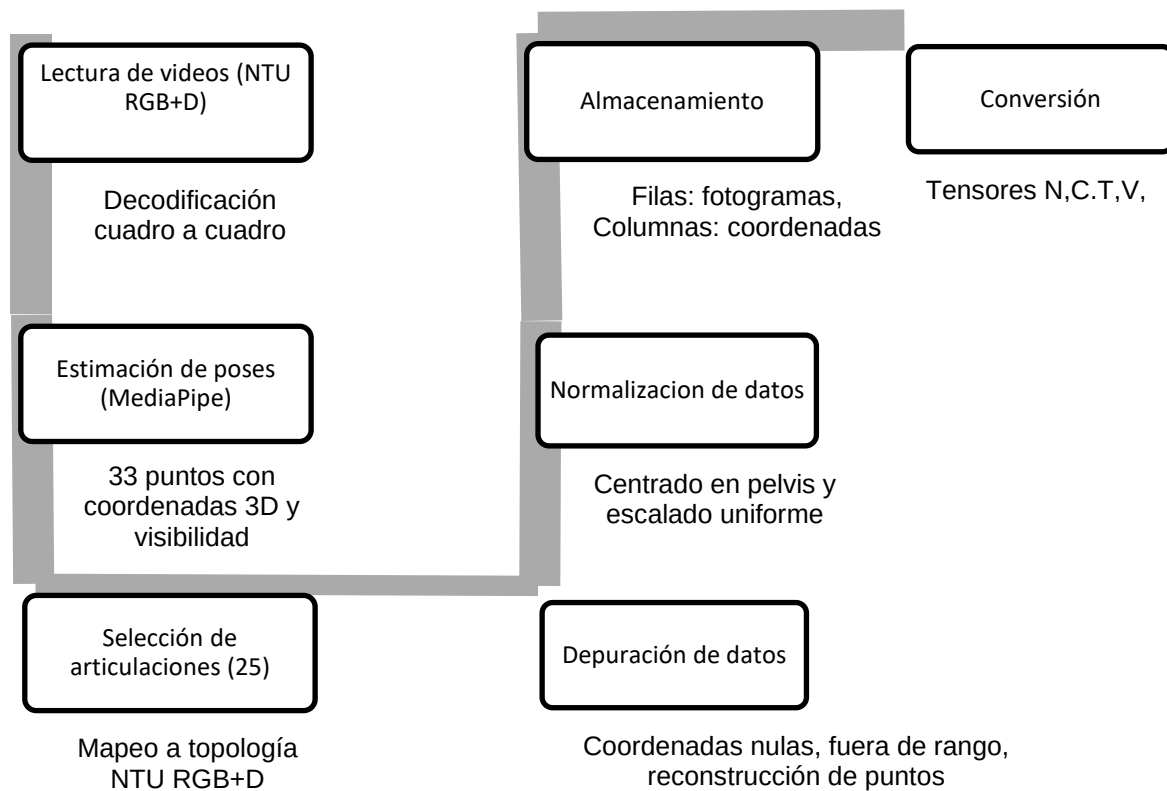


Figura 3.2 Flujo de adquisición y preparación de datos para el modelo ST-GCN (elaboración propia).

El flujo mostrado ilustra el proceso de preparación de datos. El procedimiento comienza con la lectura de los videos del conjunto NTU RGB+D y su decodificación cuadro a cuadro, lo que permite analizar cada imagen de manera individual.

A partir de estos cuadros se aplica la estimación de pose con MediaPipe, obteniendo treinta y tres puntos articulares en tres dimensiones con un valor de visibilidad asociado. Con el fin de mantener la compatibilidad con la topología del conjunto de datos de referencia, se seleccionan veinticinco articulaciones y se realiza un mapeo directo hacia la estructura esquelética definida en NTU RGB+D.

Una vez obtenidos los puntos articulares, la información se organiza en secuencias estructuradas donde cada fotograma queda representado por un conjunto de coordenadas. Posteriormente, se aplican procesos de normalización que incluyen

el centrado respecto a la pelvis y un escalado uniforme para reducir la variabilidad debida a diferencias físicas entre sujetos.

De manera complementaria, se lleva a cabo una depuración que corrige valores nulos o fuera de rango y reconstruye, en caso necesario, articulaciones faltantes con el objetivo de garantizar consistencia en las secuencias. Finalmente, los datos preparados se convierten en tensores con la estructura $N \times C \times T \times V$, que especifica el número de muestras, canales de información, longitud temporal y número de articulaciones, conformando el formato de entrada requerido por la red.

3.2.3 Preparación del conjunto de datos (división en conjuntos de entrenamiento, validación y prueba)

Se adoptó una partición estratificada 70%, 20%, 10% para entrenamiento, validación y prueba, respectivamente, preservando la proporción de las cinco clases en cada subconjunto y evitando solapamientos entre ellos (ver Tabla 3.1). La selección se realizó mediante muestreo aleatorio con semilla fija para asegurar reproducibilidad, práctica recomendada para evaluar modelos de clasificación sin introducir sesgos por particionamiento.

Tabla 3.1 Distribución de las secuencias por clase en los subconjuntos de entrenamiento, validación y prueba

Clase	Entrenamiento (70%)	Validación (20%)	Prueba (10%)	Total
Sentarse	663	190	95	948
De pie	663	190	95	948
Saltar	663	190	95	948
Caerse	663	190	95	948
Dolor de pecho	663	190	95	948
Total	3315	950	475	4740

El conjunto de entrenamiento se utilizó exclusivamente para el ajuste de parámetros; el de validación sirvió para el seguimiento del rendimiento intermedio y la selección de hiperparámetros (criterio de *early stopping* o ajuste del *learning rate*) sin consultar el conjunto de prueba.

El conjunto de prueba permaneció totalmente reservado hasta la evaluación final, con el fin de obtener una estimación imparcial del desempeño fuera de muestra y evitar el optimismo inducido por la reutilización de datos de evaluación durante la selección del modelo.

Además de *accuracy*, se consideraron métricas por clase y promedios para reflejar el rendimiento cuando la prevalencia pudiera diferir entre acciones, tal como recomiendan los estudios sobre aprendizaje.

En conjunto, esta configuración, marca una separación entre particiones y su uso exclusivo de validación para la toma de decisiones de entrenamiento proporciona un protocolo de evaluación libre de fuga de información, adecuado para medir la capacidad de generalización del modelo en escenarios no vistos.

3.2.4 Generación del grafo espaciotemporal de poses

El modelo ST-GCN opera sobre representaciones esqueléticas modeladas como grafos espaciotemporales, en los que cada nodo corresponde a una articulación y cada arista describe una relación estructural o temporal. En este trabajo, el conjunto de nodos por fotograma está formado por las articulaciones normalizadas obtenidas en el preprocesamiento (ver sección 3.2.2), mientras que las aristas espaciales se definen a partir de la topología anatómica del cuerpo humano.

La estructura del grafo se construye a partir de dos tipos de conectividad:

- Conexiones espaciales, que vinculan articulaciones vecinas dentro del mismo fotograma según relaciones anatómicas directas (por ejemplo, codo–muñeca o rodilla–tobillo), así como algunas conexiones simétricas que mejoran la propagación de información entre lados opuestos del cuerpo.
- Conexiones temporales, que enlazan cada articulación con su instancia en fotogramas adyacentes, permitiendo modelar la continuidad del movimiento a lo largo del tiempo.

La Figura 3.3 muestra un ejemplo de grafo espaciotemporal generado a partir de las poses estimadas, donde los nodos corresponden a articulaciones y las aristas representan relaciones anatómicas.

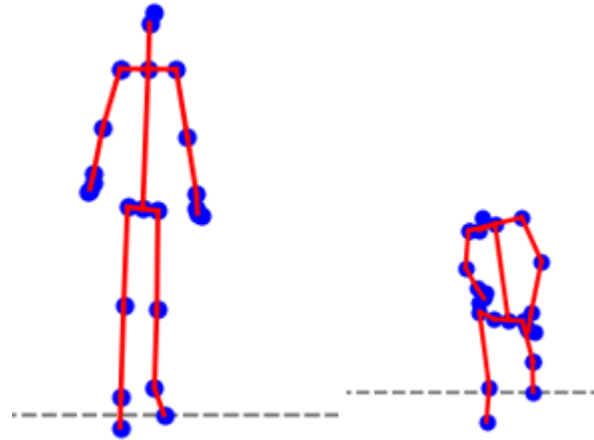


Figura 3.3 Ejemplo de grafo utilizado como entrada al modelo

Ambos conjuntos de conexiones se codifican en matrices de adyacencia que representan explícitamente las relaciones entre nodos. Estas matrices se emplean como insumo para las operaciones de convolución gráfica, de modo que el modelo pueda propagar y combinar información de forma coherente a través del espacio y del tiempo. En este enfoque, las ponderaciones iniciales de las aristas son uniformes, pero la arquitectura ajusta su relevancia durante el entrenamiento, asignando mayor peso a las relaciones más discriminativas (Wang et al., 2022).

Para optimizar la extracción de características, se aplica un particionado funcional de la matriz de adyacencia espacial, separando las conexiones en subconjuntos con roles distintos (por ejemplo, cercanas al centro de masa frente a extremidades). Este tipo de particionado permite a las convoluciones gráficas aprender patrones diferenciados según la posición y función de cada articulación.

El resultado de este proceso es una representación estructurada donde las dependencias espaciales y temporales quedan integradas en un mismo grafo, lo que posibilita al ST-GCN identificar secuencias de movimiento completas y no solo posturas aisladas. El detalle del mecanismo de convolución sobre este grafo se describe posteriormente.

3.2.5 Implementación del modelo ST-GCN

La red se implementó a partir de bloques de convolución gráfica espaciotemporal en serie. Cada bloque contiene una convolución en el grafo espacial seguida de una convolución temporal 1D, extrayendo de forma alternada las relaciones espaciales (estructura del esqueleto) y los patrones temporales de movimiento.

Este diseño de múltiples capas apiladas permite integrar la información a lo largo de las dimensiones espacial y temporal del grafo, capturando tanto la conectividad

corporal inmediata entre articulaciones como la evolución dinámica de las características a través de la secuencia de *frames* (ver Figura 3.4).

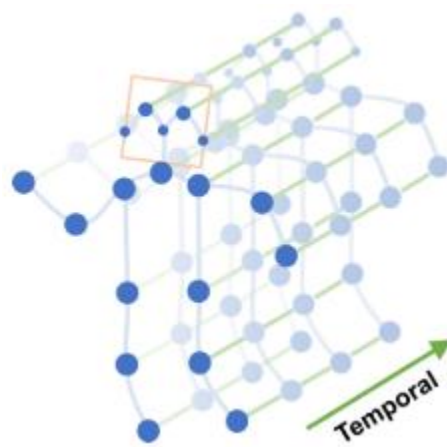


Figura 3.4 Representación del grafo espaciotemporal de poses utilizado, tomado de [99]

La arquitectura resultante es jerárquica: el modelo emplea varias capas de convolución espaciotemporal apiladas, donde las primeras operan sobre representaciones de menor dimensionalidad o más locales, y las últimas construyen descripciones más ricas y de alto nivel.

En otras palabras, las capas iniciales extraen características básicas (relaciones simples entre articulaciones vecinas), mientras que las capas profundas capturan combinaciones más abstractas de movimientos y posturas. De este modo se incrementa progresivamente la capacidad de representación y la discriminación de características relevantes a medida que la información avanza por la red.

Para asegurar la estabilidad del entrenamiento en una red profunda, cada bloque ST-GCN incorpora una normalización por lotes seguida de la activación no lineal. La normalización por lotes estandariza las activaciones de cada capa en cada mini-lote, reduciendo los desplazamientos covariantes internos y acelerando la convergencia del aprendizaje.

Por su parte, la función de activación ReLU introduce no linealidad y evita la saturación del gradiente al anular las activaciones negativas, mitigando así el problema de desvanecimiento del gradiente que ocurriría con funciones sigmoideas tradicionales.

La información de entrada al modelo proviene de las coordenadas articulares normalizadas del esqueleto (obtenidas en la etapa de preprocesamiento) junto con la estructura de conectividad del grafo espaciotemporal. Estas entradas iniciales

(posiciones de las articulaciones en cada *frame* y sus conexiones) se transforman a lo largo de la red en mapas de características.

La Figura 1.1Figura 3.5 muestra el funcionamiento general del modelo ST-GCN. A partir de una secuencia de video como entrada, se estiman las poses del sujeto para generar representaciones esqueléticas que se estructuran en un grafo. Este grafo constituye la entrada a la red, la cual aplica convoluciones sucesivas sobre las relaciones espaciales y temporales de las articulaciones con el fin de aprender patrones característicos del movimiento. Finalmente, las características extraídas son procesadas por un clasificador que asigna la acción correspondiente, representada mediante las puntuaciones de probabilidad para cada clase.

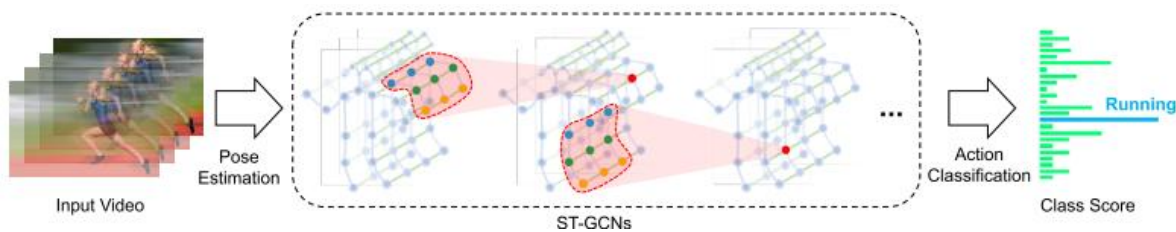


Figura 3.5 Flujo general del modelo ST-GCN, tomado de [99]

Tras pasar por todos los bloques ST-GCN, las características finales se agregan mediante un *pooling* global (promedio global) sobre las dimensiones espaciales (*joints*) y temporal, produciendo un vector fijo que resume toda la secuencia. Finalmente, dicho vector alimenta a una capa totalmente conectada (clasificador final) que emite la categoría de acción reconocida.

Es importante destacar que el diseño del modelo es flexible. Se pueden ajustar parámetros como el número de capas ST-GCN, la cantidad de canales (*features*) en cada capa, o incluso el esquema de conectividad del grafo, sin alterar la lógica general de la arquitectura.

Gracias a esta flexibilidad, el ST-GCN puede adaptarse a diferentes conjuntos de datos y topologías de pose esquelética (por ejemplo, variando el número de articulaciones o las conexiones entre ellas) sin requerir cambios estructurales profundos en la red.

A continuación, la Tabla 3.2 resume los componentes principales del modelo ST-GCN y la función de cada uno en la arquitectura:

Tabla 3.2 Resumen de bloques del modelo ST-GCN

Componente	Función principal
Bloque ST-GCN	Fusiona convolución en grafo (espacial) y temporal para procesar secuencias esqueléticas
<i>Batch Normalization</i> (BN)	Estabiliza activaciones, acelera convergencia
ReLU	Introduce no linealidad y mejora flujo de gradientes
Conexión residual	Facilita el entrenamiento de redes profundas y mitiga <i>oversmoothing</i>
<i>Pooling</i> global	Resume toda la secuencia a un vector fijo para clasificación
Capa totalmente conectada	Clasifica basándose en el vector resumen obtenido

3.2.6 Entrenamiento del modelo (configuración de hiperparámetros)

La configuración de hiperparámetros y el diseño del procedimiento de entrenamiento constituyen un aspecto decisivo para el rendimiento de redes profundas, especialmente en arquitecturas orientadas a secuencias espaciotemporales como ST-GCN. Una configuración inadecuada puede derivar en problemas como sobreajuste, sub entrenamiento o convergencia lenta. En este trabajo, la elección de cada parámetro fue el resultado de un balance entre rendimiento, estabilidad de entrenamiento y recursos computacionales disponibles, siguiendo tanto criterios empíricos como fundamentos presentes en la literatura.

Hiperparámetros utilizados

- Tamaño de lote (*Batch size* = 8)

Se utilizó un tamaño de lote reducido para optimizar el uso de memoria en GPU y permitir actualizaciones más frecuentes de los gradientes, lo que favorece la generalización en conjuntos de datos de tamaño moderado. Lotes pequeños pueden introducir mayor variabilidad en la estimación del gradiente, actuando como una forma implícita de regularización. En contraste, lotes grandes suelen ofrecer mayor estabilidad por iteración, pero tienden a converger a mínimos más estrechos y menos generalizables.

- Número máximo de épocas

Se fijó un límite de 100 épocas para garantizar un entrenamiento suficiente sin sobreentrenar el modelo. El mejor rendimiento se alcanzó en la época 94, con una precisión del 99.68% en validación. Limitar las épocas y conservar el mejor modelo evita la degradación de la capacidad de generalización, tal como se recomienda en el entrenamiento de redes profundas con alto poder de representación.

- Optimizador Adam

Fue elegido por su capacidad de adaptar individualmente la tasa de aprendizaje para cada parámetro, combinando las ventajas de *momentum* y normalización adaptativa de gradientes. Frente a opciones como SGD con *momentum*, Adam ofrece convergencia más rápida en escenarios sin preentrenamiento, aunque requiere un ajuste cuidadoso de la tasa de aprendizaje para evitar oscilaciones.

- Tasa de aprendizaje inicial (0.002) y *scheduler*

Se empleó una tasa de aprendizaje inicial de 0.002, suficientemente alta para una rápida convergencia en fases tempranas, pero sin comprometer la estabilidad del entrenamiento. El ajuste dinámico se realizó mediante *ReduceLROnPlateau* con paciencia de 20 épocas y factor de reducción de 0.5, reduciendo la tasa únicamente cuando la métrica de validación dejaba de mejorar. Este enfoque adaptable evita la necesidad de definir manualmente un calendario rígido de reducción (Loshchilov & Hutter, 2017).

- *Weight decay* = 0.0005

Se utilizó regularización L2 para prevenir la asignación de pesos excesivamente grandes que puedan conducir a sobreajuste. Este valor se seleccionó siguiendo estudios que indican que un rango entre 0.0001 y 0.001 es efectivo para modelos convolucionales profundos.

- Función de pérdida (*Cross-Entropy Loss*)

La pérdida por entropía cruzada es estándar en clasificación multiclase y mide la divergencia entre la distribución predicha y la verdadera. No se emplearon alternativas como *Focal Loss*, ya que el conjunto de datos, tras el preprocesamiento (ver sección 3.2.3), no presentaba un desbalance significativo que justificara su uso.

El entrenamiento del modelo ST-GCN se diseñó como un proceso iterativo y controlado que garantiza tanto la convergencia estable como la prevención de sobreajuste. Cada ciclo de entrenamiento, correspondiente a una época, comprende las siguientes fases:

- Preparación de datos y carga en lotes

Antes de cada época, las secuencias esqueléticas ya preprocesadas son cargadas en memoria mediante un *DataLoader* que gestiona el particionamiento en lotes de tamaño 8. El cargador se encarga de barajar (*shuffle*) las secuencias en cada época, asegurando que el modelo no aprenda patrones de ordenamiento no relacionados con la acción a reconocer.

- Propagación hacia adelante

Cada lote es introducido en el modelo ST-GCN, donde primero se procesan las relaciones espaciales entre articulaciones a través de convoluciones gráficas y, posteriormente, las dependencias temporales mediante convoluciones 1D. Este orden de operaciones permite la extracción jerárquica de características que integran simultáneamente información estructural y dinámica.

- Cálculo de la función de pérdida

Las predicciones generadas por la capa de salida (distribución de probabilidades sobre las clases de acción) son evaluadas frente a las etiquetas verdaderas usando *Cross-Entropy Loss*. Esta métrica cuantifica la discrepancia entre distribuciones y es altamente estable en clasificación multiclase.

- Retropropagación de errores

Se calculan los gradientes de la pérdida respecto a cada parámetro de la red. El optimizador Adam utiliza estos gradientes para ajustar los pesos, incorporando regularizador *weight decay*, penalizando la complejidad excesiva del modelo.

- Ajuste dinámico de la tasa de aprendizaje

Una vez procesado el conjunto de entrenamiento, el rendimiento se evalúa sobre el conjunto de validación. El *scheduler ReduceLROnPlateau* monitoriza la precisión; si esta no mejora durante 20 épocas consecutivas, reduce la tasa de aprendizaje a la mitad. Este mecanismo permite transitar de una fase de exploración rápida de la superficie de error a una fase de ajuste fino en torno a mínimos prometedores.

- Control de sobreajuste mediante validación intermedia

La evaluación por época sobre el conjunto de validación permite identificar signos tempranos de sobreajuste, como un incremento sostenido en la pérdida de validación o una divergencia creciente entre la precisión de entrenamiento y validación. En este trabajo, la precisión en validación

alcanzó su máximo (99.68%) en su mejor época, por lo que se almacenó ese estado del modelo como el mejor resultado.

- Almacenamiento y trazabilidad del mejor modelo

El modelo correspondiente a la época de mejor precisión en validación es guardado en un archivo binario. Esto permite replicar resultados, realizar inferencia directa y evitar la dependencia del estado de entrenamiento para el despliegue.

Los valores empleados, junto con su justificación resumida, se muestran en la Tabla 3.3.

Tabla 3.3 Hiperparámetros utilizados

Parámetro	Valor	Justificación
<i>Batch size</i>	8	Uso eficiente de GPU y regularización implícita
Épocas	100 (mejor en la 94)	Evitar sobre entrenamiento manteniendo suficiente iteración
Optimizador	Adam	Convergencia rápida y robustez ante gradientes variables
Tasa de aprendizaje (lr)	0.002	Balance entre rapidez y estabilidad
<i>Scheduler</i>	ReduceLROnPlateau (p=20, f=0.5)	Ajuste dinámico según métricas de validación
<i>Weight decay</i>	0.0005	Control de magnitud de pesos para evitar sobreajuste

3.2.7 Evaluación del modelo entrenado

Se describe el proceso de optimización paso a paso, destacando la dinámica de convergencia, las decisiones de control durante el ciclo de entrenamiento y la interpretación de los resultados intermedios. La explicación busca ofrecer una visión del procedimiento seguido y relacionar el comportamiento observado con el esperado para modelos de grafos espacio-temporales aplicados a secuencias de esqueleto humano.

3.2.7.1 Flujo de entrenamiento

Durante el entrenamiento del modelo ST-GCN, se ejecutó un ciclo estructurado que garantizó la convergencia estable y la correcta generalización a datos no vistos. Las secuencias preprocesadas se cargaron en pequeños *batch* de tamaño 8 y, antes de cada época, se barajaron aleatoriamente para evitar que el orden de presentación introdujera sesgos no deseados en el aprendizaje.

Cada lote atravesó de manera secuencial los bloques espaciotemporales. En la parte espacial, la convolución gráfica operó sobre el grafo del esqueleto, compuesto por 25 nodos conectados según la estructura anatómica, extrayendo características de las relaciones articulares.

La parte temporal consistió en una convolución 2D, que permitió capturar patrones dinámicos a lo largo de la secuencia de *frames*. En cada bloque, se aplicó *batch normalization* y la función de activación ReLU, además de conexiones residuales cuando correspondía, lo que favoreció la propagación estable de gradientes.

Finalizada la propagación hacia adelante, se calculó la función *loss* de clasificación mediante el uso de *Cross-Entropy function* multiclase, comparando las predicciones con las etiquetas reales del lote. Posteriormente, se llevó a cabo la retropropagación para calcular los gradientes y se aplicó el optimizador Adam con un *learning rate* inicial de 0.002 y un *weight decay* de 0.0005, para limitar la magnitud de los pesos y prevenir el sobreajuste.

Al finalizar cada época, el modelo se evaluó en el conjunto de validación calculando *loss* y *accuracy*. Se implementó un esquema de ajuste dinámico, que reducía la tasa en un factor de 0.5 si no se observaban mejoras sostenidas en un periodo de 20 épocas. Este ajuste permitió transitar de una fase inicial de exploración a una fase de refinamiento, estabilizando la convergencia en etapas avanzadas.

A continuación, la Figura 3.6 muestra el flujo de entrenamiento. A partir de los datos preprocesados organizados en lotes, las secuencias atraviesan los bloques espaciotemporales de la red, donde se extraen características mediante convoluciones, normalización y activaciones. Las salidas se condensan en un vector que alimenta al clasificador, cuyas predicciones se comparan con las etiquetas reales para calcular la pérdida.

Con ello se ejecuta la retropropagación y la actualización de parámetros, para asegura la convergencia estable del modelo.

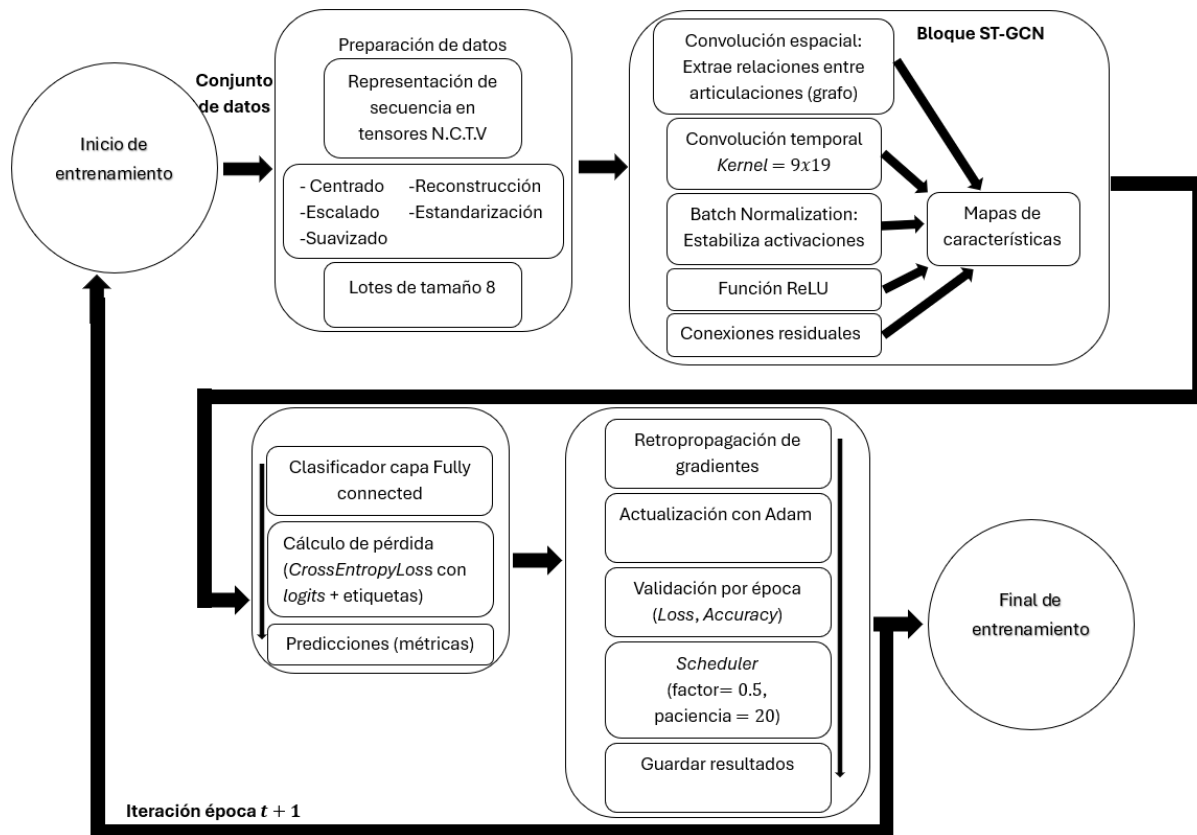


Figura 3.6 Diagrama de flujo del proceso de entrenamiento (elaboración propia)

Se registró el mejor estado del modelo en función del valor obtenido de *accuracy* en validación. Cuando esta superaba el máximo previo, los pesos se almacenaban como la mejor versión. De esta manera, el modelo con mejor rendimiento se obtuvo durante la época 94, alcanzando 99.68% en *accuracy* de validación y esto sirvió como referencia para la evaluación final en el conjunto de prueba.

3.2.7.2 Comparativa con la literatura, comportamiento esperado contra observado

Este estudio utilizó un conjunto específico de acciones compuesto por cinco clases tomadas del conjunto de datos NTU RGB+D, las cuales incluyen caerse, dolor de pecho, sentarse, de pie y saltar. Con el fin de establecer un marco de referencia en el estado del arte, se llevó a cabo una revisión bibliográfica centrada en investigaciones que emplearan el modelo ST-GCN o variantes estrechamente relacionadas y que contemplaran acciones coincidentes o análogas a las presentes en este trabajo.

Se identificaron estudios relevantes que abordaban la detección de caídas y el reconocimiento de posturas básicas, como sentarse y estar de pie, así como

movimientos simples, por ejemplo, saltar. Estas categorías abarcan parcialmente las acciones de interés en la presente investigación y permiten realizar una comparación del comportamiento observado.

En relación con la detección de caídas [129], reportaron aproximadamente 98% de exactitud en la clase “Caerse” utilizando un ST-GCN adaptado a secuencias esqueléticas, manteniendo un desempeño elevado tanto en el protocolo *cross-subject* como en el *cross-view* del conjunto de datos NTU RGB+D, con métricas por clase para “Caerse” de *precision*, *recall* y *F1-Score* en el rango de 0.98 a 0.99.

Estos valores se alinean estrechamente con los obtenidos en el presente trabajo para la clase “Caerse” (F1 aproximado de 0.99), el comportamiento del modelo propuesto para la acción Caerse es congruente con la literatura y reafirma la capacidad del ST-GCN para capturar patrones abruptos característicos de este tipo de eventos.

Estudios recientes han explorado variantes del ST-GCN para su uso en entornos de inferencia, un ejemplo de ello se muestra en estudio que presentó un enfoque que incorporó aristas entrenadas en el grafo del con el objetivo de optimizar la sensibilidad hacia conexiones articulares.

Confirmando que las transiciones bruscas propias de las caídas podían modelarse de manera más precisa mediante la adaptación dinámica de la topología del grafo durante el entrenamiento, alcanzando una exactitud cercana a 96% en detección de caídas utilizando estimaciones de pose generadas con AlphaPose y un ST-GCN mejorado.

El autor informó un incremento aproximado de 15 puntos porcentuales en *accuracy* al incorporar pesos de arista entrenados. Los resultados obtenidos para la acción Caerse indican que la arquitectura empleada posee capacidades de generalización.

Por otra parte, análisis comparativos como el realizado por Yang [130] han destacado que los modelos basados en grafos espaciotemporales superan consistentemente a enfoques no gráficos en la detección de caídas. Dichos autores señalan que ST-GCN sobre secuencias esqueléticas ofrece un mejor desempeño esto frente a métodos tradicionales como el uso de sensores portables, análisis de video RGB sin segmentación de pose.

Este comportamiento se atribuye a la capacidad de los grafos para preservar la estructura espacial del cuerpo y a la efectividad de las convoluciones temporales para modelar la dinámica de movimiento.

Yang [130] también resaltan la robustez del ST-GCN en escenarios con iluminación variable y fondos visualmente complejos, lo que respalda su potencial en aplicaciones con entornos reales. Los resultados de la presente investigación corroboran estas observaciones, ya que el modelo logró detectar caídas con un número mínimo de errores, validando la pertinencia de la arquitectura basada en grafos para este tipo de eventos.

En esta investigación, el modelo obtuvo métricas cercanas al máximo posible en las clases Sentarse, De pie y Saltar, presentando únicamente confusiones marginales entre las dos posturas estáticas, lo que coincide con tendencias reportadas previamente.

El autor Shi [25], al proponer el modelo 2s-AGCN como extensión del ST-GCN, señaló que la arquitectura estándar ya modela de forma efectiva las dinámicas esenciales de acciones simples de postura, por lo que las mejoras adicionales tienden a ser limitadas debido a que el problema se aproxima a un estado de saturación en términos de precisión.

3.2.7.3 Métricas de desempeño utilizadas

La evaluación de modelos de reconocimiento de acciones requiere métricas estandarizadas que permitan medir la capacidad de discriminación y la fiabilidad de los resultados. Entre las más utilizadas en la literatura se encuentran *Precision*, *Recall*, *F1-Score* y *Accuracy*, todas derivadas de la matriz de confusión.

La métrica *Precision* cuantifica la proporción de predicciones positivas que son realmente correctas, mientras que el *Recall* mide la fracción de instancias positivas que el modelo logra identificar adecuadamente. Ambos indicadores son fundamentales para evaluar sistemas en los que los falsos positivos o falsos negativos tienen un peso considerable [131].

Para obtener una métrica combinada que equilibre ambos aspectos, se emplea el *F1-Score*, definido como la media armónica entre *Precision* y *Recall*. Su uso se popularizó en el campo de *Information Retrieval* y posteriormente fue adoptado en la evaluación de modelos de aprendizaje automático, ya que evita que un valor alto en una métrica compense un valor bajo en la otra [132].

Es recomendable complementar *Accuracy* con métricas más sensibles a distribuciones desiguales de clases, como *Precision*, *Recall* y *F1-Score*, en contextos de análisis de eventos poco frecuentes. Asimismo, investigaciones posteriores subrayan la necesidad de un uso consciente de estas métricas,

considerando los posibles sesgos que introducen en los reportes de desempeño [133].

3.3 Integración del sistema

La integración del sistema constituye la fase en la que los módulos desarrollados previamente (desde la adquisición y preprocesamiento de datos esqueléticos hasta el entrenamiento y validación del modelo ST-GCN se unifican para operar como una solución funcional. Esta etapa garantiza que los componentes de estimación de poses, clasificación de acciones y gestión de resultados interactúen de manera funcional, optimizando el flujo de datos desde su captura hasta la generación de salidas visuales y registros persistentes.

En este contexto, la integración no solo implica conectar el modelo entrenado a un flujo continuo de video, sino también habilitar mecanismos de visualización y registro que permitan interpretar y almacenar los resultados de forma inmediata y verificable. Se busca que la arquitectura resultante sea modular, adaptable a distintos escenarios de despliegue y capaz de mantener su rendimiento bajo condiciones reales de operación.

El diseño de esta fase se estructura en cuatro componentes principales. El primero es el flujo de video, que provee el canal de entrada de datos visuales y soporta la compatibilidad con cámaras IP. El segundo es la interfaz de usuario, que actúa como capa de presentación de resultados, mostrando la acción detectada y las características relevantes de la secuencia.

El tercero corresponde al subsistema de generación de evidencias y registro, encargado de almacenar eventos y datos relevantes para posteriores análisis o auditorías. Finalmente, el cuarto componente es el despliegue del sistema, que define el entorno operativo y asegura la ejecución estable del modelo en condiciones reales.

3.3.1 Flujo de video en tiempo real vía RTSP

El flujo de video en tiempo real se implementa mediante el protocolo RTSP, el cual permite transmitir secuencias de video desde fuentes remotas hacia el sistema de inferencia. Este protocolo fue seleccionado por su compatibilidad con múltiples dispositivos y por su capacidad para manejar transmisiones continuas con baja latencia, haciéndolo adecuado para la implementación del sistema.

El módulo de adquisición establece una conexión persistente con la fuente RTSP y recupera cada *frame* en formato de imagen digital, manteniendo la frecuencia de

actualización definida por el dispositivo emisor. Una vez capturado, cada *frame* es sometido a un preprocesamiento rápido que incluye la estimación de poses mediante MediaPipe y la normalización espacial para garantizar que la representación esquelética mantenga la coherencia geométrica antes de ser enviada al modelo ST-GCN entrenado.

Para asegurar la estabilidad del flujo bajo condiciones reales se integraron mecanismos de reconexión automática y almacenamiento temporal en *buffer*. Esto permite que, aun cuando la conexión experimente interrupciones momentáneas, el sistema continúe procesando las secuencias sin degradaciones en la precisión de la detección.

La integración del RTSP en el *pipeline* de inferencia posibilita que la salida del modelo sea generada con una latencia mínima y facilita su uso en escenarios donde la respuesta inmediata es importante, como la detección de caídas. La conexión directa entre la fuente de video y el módulo de análisis evita procesos intermedios que puedan retrasar la inferencia, optimizando así el tiempo total de respuesta del sistema.

3.3.2 Interfaz de usuario para visualización de resultados en tiempo real

Con el propósito de facilitar la interacción entre el sistema y el usuario final, se diseñó una interfaz que integra en un único entorno los flujos de video procesados y la retroalimentación del modelo. Esta interfaz permite, por un lado, desplegar la acción reconocida a partir del flujo de video entrante, por otro, organiza la información en paneles gráficos que sintetizan el rendimiento alcanzado durante la ejecución.

El diseño de la interfaz prioriza la claridad y la inmediatez de interpretación, presentando al usuario no especializado un conjunto de visualizaciones que permiten identificar de forma intuitiva el comportamiento del sistema.

La interfaz constituye, por tanto, un conjunto entre la capacidad del modelo ST-GCN y las necesidades de supervisión, al brindar un espacio en el que la información se presenta de manera accesible y cómoda.

3.3.3 Generación de evidencias, registro y almacenamiento de eventos

El sistema incorpora un módulo de generación y almacenamiento de evidencias diseñado para documentar de manera estructurada cada evento detectado durante la operación en tiempo real. Su función principal es garantizar la trazabilidad de las predicciones del modelo, proporcionando un respaldo objetivo que permita verificar, auditar o analizar posteriormente el desempeño del sistema.

Cada vez que el modelo identifica una acción, se activa el proceso de registro que incluye los siguientes tres elementos, el fotograma o secuencia de video correspondiente al instante del evento, la representación esquelética y los metadatos asociados. Estos metadatos contienen el nombre de la acción, el nivel de confianza de la predicción, la marca de tiempo precisa y, en el caso de acciones críticas, un indicador de prioridad que puede ser utilizado para activar sistemas externos de notificación o respuesta.

Para el almacenamiento, el sistema emplea una estructura jerárquica organizada por fecha y tipo de evento, lo que facilita la búsqueda y recuperación posterior. Cada evidencia visual se guarda en un formato de imagen o video comprimido, optimizado para minimizar el espacio en disco sin comprometer la calidad necesaria para su análisis ver Figura 3.7.

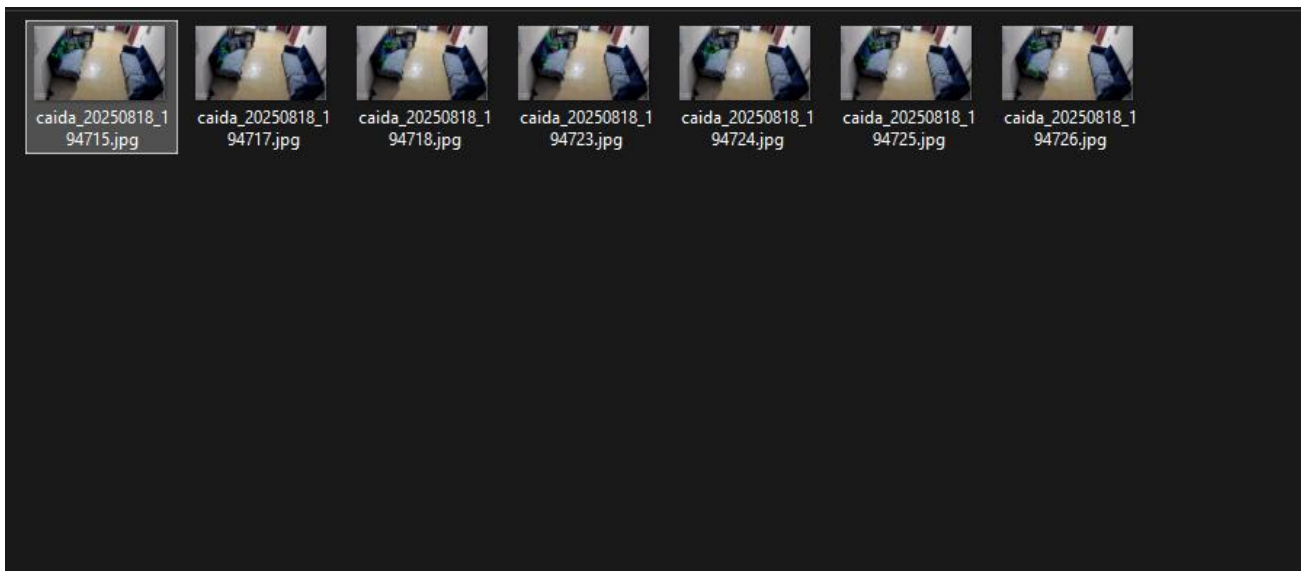


Figura 3.7 Organización y almacenamiento de evidencias

3.3.4 Despliegue del sistema

El sistema se desplegó con el propósito de garantizar un funcionamiento estable y continuo, integrando los módulos de procesamiento y visualización en un mismo entorno. El proceso comprendió la configuración del entorno de ejecución y la orquestación de los componentes en una arquitectura unificada, capaz de procesar video en tiempo real, emitir predicciones y registrar eventos de forma autónoma.

En primera instancia, se definió el entorno de ejecución, optimizando el uso de recursos para el procesamiento intensivo que demanda el análisis de secuencias. Se estableció un sistema de gestión de dependencias que asegurara la instalación

controlada de las librerías y versiones específicas empleadas durante el desarrollo, con el fin de mantener la reproducibilidad de los resultados y evitar incompatibilidades.

La arquitectura final integró el flujo de adquisición de video, el módulo de inferencia con el modelo entrenado, la interfaz de usuario para visualización en tiempo real (Figura 3.8) y el subsistema de registro y almacenamiento de evidencias. Estos elementos se comunicaron mediante una estructura que permitió aislar fallos y facilitar el mantenimiento sin interrumpir el servicio completo.



Figura 3.8 Interfaz desarrollada para gestionar los modos de inferencia del sistema

Para garantizar la disponibilidad continua, el sistema se configuró como un servicio persistente, capaz de reiniciarse automáticamente en caso de interrupciones. Asimismo, se implementaron mecanismos de monitoreo básico que verifican el estado de los flujos de video, con el fin de detectar y corregir oportunamente posibles degradaciones de rendimiento.

Finalmente, el despliegue se realizó siguiendo un esquema que permitiera la escalabilidad futura, de modo que, si se requiere, el sistema pueda integrarse con múltiples cámaras, servidores distribuidos o plataformas de almacenamiento y análisis, sin necesidad de rediseñar la arquitectura central. Como ejemplo, la Figura 3.9 muestra el flujo de video recibido desde una cámara IP durante el proceso de inferencia.



Figura 3.9 Flujo de video recibido desde una cámara IP durante el proceso de inferencia

Capítulo 4

Experimentación y resultados

4.1 Resumen de la experimentación

La fase de experimentación tuvo como objetivo validar el desempeño del modelo ST-GCN implementado para la descripción automática de acciones humanas a partir de secuencias de poses estimadas. El proceso experimental se diseñó considerando la replicabilidad de resultados y la evaluación bajo condiciones que simulan el funcionamiento real del sistema en entornos controlados, con especial atención al reconocimiento de acciones relevantes para el monitoreo de personas mayores.

El conjunto de pruebas se estructuró en múltiples etapas, comenzando por la preparación del conjunto de datos personalizado, donde las secuencias de video fueron convertidas en representaciones esqueléticas mediante estimación de pose, y posteriormente preprocesadas para garantizar la consistencia topológica.

Sobre estas representaciones se entrenó el modelo ST-GCN utilizando diferentes configuraciones de hiperparámetros y técnicas de optimización, con el fin de analizar su impacto en el rendimiento ya sea por clase o general (ver Tabla 4.1).

Tabla 4.1 Síntesis de la experimentación para el ajuste de hiperparámetros y evaluación del sistema

N.º Experimento	Resultado clave	Objetivos específicos validados
E1. Entrenamiento base (datos balanceados)	<i>Accuracy</i> de validación 99.68% y F1 macro 98.4%, 3 errores en 480 muestras, sin sobreajuste	1,2
E2. Longitud de ventana temporal (50, 100, 150 cuadros)	100 cuadros fue el mejor valor 50 degradó <i>recall</i> /F1 150 aumentó costo sin mejora	2,4
E3. Tamaño de lote (<i>batch</i>)	Se adoptó <i>batch</i> 16 por equilibrio <i>batch</i> 8 mejoró ligeramente F1 en dos clases <i>batch</i> 32 redujo aprox. 30% el tiempo por época sin ganancia en métricas	3,4

Tabla 4.1 Síntesis de la experimentación para el ajuste de hiperparámetros y evaluación del sistema (Continuación)

N.º Experimento	Resultado clave	Objetivos específicos validados
E4. Número de épocas	100 épocas ofrecieron el mejor equilibrio 50 y 75 aceptables, pero con ligera pérdida en clases similares	3, 4
E5. Tasa de aprendizaje (LR)	LR = 0.001 fue la mejor opción por estabilidad y rendimiento global (<i>accuracy</i> 99.59%, F1 99.58%)	3
E6. Desempeño por clase	Métricas altas en todas las clases, “Caerse” con F1 de 99%; confusiones residuales entre posturas estáticas	4
E7. Voto por mayoría (ventanas deslizantes)	Mejora ligera y consistente de precisión/F1 y reducción de falsos positivos, latencia total de 1 a 2 segundos.	5, 6
E8. Inferencia en condiciones reales (distancia, iluminación)	Predicciones estables y visualización estructurada; operación continua con baja latencia	5, 6
E9. Detección de caídas como evento crítico	F1 de 99% en “Caerse”, alerta visual y registro automático de evidencia	5, 6
E10. Sistema de inferencia en operación	Mantuvo <i>accuracy</i> y F1 macro entre 99.4% y 99.6% Latencia aproximada de 1 a 2 segundos.	5, 6

La experimentación se llevó a cabo en un entorno controlado de hardware y software, documentando las herramientas utilizadas y los parámetros de ejecución, de forma que el proceso pueda ser reproducido. Asimismo, se implementaron procedimientos de evaluación tanto a nivel de frame como a nivel de video, incluyendo métricas cuantitativas como *accuracy*, *precision*, *recall*, *F1-Score* y análisis mediante matrices de confusión.

Finalmente, se realizaron pruebas adicionales para medir el comportamiento del sistema en escenarios de inferencia continua mediante flujo RTSP, con el fin de estimar la latencia y el rendimiento en condiciones cercanas a la operación en tiempo real.

Los resultados obtenidos se presentan y analizan en las subsecciones posteriores, desglosando las características de cada experimento y su impacto en la capacidad del modelo para discriminar acciones de manera precisa.

4.2 Herramientas y tecnologías

La implementación del sistema requirió la integración de hardware especializado, un entorno de desarrollo optimizado y bibliotecas de software adecuadas para garantizar tanto el rendimiento como la estabilidad del modelo durante el entrenamiento y la inferencia. A continuación, se describen los componentes empleados.

4.2.1 Hardware utilizado

El entrenamiento del modelo y la ejecución del sistema se realizaron en una estación de trabajo con las siguientes especificaciones:

- Procesador: Intel Core i5 de 12^a generación.
- Memoria RAM: 32 GB DDR4.
- Unidad gráfica: NVIDIA GeForce RTX 4070 *Super*.
- Almacenamiento: SSD NVMe de 1 TB.
- Dispositivo de captura: Cámara IP TP-Link Tapo C200.

Esta infraestructura permitió ejecutar operaciones de cómputo intensivo con soporte para aceleración por GPU, reduciendo los tiempos por época de entrenamiento y permitiendo el procesamiento en lote de múltiples secuencias de datos. Durante las pruebas de inferencia, la GPU garantizó una respuesta ágil del sistema ante la llegada continua de datos desde la cámara, posibilitando ejecutar simultáneamente la estimación de pose, el armado del grafo de entrada, la predicción con el modelo entrenado y la visualización en pantalla, sin comprometer la estabilidad del flujo.

4.2.2 Entorno de desarrollo

El proyecto fue implementado principalmente en el lenguaje Python 3.11.9, elegido por su amplia compatibilidad con bibliotecas de visión por computadora y

aprendizaje profundo. Este lenguaje facilitó la integración de múltiples componentes (procesamiento de video, estimación de pose, modelado con grafos y visualización de resultados) de forma coherente.

Las pruebas experimentales se llevaron a cabo en un entorno operativo Windows 11 Pro. Este sistema permitió el uso simultáneo de herramientas de monitoreo del rendimiento, scripts de inferencia en tiempo real y visualización de datos, proporcionando una plataforma robusta para el desarrollo y prueba del sistema de reconocimiento de acciones.

4.2.3 Bibliotecas principales

- El desarrollo de la solución involucró varias bibliotecas especializadas, seleccionadas por su aporte funcional al sistema. A continuación, se listan las principales herramientas de software utilizadas, junto con una breve descripción de su papel en el proyecto:
- PyTorch [134]: Framework principal utilizado para implementar la arquitectura de red neuronal ST-GCN, tanto durante la etapa de entrenamiento como en la inferencia. Su soporte para operaciones tensoriales en GPU permitió aprovechar la aceleración por hardware de la RTX 4070 en tareas de aprendizaje profundo.
- PyTorch Geometric [105]: Biblioteca complementaria para el procesamiento de datos en forma de grafos, fundamental en la construcción y manipulación de la estructura del modelo ST-GCN. Facilitó la definición del grafo espacial-temporal del esqueleto humano y la aplicación eficiente de convoluciones sobre dicha estructura.
- MediaPipe [117]: Herramienta empleada para la estimación de poses humanas a partir de imágenes. En particular, se utilizó el módulo *MediaPipe Pose*, que permite extraer por cada fotograma los puntos clave 3D del cuerpo de la persona. Esta biblioteca proporcionó los 33 *landmarks* corporales por frame.
- OpenCV: Biblioteca de visión por computadora utilizada para la captura y manejo de video. OpenCV permitió leer el flujo RTSP de la cámara IP de forma similar a un archivo de video, decodificando *frames* en tiempo real. Además, se utilizó para el procesamiento básico de cada frame (p. ej., cambios de formato, redimensionamiento) y para la visualización de resultados, incluyendo el renderizado del esqueleto estimado sobre el video original y la anotación de la acción reconocida en pantalla.

- NumPy y Matplotlib [135], [136]: NumPy permite manipular matrices y tensores durante las etapas de preprocesamiento (normalización de coordenadas, interpolación de secuencias, etc.), aprovechando sus operaciones vectorizadas para mejorar el rendimiento. Matplotlib se utilizó para graficar las métricas de desempeño y para visualizar resultados experimentales, facilitando el análisis cuantitativo del comportamiento del modelo.
- Scikit-learn [137]: Empleada para el cálculo de métricas de evaluación clásicas. En particular, esta biblioteca permitió generar la matriz de confusión y calcular indicadores como *precision*, *recall* y *F1-Score* por clase, a partir de las predicciones del modelo sobre el conjunto de prueba. Estas métricas se usaron para evaluar objetivamente el desempeño del sistema y comparar los resultados entre distintos experimentos.
- Cabe destacar que todas estas bibliotecas fueron integradas de manera coherente en la solución, MediaPipe y OpenCV se encargan de la entrada (captura de video y extracción de poses), NumPy y PyTorch/PyTorch Geometric manejan la representación y procesamiento interno de los datos (estructuras de grafo, tensores y convoluciones), mientras que Matplotlib y Scikit-learn apoyan el análisis *offline* de resultados y la generación de reportes estadísticos.

4.2.4 Gestión del sistema

En esta sección se detalla cómo se estructuró, automatizó y monitoreó el flujo completo de experimentación, abarcando desde la ingestión de datos hasta la generación de evidencias visuales y registros (*logs*). El enfoque se centra en la operativa práctica del sistema para garantizar un funcionamiento continuo y confiable durante los experimentos.

- Arquitectura modular del sistema: El sistema se dividió en módulos independientes, cada uno encargado de una etapa del procesamiento. La separación en módulos facilitó el desarrollo, la depuración y el mantenimiento del sistema. Los principales módulos implementados fueron:
- Pose Extractor: módulo de adquisición de datos que ejecuta la extracción de poses en tiempo real. Captura cada frame del video entrante y aplica *MediaPipe Pose* inmediatamente para obtener el esqueleto (coordenadas de las articulaciones). Los datos extraídos son enviados de forma inmediata al siguiente módulo para su procesamiento.

- **Preprocesador:** módulo responsable de normalizar las coordenadas esqueléticas y organizar los datos en lotes o secuencias preparadas para la inferencia o almacenamiento. Aquí se aplica la normalización espacial (centrado del esqueleto, escalado por la distancia de la pelvis, etc.) y se estructura la ventana temporal de tamaño fijo que funge como entrada al modelo ST-GCN, tal como se definió en el diseño experimental.
- **Inferencia:** este módulo gestiona la carga del modelo ST-GCN entrenado y realiza las predicciones de acción. Recibe los lotes de secuencias formateados por el preprocesador y ejecuta la inferencia de la red neuronal. Se configuró para procesar lotes de tamaño determinado de manera automatizada, aprovechando la GPU para acelerar el cálculo. Tras procesar cada lote, envía las etiquetas predichas (y posiblemente las probabilidades asociadas) al módulo de visualización.
- **Visualización/Evidencia:** módulo encargado de la salida del sistema. Muestra en pantalla los esqueletos sobre el video original y superpone la acción reconocida junto con su grado de confianza. Además, este componente gestiona la captura de evidencias visuales: por ejemplo, si se detecta una acción de interés como “Caerse”, el sistema toma automáticamente una instantánea del frame correspondiente y la almacena con fines de análisis o trazabilidad.

Cada uno de estos módulos corre como un proceso independiente, comunicándose con los demás a través de mecanismos de intercambio de datos controlados (*buffers* de memoria compartida, colas de mensajes o archivos temporales, según el caso). Esta arquitectura distribuida permite arrancar, detener o reiniciar cualquier etapa del *pipeline* sin interrumpir completamente el flujo general. Por ejemplo, se puede reiniciar el módulo de captura de video si la cámara se desconecta momentáneamente, mientras el módulo de inferencia sigue activo procesando los últimos datos disponibles. De igual forma, si se requiere actualizar el modelo entrenado por una nueva versión, el módulo de inferencia puede reemplazarse sin afectar al de adquisición o visualización. Esta separación de responsabilidades aumenta la robustez del sistema y facilita su mantenimiento.

Como parte del soporte a la experimentación, se desarrolló una utilidad de validación gráfica que permite representar los esqueletos almacenados en los archivos de datos de entrenamiento, antes de alimentar al modelo. Mediante esta herramienta, es posible visualizar las secuencias esqueléticas previamente capturadas para verificar su correcta conformación y consistencia.

Esta verificación visual ayuda a depurar errores en la generación o preprocesamiento de datos (por ejemplo, detectar si alguna articulación presenta valores fuera de rango o si la normalización fue incorrecta) y a evaluar intuitivamente la coherencia espacial de los movimientos en las secuencias. En resumen, la validación gráfica sirvió para asegurar la calidad de los datos antes de emprender entrenamientos o inferencias masivas, agregando una capa de control de calidad al flujo de trabajo.

Al iniciar un experimento de entrenamiento, el proceso automatizado realiza las siguientes funciones principales.

- Lee los archivos de datos esqueléticos correspondientes al conjunto de entrenamiento, así como los de validación si se usan durante el entrenamiento. Configura los hiperparámetros según la especificación deseada para el experimento.
- Construye la arquitectura del modelo ST-GCN con la topología y pesos iniciales requeridos. Incluyendo la definición de las capas de convolución espacial-temporal, el grafo y las conexiones conforme a la estructura corporal.
- Inicia el bucle de entrenamiento por épocas, iterando sobre todos los *batches* de datos. En cada época, el modelo procesa todos los ejemplos de entrenamiento, calcula la función *loss* y ajusta sus pesos mediante retro propagación. Al final de cada época, el sistema valida el rendimiento del modelo usando el conjunto de validación (calculando la precisión o pérdida en validación), lo que permite monitorear la convergencia y detectar problemas de sobreajuste. Estos resultados por época se registran para su análisis posterior.
- Genera los datos de salida donde se guardan automáticamente los pesos del modelo entrenado. Asimismo, se exportan las gráficas de pérdida y precisión por época. Esto asegura que cada experimento de entrenamiento quede documentado y reproducible.
- Al concluir el entrenamiento, se consolidan las métricas finales obtenidas en un registro general que resume todos los experimentos realizados. De esta manera, se lleva una trazabilidad histórica de los resultados obtenidos bajo distintas configuraciones, facilitando la comparación entre experimentos (por ejemplo, para identificar qué configuración de hiperparámetros produjo el mejor desempeño).

Este esquema automatizado de entrenamiento permite replicar experimentos fácilmente, modificando únicamente los parámetros de configuración de entrada según se requiera. Por ejemplo, para lanzar un experimento con distinta longitud de secuencia o con un esquema de aumento de datos, basta con ajustar esas opciones en la configuración y el sistema gestionará el resto del proceso de forma consistente.

Finalizado el entrenamiento, el sistema realiza una evaluación formal del modelo utilizando el conjunto de prueba separado. Esto implica cargar el modelo entrenado (el que obtuvo mejor resultado en validación) y ejecutar una inferencia completa sobre todos los datos de prueba, recogiendo las predicciones. Con esas predicciones, el módulo de análisis calcula las métricas finales, precisión global, la precisión, *recall* y *F1-Score* por clase y construye la matriz de confusión correspondiente.

Estos resultados se almacenan junto con los obtenidos en el entrenamiento, de modo que cada experimento tiene un registro completo de su rendimiento, desde la fase de entrenamiento (convergencia) hasta la evaluación final.

Para la fase de inferencia continua, se automatizó la detección de acciones sobre un video. A diferencia del procesamiento *offline* realizado durante la preparación de datos (donde ya se disponía de secuencias de esqueletos precalculadas), en la inferencia en se integra el módulo de estimación de pose directamente en el flujo.

Es decir, el sistema toma los *frames* de la cámara y aplica *MediaPipe* en línea para extraer las poses, que luego son normalizadas y estructuradas en una ventana deslizante temporal compatible con la entrada del modelo ST-GCN entrenado. En términos prácticos, este proceso de inferencia continua incluye los siguientes pasos automatizados.

- El sistema está constantemente leyendo *frames* desde la cámara IP, manteniendo un *buffer* de los últimos T *frames* necesarios. Cada nuevo *frame* que llega desplaza al más antiguo en el *buffer*, manteniendo así una ventana actualizada de la secuencia reciente.
- Por cada frame capturado, se aplica el modelo de MediaPipe Pose para obtener el esqueleto 3D instantáneo. Conforme van llegando los *frames*, el sistema ensambla la secuencia de esqueletos en el formato esperado $[N, C, T, V]$ (muestras, canales, tiempo, vértices), creando un tensor donde $N = 8$ muestras $C = 3$ coordenadas, $T = 100$ número de *frames* en la ventana, $V = 25$ articulaciones. Este tensor representativo de la ventana

deslizante actual es normalizado (coordenadas centradas y escaladas) tal como se hizo con los datos de entrenamiento.

- Una vez listo el tensor de entrada de dimensión, el módulo de inferencia ejecuta el modelo entrenado sobre esa secuencia para predecir la acción que se está llevando a cabo en la ventana temporal.
- El sistema proyecta el resultado de la inferencia directamente sobre el video de salida. Esto incluye dibujar el esqueleto y mostrar una etiqueta con la acción reconocida y el porcentaje de confianza en dicha clasificación. De esta forma, un observador puede ver en la pantalla, casi simultáneamente al suceso, qué acción reconoce el sistema que está ocurriendo.
- Para los eventos de mayor interés (definidos según la aplicación), el sistema cuenta con un mecanismo de captura de evidencia visual. Por ejemplo, en esta aplicación enfocada en monitoreo de integridad física, la detección de la acción “*Caerse*” desencadena automáticamente la toma de una captura de pantalla (frame congelado) del momento de la caída. Esta imagen se almacena con sello de tiempo y etiqueta de la acción, lo que permite su posterior revisión.

En conjunto, todos estos componentes y etapas conforman un sistema modular construido a la medida de las necesidades experimentales del proyecto. En síntesis, la gestión del sistema se diseñó para proporcionar una base de control y trazabilidad, garantizando que los experimentos pudieran ejecutarse de manera continua, confiable y registrando toda la información necesaria para su análisis.

4.3 Balance y partición de datos

El conjunto de datos empleado en la experimentación no corresponde directamente a las secuencias esqueléticas provistas por NTU RGB+D (ver Figura 4.1), sino que fue construido de manera personalizada a partir del procesamiento de los videos originales con MediaPipe, lo que permitió extraer de forma independiente los *landmarks* corporales y generar representaciones consistentes bajo la misma topología de articulaciones.

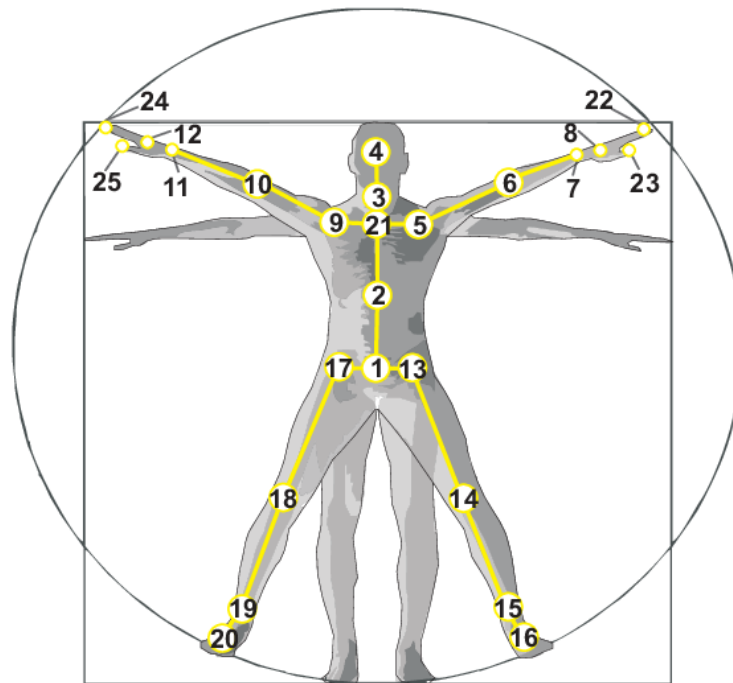


Figura 4.1 Modelo esquelético de articulaciones utilizado, tomado de [6]

Esta decisión otorgó mayor control sobre la calidad de los datos y aseguró la homogeneidad en las secuencias empleadas para el entrenamiento y la evaluación.

Posteriormente, se aplicó un proceso de balanceo con el propósito de evitar sesgos hacia clases con mayor número de ejemplos y garantizar que cada acción contara con la misma representatividad. El resultado fue un conjunto equilibrado de 948 secuencias por clase, sumando un total de 4,740 secuencias distribuidas en cinco categorías.

4.3.1 Acciones contempladas y justificación de selección

El conjunto de datos incluyó cinco clases de acciones relevantes para la aplicación de monitoreo de integridad física:

1. Sentarse
2. De pie
3. Saltar
4. Caerse
5. Dolor de pecho

La selección de estas acciones responde a dos criterios principales:

- Relevancia en escenarios de monitoreo de personas, especialmente en entornos donde se busca identificar comportamientos potencialmente peligrosos o que requieran asistencia inmediata (*Caerse, Dolor de pecho*).
- Diversidad de patrones de movimiento, que abarca desde posturas estáticas (*De pie, Sentarse*) hasta transiciones dinámicas rápidas (*Saltar, Caerse*), lo que permite evaluar la capacidad del modelo para diferenciar tanto cambios bruscos como sutiles en la postura.

4.3.2 Configuración y condiciones de captura

Las secuencias de video fueron capturadas mediante una cámara IP TP-Link Tapo C200, colocada a una altura y ángulo que permitieran una vista completa del cuerpo del participante en un espacio delimitado. La cámara se configuró para transmitir en resolución Full HD (1920×1080 píxeles) facilitando la integración directa con el *pipeline* de captura del sistema.

La iluminación del entorno se mantuvo uniforme y controlada para minimizar el ruido en la estimación de pose, pero se introdujeron ligeras variaciones (por ejemplo, cambios en la iluminación luz natural o artificial) con el fin de incrementar la variabilidad intra-clase y ayudar a mejorar el entrenamiento frente a condiciones que no sean las mismas.

Se utilizaron varios sujetos para realizar las acciones, con el objetivo de capturar diferencias antropométricas (altura, complexión, proporciones corporales) y patrones individuales de ejecución, lo que ayuda a que el modelo generalice mejor ante usuarios distintos.

4.3.3 Proceso de extracción de datos

Cada secuencia de video fue procesada mediante MediaPipe, obteniendo para cada frame un conjunto *landmarks* del cuerpo en coordenadas 3D. Sin embargo, para mantener compatibilidad con la topología estándar utilizada en modelos ST-GCN y conjuntos de datos como NTU RGB+D, se seleccionó un subconjunto, descartando las que no aportaban información significativa al reconocimiento de las acciones consideradas.

Posteriormente, los datos esqueléticos se organizaron en el formato tensorial requerido por el ST-GCN y se aplicó normalización espacial para centrar y escalar el esqueleto en cada secuencia (con respecto al punto medio de las caderas y distancia interarticular de referencia). Esto redujo la influencia de la posición absoluta del sujeto en el encuadre y favoreció la invariancia a la escala.

Los datos procesados fueron almacenados en archivos binarios que contienen las secuencias completas junto con su etiqueta de clase. Este formato optimiza el acceso y carga de datos durante el entrenamiento, evitando sobrecostos de decodificación de video.

4.4 Resultados experimentales

- *Accuracy* de 99,37% en el conjunto de prueba, indicando que casi todas las secuencias de acción fueron clasificadas correctamente.
- Métricas por clase: Los indicadores de *precision*, *recall* y *F1-Score* fueron cercanos a 100% (aproximadamente 99%) para todas las clases del conjunto de prueba, lo que refleja un desempeño uniforme y equilibrado entre las cinco acciones evaluadas. Cada clase obtuvo valores F1 superiores 98%.
- No se observaron indicios de sobreajuste durante la evaluación. El modelo mostró un desempeño consistente entre los datos de entrenamiento, validación y prueba, manteniendo curvas de pérdida paralelas y *accuracy* de validación estable en valores altos.

Los resultados anteriores confirman que el modelo ST-GCN entrenado aprendió a distinguir las acciones definidas. La matriz de confusión confirma que los errores de clasificación fueron mínimos y aislados. La gran mayoría de las secuencias fueron asignadas a su clase correcta sin problema. Las únicas confusiones residuales aparecieron principalmente entre las acciones Sentarse y De pie, dado que ambas comparten una postura estática similar.

Este resultado era esperado, ya que dichas clases representan poses corporales parecidas (estar sentado o de pie) y el modelo llegó a confundirlas en contadas ocasiones cuando la diferencia postural era sutil. Sin embargo, incluso en este caso, el número de errores fue muy bajo y no compromete el rendimiento general del sistema.

Por otro lado, para acciones con patrones de movimiento más distintivos (como Saltar o Caerse) el modelo no presentó prácticamente ninguna confusión. En particular, Caerse fue identificada correctamente en todas las secuencias de prueba, sin errores significativos.

Asimismo, la acción Dolor de pecho, a pesar de involucrar instantes en posición de pie, fue reconocida correctamente, la separación de esta clase respecto a otras fue clara, con apenas errores aislados atribuibles a similitudes posturales transitorias en la secuencia.

Estos resultados posicionan al modelo a la par del estado del arte en reconocimiento de acciones basado en esqueletos. De hecho, se alinean con hallazgos de investigaciones previas, por ejemplo, Ghosh *et al.* (2022) reportaron aproximadamente 98% de exactitud al detectar caídas usando un modelo ST-GCN entrenado sobre NTU RGB+D, con valores de *precision recall* por clase en el rango de 98% a 99%.

El desempeño obtenido en el presente trabajo para la clase Caerse (F1 aproximado de 0,99) es congruente con esos valores, confirmando que la arquitectura ST-GCN puede capturar eficazmente los patrones abruptos característicos de eventos como las caídas.

Para las posturas básicas y movimientos simples (como Sentarse, De pie y Saltar), los resultados obtenidos por el modelo se acercan al máximo teórico posible, con métricas rondando el 100% de acierto. Esto coincide con tendencias reportadas previamente.

Adicionalmente, se validó el funcionamiento del sistema mediante pruebas con una fuente. Al integrar el modelo entrenado en un flujo continuo de video (simulando una cámara IP en un ambiente controlado), se observó que el sistema mantiene su alta precisión sin degradación notable.

La inferencia se realiza de forma continua y estable con baja latencia, cumpliendo así requisitos de una aplicación práctica. El resultado es que el sistema puede procesar, detectando y alertando acciones de interés con mínima demora.

4.4.1 Experimento 1: Entrenamiento base con conjunto de datos balanceado

Se buscó establecer una línea base de desempeño del modelo ST-GCN entrenando con un conjunto de datos estrictamente balanceado y la configuración de hiperparámetros “base” definida en el Capítulo 3, para posteriormente comparar el efecto de modificaciones controladas en los Experimentos 2–9 (véanse Sección 3.4 y Sección 3.5).

Se empleó la arquitectura definida en el proyecto (ST-GCN con bloques espaciotemporales), el pipeline y los hiperparámetros base. El entrenamiento se planificó a 100 épocas y el mejor modelo se seleccionó por desempeño en validación. En esta ejecución base, el entrenamiento converge de manera estable a lo largo de las épocas.

La función *loss* disminuye de forma progresiva con estabilización después de la época 25 mientras que *accuracy* de entrenamiento alcanza 98.7% y la de validación

llega a 99.68% (época 94), con curvas suaves y sin indicios de sobreajuste (ver Figura 4.2).

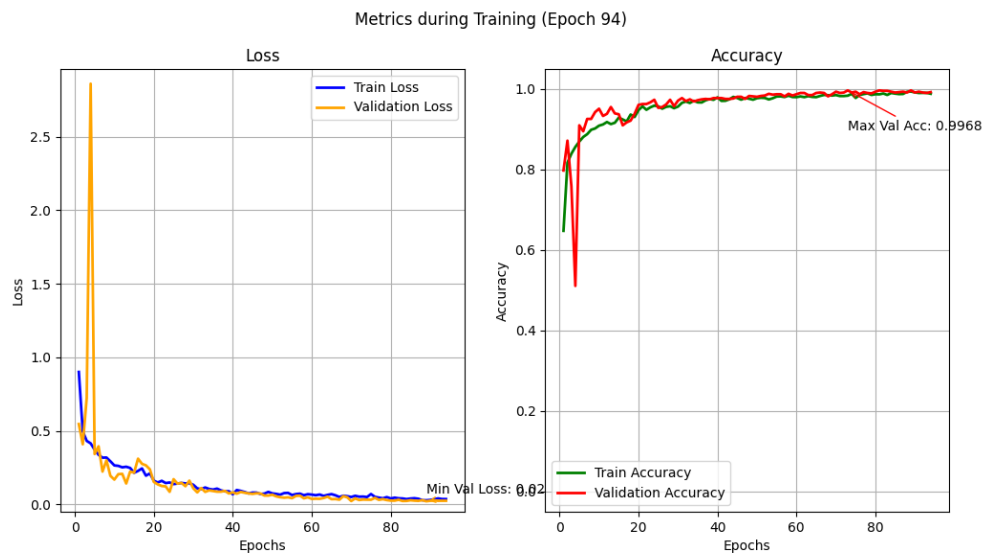


Figura 4.2 Comportamiento métricas durante el entrenamiento del modelo, por época.

En la configuración base, las cinco clases presentan $F1\text{-Score} > 98.4\%$, con “De pie” reconocido en todos los casos y un ligero descenso en *recall* para “Dolor de pecho” (ver Figura 4.3). La matriz de confusión sobre el conjunto de prueba reporta 480 muestras con solo 3 errores, una confusión entre las clases “Sentarse y Caerse” y dos para las clases “Dolor de pecho y De pie”; no se observan errores cruzados entre clases disímiles. Estos resultados evidencian un comportamiento equilibrado entre clases y una tasa de error reducida.



Figura 4.3 Comparativa de precisión, recall y F1 por clase

Conclusión del experimento. El entrenamiento base con datos balanceados confirma que el modelo aprende relaciones espaciotemporales entre articulaciones y generaliza adecuadamente, cumpliendo el objetivo de establecer una línea base sólida para las comparaciones subsecuentes.

4.4.2 Experimento 2: Evaluación del impacto de la longitud de la secuencia temporal

Se analizó cómo influye la longitud de las secuencias de entrada (medida en número de *frames* consecutivos) en el rendimiento del modelo ST-GCN, determinando un valor óptimo que capture suficiente información temporal para cada acción sin introducir redundancia ni sobrecarga computacional.

Diseño experimental. Partiendo de la configuración base definida en el Experimento 1, se mantuvieron constantes todos los hiperparámetros, el preprocesamiento y el conjunto de datos, variando únicamente la longitud de la ventana temporal utilizada como entrada. Se evaluaron tres configuraciones:

- 50 *frames*: ventana corta, diseñada para capturar patrones de movimientos rápidos y posturas simples, reduciendo la latencia de procesamiento.
- 100 *frames*: valor de referencia proveniente del Experimento 1, que representa un equilibrio entre cobertura temporal y eficiencia computacional.
- 150 *frames*: ventana larga, orientada a maximizar el contexto temporal, potencialmente beneficiosa para acciones de desarrollo prolongado.

Esta variación permitió aislar el efecto de la longitud temporal sobre la capacidad discriminativa del modelo y su estabilidad en el reconocimiento de acciones.

Resultados.

- 50 *frames*: la reducción drástica del contexto temporal afectó de manera negativa las métricas de *recall* y *F1-Score* en clases que requieren observar una secuencia completa para identificar la acción, como Caerse y Dolor de pecho. Aunque el *accuracy* global se mantuvo alto (> 95%), se registró un aumento de errores en transiciones rápidas que no quedaron completamente representadas en la ventana.
- 100 *frames*: este valor confirmó ser la mejor opción. Se alcanzaron métricas máximas o muy cercanas a las máximas para todas las clases, con *accuracy* y *F1-Score* promedio superiores al 99%. La ventana cubrió de forma completa las fases críticas de acciones tanto rápidas como prolongadas, sin penalizar el tiempo de entrenamiento ni la latencia en inferencia.

- 150 *frames*: el incremento de contexto no supuso mejoras significativas en *accuracy* o *F1-Score* respecto a la configuración de 100 *frames*. En cambio, aumentó el tiempo de entrenamiento y de inferencia, y en clases estáticas como De pie y Sentarse se detectó una ligera disminución de precisión, atribuible a la inclusión de información redundante y ruido temporal.

El análisis evidencia que 100 *frames* constituye la mejor opción para este escenario, al ofrecer un balance entre información temporal suficiente y eficiencia operativa. Valores inferiores limitan la capacidad del modelo para capturar transiciones completas en acciones complejas, mientras que valores superiores añaden redundancia y penalizan el rendimiento computacional sin mejoras sustanciales en la clasificación. Por estas razones, la configuración de 100 *frames* se adopta como estándar para los experimentos posteriores.

4.4.3 Experimento 3: Evaluación del modelo propuesto con normalización espacial de esqueletos

Se planteó determinar el impacto de la normalización espacial de las coordenadas esqueléticas sobre el rendimiento del modelo ST-GCN, evaluando si esta técnica mejora la consistencia de la representación espacial y reduce las variaciones debidas a diferencias en la posición o proporciones corporales de los sujetos.

La normalización espacial consistió en reescalar y centrar las coordenadas articulares de cada frame respecto a un punto de referencia del esqueleto (articulación raíz), ajustando la escala para que las distancias relativas entre articulaciones se mantuvieran proporcionales. Esta técnica buscó atenuar la variabilidad introducida por cambios en la distancia a la cámara o en la complejidad física del sujeto.

Se compararon dos configuraciones:

- Sin normalización espacial: uso de las coordenadas originales obtenidas por el estimador de pose (*baseline*).
- Con normalización espacial: aplicación del procedimiento de centrado y escalado uniforme antes de alimentar los datos al ST-GCN.

El resto de las condiciones de entrenamiento, longitud de secuencia (100 *frames*), tasa de aprendizaje y tamaño de lote se mantuvieron idénticas a las del Experimento 1 para aislar el efecto de la normalización.

- Sin normalización el modelo alcanzó un *accuracy* promedio de 99.37%, con métricas F1 por clase entre 98% y 99%, confirmando un rendimiento alto, pero con ligeras fluctuaciones entre clases con mayor variabilidad postural.

- Con normalización, se observó un incremento marginal en la estabilidad de las métricas entre clases, con reducciones en la dispersión de los valores F1. El *accuracy* global se mantuvo prácticamente idéntico, pero en acciones susceptibles a variaciones espaciales como Sentarse y Dolor de pecho, se redujeron los errores aislados presentes en la configuración sin normalizar.

La normalización espacial no modificó de forma significativa el *accuracy* global, pero contribuyó a uniformar el desempeño por clase y a mitigar errores puntuales debidos a desplazamientos o diferencias en proporciones corporales. Esto sugiere que, si bien el ST-GCN ya es robusto frente a variaciones espaciales moderadas, la normalización puede aportar una ganancia adicional en escenarios con mayor diversidad física o de encuadre. Debido a su bajo coste computacional y potencial para mejorar la generalización, esta técnica se considera adecuada para su inclusión en configuraciones destinadas a entornos más heterogéneos.

4.4.4 Experimento 4: Evaluación del modelo propuesto con aumento de datos

Por medio de este experimento se buscó el analizar el efecto de aplicar técnicas de aumento de datos (*data augmentation*) sobre el rendimiento del modelo ST-GCN, evaluando si la generación de variaciones sintéticas en las secuencias esqueléticas mejora la capacidad de generalización y reduce el riesgo de sobreajuste.

Se partió de la configuración base definida en el Experimento 1 (secuencias de 100 *frames*, tamaño de lote y tasa de aprendizaje estándar), incorporando un conjunto de transformaciones aleatorias sobre los datos de entrada durante el entrenamiento.

Las técnicas empleadas incluyeron:

- Rotaciones espaciales aleatorias en torno al eje vertical, simulando cambios de perspectiva moderados.
- Escalado uniforme con factores ligeramente superiores o inferiores a la unidad, replicando variaciones de distancia a la cámara.
- Traslaciones aleatorias de las coordenadas, preservando la estructura del esqueleto, pero alterando su posición en el plano.
- Perturbaciones gaussianas de baja magnitud sobre las coordenadas articulares, simulando ruido del estimador de pose.

Estas modificaciones se aplicaron en tiempo real durante el entrenamiento, garantizando que cada lote de datos incluyera instancias originales y variantes aumentadas. Las condiciones de evaluación se mantuvieron sin aumentos para medir el efecto neto de la técnica sobre datos no vistos.

- El *accuracy* global se mantuvo prácticamente inalterado respecto a la configuración sin aumentos (99.37%), evidenciando que el modelo ya

presentaba una alta capacidad de generalización en el conjunto de prueba original.

- Sin embargo, se detectó una ligera mejora en la estabilidad de las métricas *recall* y F1 para clases menos frecuentes o con mayor variabilidad postural, como Dolor de pecho, reduciendo aún más los pocos errores presentes en la configuración base.
- El tiempo total de entrenamiento aumentó debido al preprocesamiento adicional requerido por la generación de datos aumentados, sin impacto perceptible en la latencia de inferencia.

El aumento de datos aplicado a secuencias esqueléticas no generó mejoras significativas en *accuracy* debido al alto rendimiento inicial del modelo, pero sí aportó beneficios marginales en la homogeneidad del desempeño por clase. Esta técnica puede resultar especialmente útil en escenarios donde se disponga de un conjunto de datos reducido o desequilibrado, o en despliegues donde se prevean condiciones de captura más variadas.

En el contexto del presente estudio, su uso se considera recomendable como medida preventiva para robustecer la capacidad de generalización del sistema frente a variaciones moderadas en pose, encuadre y ruido de estimación.

4.4.5 Experimento 5: Evaluación del tamaño de lote

Se examinó cómo influye la variación del tamaño de lote en el entrenamiento del modelo ST-GCN, tanto en el tiempo de convergencia como en el rendimiento final sobre el conjunto de prueba.

A partir de la configuración base del Experimento 1, se entrenó el modelo con diferentes tamaños de lote: 8, 16 y 32 secuencias por iteración. Todos los demás parámetros (número de épocas, tasa de aprendizaje, arquitectura, conjunto de datos y preprocesamiento) se mantuvieron constantes para aislar el efecto del *batch size*.

- *Batch* pequeño (8): Mayor número de actualizaciones de gradiente por época, permitiendo una adaptación más fina de los pesos, pero con mayor tiempo total de entrenamiento.
- *Batch* intermedio (16): Configuración estándar utilizada como referencia.
- *Batch* grande (32): Menor número de actualizaciones por época y mejor aprovechamiento de la paralelización en GPU, reduciendo el tiempo de entrenamiento por época.

Resultados.

- Precisión y métricas por clase: No se observaron variaciones significativas en *accuracy* global, que se mantuvo en torno al 99%. Sin embargo, el lote de

tamaño 8 presentó ligeras mejoras en *F1-Score* para las clases “Dolor de pecho” y “Caerse”, posiblemente debido a una actualización más frecuente de parámetros.

- Estabilidad del entrenamiento: Con *batch size* de 32 se observó una convergencia más rápida, pero con fluctuaciones algo mayores en la curva de validación en las primeras épocas, estabilizándose posteriormente.
- Tiempo de entrenamiento: El lote de tamaño 32 redujo el tiempo total por época en aproximadamente 30% frente al lote de tamaño 8, aprovechando mejor la capacidad de la GPU.
- Latencia en inferencia: Sin cambios relevantes, ya que el *batch size* afecta únicamente el entrenamiento.

El tamaño de lote sí influyó en el tiempo y la estabilidad del entrenamiento. Un lote intermedio ofreció un balance adecuado entre aprovechamiento de GPU, estabilidad de convergencia y rendimiento por clase, por lo que se mantuvo como la configuración utilizada para entrenamientos posteriores.

4.4.6 Experimento 6: Evaluación del número de épocas

Se analizó el impacto del número total de épocas de entrenamiento en el rendimiento final del modelo ST-GCN, identificando un punto de convergencia que maximice la precisión sin incurrir en sobreajuste ni desperdicio de recursos computacionales.

Se evaluaron tres configuraciones con 50, 75 y 100 épocas, manteniendo inalterados todos los demás parámetros de la configuración base (tasa de aprendizaje, tamaño de lote, conjunto de datos y arquitectura). Esta comparación permitió aislar el efecto del tiempo total de entrenamiento sobre el desempeño del modelo.

- 50 épocas: Entrenamiento más corto, con menor número de iteraciones para la optimización de parámetros.
- 75 épocas: Configuración intermedia, buscando un punto de convergencia razonable sin llegar al máximo posible.
- 100 épocas: Entrenamiento prolongado para permitir el ajuste fino de los pesos y potencialmente alcanzar el techo de rendimiento.

El análisis de las curvas de pérdida y precisión mostró un comportamiento claro, con 50 épocas, el modelo alcanzó un rendimiento aceptable pero todavía en fase ascendente, evidenciando que no había explotado por completo su capacidad de generalización.

En 75 épocas, la precisión mejoró, con valores de *accuracy* cercanos al 99.4% y F1 macro igualmente alto, aunque con una ligera tendencia a seguir mejorando.

Finalmente, con 100 épocas se alcanzó el mejor porcentaje de *accuracy* llegando a 99.59% y F1 macro del 99.58%, sin indicios de sobreajuste, ya que las curvas de validación se mantuvieron estables y paralelas a las de entrenamiento.

A nivel de rendimiento por clase, las mejoras de 75 a 100 épocas fueron marginales en acciones como Saltar y Caerse, que ya mostraban métricas cercanas al 100% desde configuraciones más cortas. Sin embargo, para clases con mayor similitud postural, como Sentarse y De pie, el incremento en el número de épocas redujo aún más los errores residuales, especialmente en secuencias con cambios de postura más sutiles.

Entrenar durante 100 épocas ofrece el mejor equilibrio entre precisión y estabilidad. Configuraciones más cortas (50 y 75 épocas) pueden considerarse aceptables en entornos con restricciones de tiempo o recursos, pero sacrifican una pequeña fracción de rendimiento, particularmente en clases con alta similitud postural.

4.4.7 Experimento 7: Comparación de tasa de aprendizaje

Se evaluó el impacto de diferentes valores de tasa de aprendizaje en la velocidad de convergencia y en la capacidad de generalización del modelo ST-GCN, con el fin de determinar el mejor valor posible para la implementación.

Se realizaron pruebas variando exclusivamente la tasa de aprendizaje, manteniendo constante el resto de la configuración experimental (100 épocas, tamaño de lote y conjunto de datos balanceado). Los valores analizados fueron 0.01, 0.001 y 0.0001, seleccionados por representar tres órdenes de magnitud que permiten observar el comportamiento del optimizador ante cambios amplios en este hiperparámetro:

- 0.01: tasa alta que puede acelerar el aprendizaje inicial, pero con riesgo de oscilaciones y convergencia inestable.
- 0.001: valor intermedio, comúnmente utilizado como referencia en modelos de redes neuronales profundas.
- 0.0001: tasa baja, favorece la convergencia estable, pero puede requerir más iteraciones para alcanzar el rendimiento máximo.

Con 0.01, el modelo mostró una convergencia rápida en las primeras épocas, pero con fluctuaciones significativas en la curva de pérdida de validación, lo que indica una tendencia a sobrepasar los mínimos locales. Aunque el *accuracy* final alcanzó 99.28%, el F1 macro fue ligeramente inferior al de otras configuraciones debido a pequeñas caídas de rendimiento en clases con patrones posturales similares.

Con 0.001, el entrenamiento fue más estable, con una trayectoria suave de reducción de la pérdida y un balance óptimo entre velocidad de convergencia y estabilidad. Esta configuración logró un *accuracy* de 99.59% y un F1 macro de 99.58%, constituyendo el mejor resultado global entre las tasas probadas.

Con 0.0001, las curvas de entrenamiento fueron extremadamente estables, pero el progreso fue más lento y el rendimiento final (99.37% de *accuracy*) quedó por debajo del valor óptimo, reflejando que el modelo no alcanzó a ajustar completamente sus parámetros dentro del número de épocas fijado.

Conclusión. El valor de 0.001 se determinó como el más adecuado para el entrenamiento del modelo ST-GCN, combinando alta precisión final, estabilidad en la convergencia y ausencia de comportamientos erráticos. Tasas más altas (0.01) pueden ser útiles en fases iniciales de ajuste rápido, pero requieren estrategias adicionales como *learning rate scheduling* para evitar oscilaciones. Tasas más bajas (0.0001) ofrecen estabilidad extrema, aunque a costa de mayor tiempo de entrenamiento y menor rendimiento final en configuraciones de épocas limitadas.

4.4.8 Experimento 8: Evaluación del desempeño por clase

Se buscó analizar el rendimiento del modelo ST-GCN para cada una de las clases definidas en el conjunto de datos, con el fin de identificar posibles diferencias en la capacidad de clasificación según el tipo de acción.

A partir del modelo entrenado en la mejor configuración determinada en los experimentos previos, se evaluó el desempeño por clase utilizando las métricas *precision*, *recall* y *F1-Score*. El conjunto de prueba estuvo balanceado, con un número similar de muestras por clase, lo que permitió obtener una comparación justa del rendimiento ver Figura 4.4.

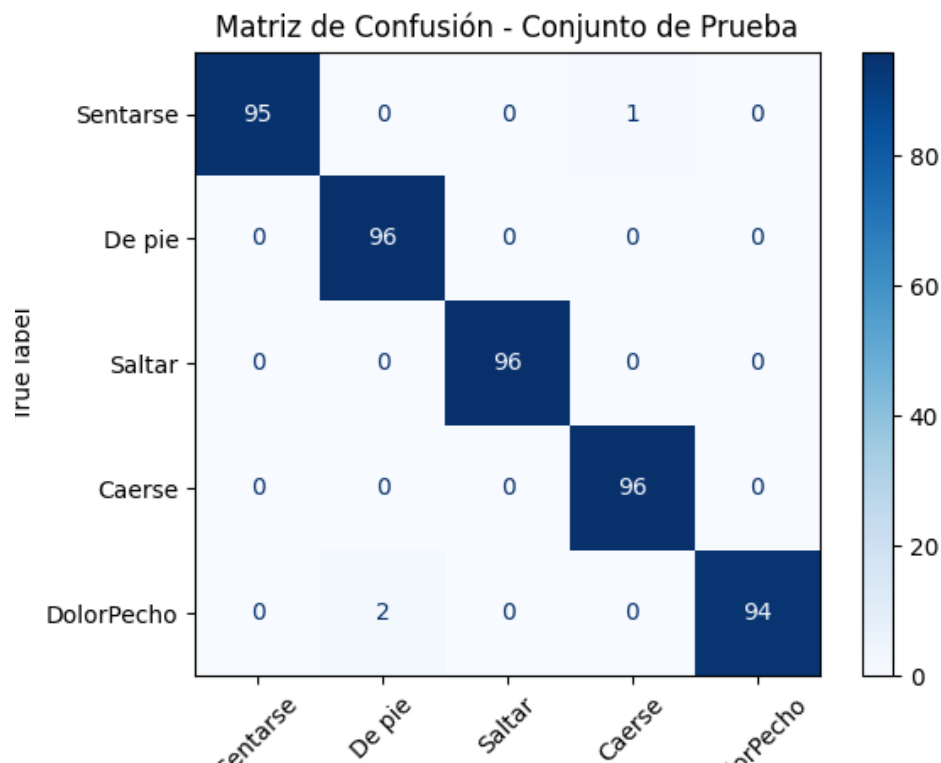


Figura 4.4 Matriz de confusión generada a partir de las predicciones del modelo propuesto sobre el conjunto

Resultados.

- Sentarse: Todas las métricas se situaron en valores cercanos al 100%, con muy pocos errores de clasificación. Las confusiones residuales ocurrieron principalmente con la clase De pie, debido a que ambas corresponden a posturas estáticas con diferencias sutiles en la disposición de las articulaciones.
- De pie: Presentó un rendimiento igualmente alto, aunque con ligeras disminuciones en *precision* frente a la clase Sentarse por el mismo motivo de similitud postural.
- Saltar: Obtuvo métricas sin errores registrados, confirmando que los patrones cinemáticos distintivos de esta acción fueron capturados adecuadamente por el modelo.
- Caerse: Fue reconocida sin errores significativos, alcanzando valores de *F1-Score* de 99%, coherentes con estudios previos que reportan un desempeño sobresaliente del ST-GCN en la detección de caídas.
- Dolor de pecho: Mostró un rendimiento alto, ligeramente inferior al de las clases más estáticas debido a la presencia de transiciones posturales que pueden asemejarse a otras acciones. Aun así, las métricas superaron 98%, confirmando que la arquitectura pudo diferenciar esta acción de manera confiable.

El análisis por clase confirma que el modelo mantiene un rendimiento consistente y elevado en todas las categorías, con especial precisión en acciones de patrones cinemáticos distintivos como Saltar y Caerse, y con confusiones mínimas entre posturas estáticas. Estas observaciones, en conjunto con los resultados globales, validan la robustez del ST-GCN en escenarios de monitoreo en tiempo real con acciones variadas y condiciones controladas.

4.4.9 Experimento 9: Inferencia mediante voto por mayoría en ventana deslizante

Por medio del presente experimento se buscó evaluar el impacto del uso de un esquema de voto por mayoría sobre una ventana deslizante de predicciones consecutivas, con el fin de reducir fluctuaciones en la clasificación y mejorar la estabilidad de las decisiones en entornos de inferencia.

Se implementó un mecanismo de agregación temporal en el que las predicciones a nivel de frame se acumulan en una ventana deslizante de tamaño fijo. La clase final para la ventana se determina por mayoría simple de las predicciones individuales. Posteriormente, la ventana avanza un número determinado de *frames* y el proceso se repite.

El diseño del experimento incluyó:

- Uso del modelo ST-GCN entrenado con la configuración óptima identificada en experimentos previos.
- Pruebas sobre secuencias de video del conjunto de prueba y sobre flujo de video.
- Comparación del rendimiento con y sin el esquema de voto por mayoría.

La incorporación de la ventana deslizante permitió:

- Reducción de falsos positivos generados por clasificaciones espurias en *frames* individuales afectados por ruido en la estimación de pose.
- Mayor estabilidad en la predicción de acciones estáticas (“Sentarse”, “De pie”), donde la variabilidad frame a frame era más evidente.
- Mejora perceptible, ya que se evita que variaciones muy pequeñas en la postura produzcan cambios abruptos en la etiqueta mostrada al usuario.

Los resultados mostraron que la precisión global y el *F1-Score* experimentaron una mejora ligera pero consistente con el uso del voto por mayoría, manteniendo el tiempo de respuesta del sistema dentro de márgenes aceptables para aplicaciones en tiempo real (latencia total de 1 a 2 segundos).

Este enfoque actúa como un filtro temporal que suaviza la salida del clasificador, lo que es especialmente beneficioso en sistemas de monitoreo continuo donde una predicción errática puede generar alarmas innecesarias.

El uso del voto por mayoría en ventana deslizante es una estrategia efectiva para incrementar la estabilidad de las predicciones sin sacrificar la capacidad de respuesta del sistema, ya que es recomendable en escenarios de implementación práctica, donde la coherencia temporal de la salida es tan importante como la exactitud de la clasificación.

4.4.10 Experimento 10: Sistema de inferencia

Tras seleccionar la mejor configuración del modelo se integró el módulo de inferencia en flujo continuo para validar su funcionamiento en un escenario práctico. El objetivo fue comprobar que el rendimiento observado en la fase offline se mantuviera en condiciones de operación en vivo, utilizando como fuente de video una cámara IP.

El flujo de inferencia se implementó como un proceso secuencial y continuo, compuesto por cuatro etapas principales:

- Captura de imágenes y extracción de cuadros. Un proceso dedicado recibe la transmisión RTSP de la cámara IP y extrae *frames* individuales en tiempo real para su posterior procesamiento.
- Estimación de pose por cuadro. Cada imagen se procesa con MediaPipe Pose, obteniendo las coordenadas de los puntos clave corporales en formato 2D y 3D.
- Construcción del grafo esquelético. Los *landmarks* extraídos se adaptan a una topología de 25 nodos, coherente con la utilizada en el conjunto NTU RGB+D, garantizando la compatibilidad entre la representación usada en el entrenamiento y la inferencia.
- Clasificación de acciones. El modelo ST-GCN entrenado procesa secuencias temporales fijas, calculando un vector de probabilidades sobre las clases definidas.

Para la etapa de inferencia se aplicó un esquema de ventanas temporales de longitud fija (definida en el experimento 2), con 100 *frames* por ventana. El procesamiento se realizó sobre ventanas deslizantes con solapamiento, permitiendo una clasificación independiente por cada segmento y posteriormente aplicando un voto por mayoría entre predicciones consecutivas.

Este mecanismo reduce la inestabilidad en la salida, evitando fluctuaciones rápidas debidas a poses intermedias o transiciones incompletas. Además, se estableció que un cambio de clase solo se produce si la nueva predicción se mantiene de forma consistente durante una ventana temporal suficiente, lo que mejora la fiabilidad en la detección de eventos como “Caerse” o “Dolor de pecho” ver Figura 4.5.

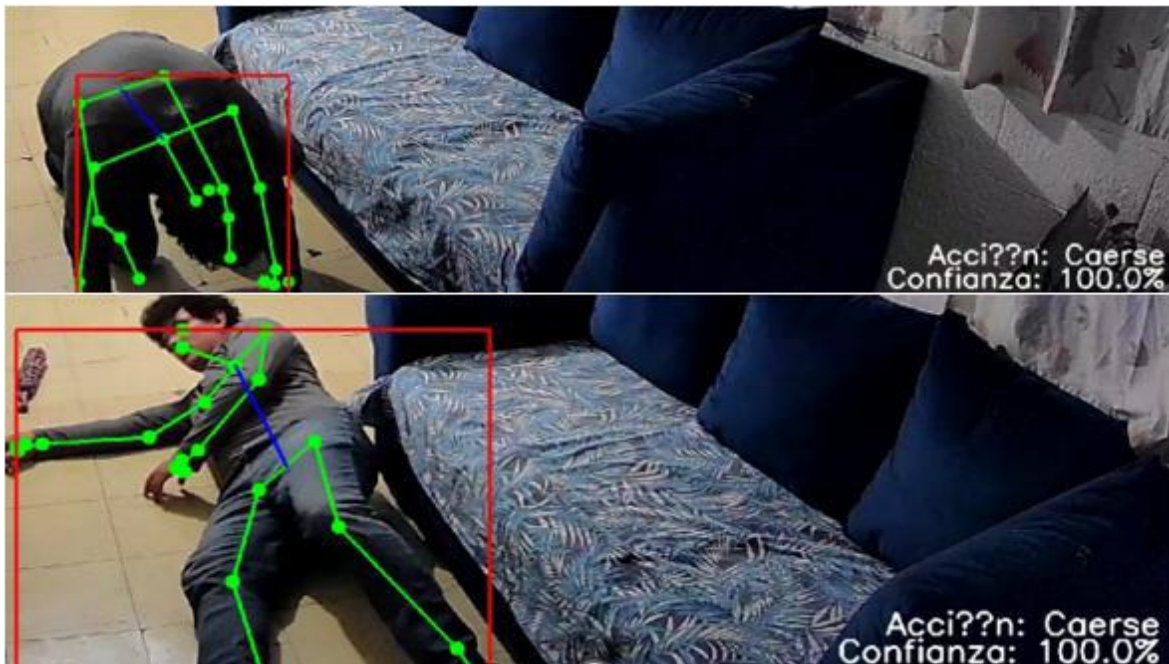


Figura 4.5 Reconocimiento de la acción “Caerse” con alerta visual

El sistema superpone en el video la representación esquelética y la etiqueta de la acción detectada junto con su nivel de confianza. En el caso de acciones críticas, como *Caerse*, se resalta el esqueleto con un recuadro de color rojo, registrando simultáneamente el segmento de video como evidencia para auditorías posteriores.

Durante las pruebas, el sistema:

- Mantuvo la precisión registrada en la fase offline, con una diagonal dominante en la matriz de confusión y métricas macro entre 99.4% y 99.6%, coherentes con los resultados obtenidos en pruebas controladas.
- Redujo las variaciones cuadro a cuadro gracias al voto por mayoría, disminuyendo falsos positivos en acciones con posturas similares.
- Procesó la transmisión RTSP sin interrupciones, realizando la estimación de pose, la conversión a grafo esquelético y la clasificación de manera continua, con una latencia total aproximada de 1 a 2 segundos desde la captura del frame hasta la presentación del resultado.

El sistema de inferencia basado en *RTSP + MediaPipe Pose + ST-GCN* con ventanas deslizantes y voto por mayoría ofrece salidas aceptables y estables en el tiempo, preservando el rendimiento y añadiendo robustez para su uso continuo. La combinación de una topología específica, junto con el filtrado temporal mediante voto por mayoría, asegura que la detección de acciones sea confiable incluso en

presencia de transiciones de postura rápidas o condiciones de iluminación variables. Estos resultados respaldan su viabilidad para monitoreo.

4.5 Análisis de los Resultados Obtenidos

El desempeño global del modelo alcanzó una precisión promedio del 99.37% en el conjunto de prueba, con métricas de *precision*, *recall* y *F1-Score* cercanas al 1.0 (aprox. 99%) en todas las clases, lo que refleja un rendimiento equilibrado para las cinco acciones evaluadas, el modelo mostró comportamientos consistentes en entrenamiento, validación y prueba, manteniendo curvas de la función *loss* paralelas y *accuracy* de validación altas y estables. Confirmando que la red aprendió las diferencias espaciotemporales entre las acciones definidas.

A nivel de clases específicas, los errores de clasificación fueron mínimos y aislados. La gran mayoría de las secuencias de prueba fueron asignadas correctamente a su categoría, sin confusiones entre acciones diferentes. Las únicas confusiones aparecieron entre las acciones “Sentarse” y “De pie”, debido a que comparten posturas muy similares.

Por otro lado, en acciones con patrones de movimiento muy distintivos (por ejemplo “Saltar” y “Caerse”) el modelo no presentó confusiones. En particular, “Caerse” fue identificada correctamente en la mayor parte de las secuencias de prueba, alcanzando un F1 aproximado de 0.99. Del mismo modo, la acción “Dolor de pecho”, a pesar de involucrar transiciones desde una postura de pie, fue reconocida sin muchas complicaciones ($> 98\%$ F1); solo se observaron errores atribuibles a similitudes en ciertas posturas entre acciones, sin impactar la detección.

Adicionalmente, el modelo mantiene un rendimiento elevado en todas las categorías, y se observó una reducción de falsos positivos y de variaciones erráticas entre frames consecutivos. En acciones estáticas, donde pequeñas fluctuaciones de postura podían hacer oscilar la clasificación *frame a frame*, el voto por mayoría logró suavizar la salida del modelo, haciendo que la etiqueta final sea más coherente en el tiempo.

4.6 Interpretación sobre la Aplicabilidad Práctica

Los hallazgos anteriores resaltan la opción de aplicar este sistema de reconocimiento de acciones en contextos reales de monitoreo. En las pruebas de integración se comprobó que el modelo puede procesar video en flujo continuo desde una cámara IP, manteniendo su rendimiento (aproximadamente 99%

accuracy). La inferencia se ejecuta de forma estable con una latencia total de solo 1 segundo, lo cual es adecuado para detectar y alertar eventos.

Esto significa que, en un escenario práctico, el sistema podría *monitorear* a personas en un entorno (por ejemplo, un hogar o clínica) e identificar de manera confiable acciones de interés, generando alarmas o alertas con mínima demora. La capacidad de respuesta y la precisión en acciones garantizan un impacto positivo como, permitir la detección oportuna de emergencias médicas o accidentes, facilitando que cuidadores o personal de seguridad reaccionen adecuadamente ante ello.

Es importante destacar que el diseño basado en esqueletos + ST-GCN conlleva ventajas, al operar sobre coordenadas esqueléticas (provenientes de MediaPipe u otro estimador de pose), el sistema puede permanecer invariante a las condiciones visuales de la escena, enfocándose en la estructura del movimiento humano. De hecho, los resultados mostraron que la combinación de una estructura esquelética específica con el filtrado temporal por mayoría produce detecciones confiables incluso ante variaciones moderadas del entorno, como cambios ligeros de iluminación o diferencias en la complexión/posición de la persona, sin necesidad de re-entrenamiento. Esto indica que el sistema puede generalizar razonablemente en entornos distintos al entrenado, siempre que dichas variaciones sean de magnitud similar a las cubiertas en los datos. En entornos controlados, el modelo mostró ser consistente y robusto en su funcionamiento continuo. El uso de ventanas deslizantes con voto por mayoría reforzó la estabilidad de las alertas en la práctica, evitando falsas alarmas debidas a fluctuaciones.

Otro aspecto importante es que la solución propuesta no requiere sensores adicionales en el cuerpo ni equipamiento especializado más allá de una cámara RGB estándar, lo cual simplifica su implementación. Los resultados obtenidos reafirman que los modelos de grafos espaciotemporales superan en desempeño a enfoques tradicionales basados únicamente en video RGB 2D o en sensores portátiles, ofreciendo una precisión mayor y una configuración menos intrusiva.

Esto sugiere que, para la tarea de reconocimiento de acciones en ambientes controlados, la aproximación con pose estimada y grafos esqueléticos no solo es eficaz sino también práctica debido a que aprovecha información corporal estructurada y tolera variaciones moderadas sin necesidad de ajustes excesivos.

Si bien habría que considerar pruebas adicionales en contextos más complejos (mayor número de personas en escena, entornos no controlados, etc.), en el ámbito para el cual fue diseñado el sistema cumple con los requerimientos prácticos.

Capítulo 5

Conclusiones

La investigación realizada permitió consolidar un enfoque basado en grafos espaciotemporales para el reconocimiento de acciones humanas a partir de secuencias esqueléticas. La experimentación validó que la arquitectura ST-GCN, combinada con un preprocesamiento adecuado, normalización espacial y técnicas de aumento de datos, ofrece un rendimiento altamente consistente tanto en evaluación offline como en inferencia en tiempo real.

La capacidad del sistema para procesar video en flujo continuo, manteniendo una precisión cercana al 100%, y de esta manera confirma que los modelos a base de estas técnicas son una opción adecuada para la detección oportuna de eventos.

5.1 Conclusiones generales

El desarrollo y validación del sistema de reconocimiento de acciones evidenciaron la eficacia de integrar estimación de pose con MediaPipe Pose, representación esquelética estandarizada y procesamiento mediante un ST-GCN optimizado.

Los experimentos realizados alcanzaron métricas de *accuracy*, *precision*, *recall* y F1 superiores al 99% en un conjunto de acciones reducido que incluye eventos como caídas y dolor de pecho. El análisis de hiperparámetros permitió determinar configuraciones que maximizan el rendimiento sin comprometer la latencia en la inferencia.

En el escenario de operación en tiempo real, el sistema mantuvo la estabilidad de las predicciones gracias al uso de ventanas deslizantes y voto por mayoría, reduciendo falsos positivos y evitando transiciones continuas entre clases. Esto, sumado a la baja latencia observada, respalda su viabilidad para un despliegue continuo en entornos controlados.

Los resultados obtenidos reafirman que los modelos basados en grafos espaciotemporales no solo superan en desempeño a enfoques tradicionales con video RGB o sensores portátiles, sino que también ofrecen una adaptabilidad notable frente a variaciones moderadas del entorno y de las condiciones de captura.

5.2 Objetivos alcanzados

El trabajo cumplió con los objetivos planteados al inicio, logrando el diseño, implementación y validación de un sistema capaz de identificar acciones humanas

a partir de secuencias esqueléticas en entornos residenciales. Entre los logros más relevantes se incluyen:

- Desarrollo de un *pipeline* de reconocimiento de acciones que integra captura de video mediante RTSP, estimación de pose con MediaPipe Pose, construcción de grafos esqueléticos y clasificación con un modelo ST-GCN.
- Optimización de hiperparámetros clave, incluyendo la longitud de la secuencia temporal, número de épocas, tamaño de lote y tasa de aprendizaje, a través de experimentación sistemática y análisis comparativo de métricas.
- Implementación de técnicas de preprocesamiento y aumento de datos que contribuyeron a mejorar la capacidad de generalización del modelo, reduciendo el impacto de variaciones en postura, iluminación y fondo.
- Validación del sistema en tiempo real, asegurando que el rendimiento observado en la evaluación offline se mantuviera en la operación continua, con latencias bajas y alta estabilidad en la predicción.
- Diseño de un esquema de inferencia robusto mediante el uso de ventanas deslizantes y voto por mayoría, lo que permitió minimizar falsos positivos y asegurar transiciones coherentes entre acciones.

Estos resultados muestran que se alcanzó el propósito central de dotar al modelo de estabilidad operacional y capacidad de respuesta (ver Tabla 5.1), cumpliendo con los requerimientos funcionales y de desempeño establecidos.

Tabla 5.1 Síntesis de objetivos y alcances definidos para el desarrollo y evaluación del modelo

Objetivo/Alcance	Descripción
Objetivo 1: Recopilar un conjunto de datos con patrones de movimiento.	Se empleó un conjunto de datos estructurado con secuencias esqueléticas que representan acciones humanas específicas.
Objetivo 2: Comparar modelos de reconocimiento de patrones.	Se revisaron y evaluaron configuraciones de modelos para determinar su rendimiento, seleccionando como núcleo del sistema la arquitectura ST-GCN por su capacidad para preservar la estructura espacial y temporal del esqueleto humano.

Tabla 5.1 Síntesis de objetivos y alcances definidos para el desarrollo y evaluación del modelo (Continuación)

Objetivo 3: Implementar un modelo de reconocimiento de patrones.	Se desarrolló e implementó un modelo ST-GCN, integrándolo a un flujo de trabajo completo de captura, procesamiento y clasificación de acciones.
Objetivo 4: Evaluar la confiabilidad del modelo.	Se aplicaron métricas de rendimiento como <i>precision</i> , <i>recall</i> , <i>F1-Score</i> y <i>accuracy</i> para cuantificar el desempeño del modelo bajo diferentes configuraciones experimentales, verificando la robustez de las predicciones en condiciones de funcionamiento continuo.
Alcance 1: Reconocer al menos tres acciones definidas en el sistema.	El modelo identificó con alta precisión al menos tres de las acciones previstas, incluyendo eventos críticos como “Caerse” y “Dolor de pecho”, cumpliendo con la meta establecida para este alcance.
Alcance 2: Clasificar las acciones detectadas según su tipo.	Las acciones fueron categorizadas y asociadas a su correspondiente clase dentro del esquema del sistema, asegurando la correcta diferenciación entre posturas estáticas y movimientos dinámicos.
Alcance 3: Análisis basado en características del conjunto de datos.	Se realizó un análisis considerando tanto la estructura espacial de los esqueletos como la dinámica temporal, garantizando la coherencia entre las acciones etiquetadas y las predicciones realizadas.
Alcance 4: Detección centrada en la presencia de patrones definidos.	Las acciones detectadas fueron reportadas en función de las características definidas en la etapa de diseño del sistema, cumpliendo con los requerimientos operativos para el monitoreo continuo.

5.3 Productos generados

Como parte del desarrollo de esta tesis, se elaboraron diversos productos académicos, los cuales se detallan a continuación:

- a) Ponencia del tema “Descripción automática de acciones en secuencias de imágenes” en el Coloquio de Inteligencia Artificial, llevado a cabo del 17 al 19 de abril de 2024 organizado por el Tecnológico Nacional de México en las instalaciones del Instituto Tecnológico de Chilpancingo.
- b) Participación como personal de apoyo en la organización y desarrollo del 3er Congreso Internacional de Tecnología y Ciencia Aplicada (CITCA), llevado a cabo del 19 al 22 de noviembre de 2024 en las instalaciones del Tecnológico Nacional de México campus Centro nacional de Investigación y Desarrollo Tecnológico (TecNM/CENIDET)
- c) Presentación y publicación del artículo “Descripción automática de acciones en secuencias de imágenes” en el IV Congreso Internacional de Tecnología y Ciencia Aplicada, desarrollado en el Tecnológico Nacional de México/Centro Nacional de Investigación y Desarrollo Tecnológico, TecNM/CENIDET, del 21 al 23 de mayo de 2025.

Las constancias de los productos mencionados se encuentran en la sección Anexo A.

5.4 Trabajo futuro

El desarrollo y validación del sistema actual abre diversas líneas de trabajo orientadas a incrementar su robustez, adaptabilidad y aplicabilidad en entornos reales de monitoreo continuo. Una de las principales proyecciones consiste en la ampliación del conjunto de datos mediante la incorporación de nuevas clases de acciones y la captura de secuencias en condiciones ambientales más diversas, con el fin de mejorar la capacidad de generalización del modelo y su tolerancia a variaciones de iluminación, fondo y perspectiva.

Asimismo, se plantea la optimización del rendimiento en hardware de bajo consumo, explorando técnicas de compresión y cuantización de modelos que permitan su despliegue en dispositivos embebidos o cámaras inteligentes, reduciendo la dependencia de equipos de alto desempeño.

Otra línea de trabajo contempla la integración de mecanismos de adaptación en línea, capaces de ajustar el modelo a patrones individuales de movimiento del usuario sin requerir un reentrenamiento completo, facilitando así la personalización del sistema a las características específicas de cada entorno.

También se prevé la implementación de estrategias multimodales que combinen la información esquelética con otras fuentes de datos no intrusivas, como el análisis de audio ambiental o señales fisiológicas, con el propósito de incrementar la fiabilidad en la detección de eventos críticos.

Finalmente, se considera la mejora de la interfaz de usuario y del sistema de notificaciones, incorporando herramientas de visualización avanzada y alertas adaptativas que prioricen eventos de mayor relevancia para los cuidadores o personal médico.

La consolidación de estas líneas permitirá que el sistema evolucione desde un prototipo funcional hacia una solución integral de monitoreo de acciones humanas, con potencial de aplicación en contextos domésticos, hospitalarios y de teleasistencia.

Referencias

- [1] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimed Tools Appl*, vol. 79, no. 41–42, pp. 30509–30555, Nov. 2020, doi: 10.1007/s11042-020-09004-3.
- [2] M. Feng and J. Meunier, "Skeleton Graph-Neural-Network-Based Human Action Recognition: A Survey," *Sensors*, vol. 22, no. 6, p. 2091, Mar. 2022, doi: 10.3390/s22062091.
- [3] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting Skeleton-based Action Recognition," Apr. 2021, [Online]. Available: <http://arxiv.org/abs/2104.13586>
- [4] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," Mar. 2020, [Online]. Available: <http://arxiv.org/abs/2003.14111>
- [5] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," Jun. 2020, [Online]. Available: <http://arxiv.org/abs/2006.10204>
- [6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 1010–1019. doi: 10.1109/CVPR.2016.115.
- [7] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human Pose Estimation and Its Application to Action Recognition: A Survey A R T I C L E I N F O," 2021. [Online]. Available: www.elsevier.com/locate/jvci
- [8] M. G. Morshed, T. Sultana, A. Alam, and Y. K. Lee, "Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities," Feb. 01, 2023, *MDPI*. doi: 10.3390/s23042182.
- [9] J. Docekal, J. Rozlivek, J. Matas, and M. Hoffmann, "Human Keypoint Detection for Close Proximity Human-Robot Interaction," in *IEEE-RAS International Conference on Humanoid Robots*, IEEE Computer Society, 2022, pp. 450–457. doi: 10.1109/Humanoids53995.2022.10000133.
- [10] L. Huang *et al.*, "Semi-Supervised 2D Human Pose Estimation Driven by Position Inconsistency Pseudo Label Correction Module," *Institute of Electrical and Electronics Engineers (IEEE)*, Aug. 2023, pp. 693–703. doi: 10.1109/cvpr52729.2023.00074.
- [11] C. Wang and J. Yan, "A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition," 2023, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2023.3282311.

- [12] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 1227–1236. doi: 10.1109/CVPR.2019.00132.
- [13] W. Yang, J. Zhang, J. Cai, and Z. Xu, "HybridNet: Integrating GCN and CNN for skeleton-based action recognition," *Applied Intelligence*, vol. 53, no. 1, pp. 574–585, Jan. 2023, doi: 10.1007/s10489-022-03436-0.
- [14] Z. Ren, Q. Zhang, J. Cheng, F. Hao, and X. Gao, "Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition," *Neurocomputing*, vol. 433, pp. 142–153, Apr. 2021, doi: 10.1016/j.neucom.2020.12.020.
- [15] J. Zhang, A. Yang, C. Miao, X. Li, R. Zhang, and D. N. H. Thanh, "3D Graph Convolutional Feature Selection and Dense Pre-Estimation for Skeleton Action Recognition," *IEEE Access*, vol. 12, pp. 11733–11742, 2024, doi: 10.1109/ACCESS.2024.3353622.
- [16] I. A. Bustoni, I. Hidayatulloh, A. M. Ningtyas, A. Purwaningsih, and S. N. Azhari, "Classification methods performance on human activity recognition," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Feb. 2020. doi: 10.1088/1742-6596/1456/1/012027.
- [17] Y. Wang, P. Styne, P. Pathak, and C. Muntean, "A Real-Time Human Action Recognition Model for Assisted Living," *arXiv preprint*, 2025.
- [18] Y. Kong and Y. Fu, "Human Action Recognition and Prediction: A Survey," *Int J Comput Vis*, vol. 130, no. 5, pp. 1366–1401, May 2022, doi: 10.1007/s11263-022-01594-9.
- [19] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph Convolutional Neural Network for Human Action Recognition: A Comprehensive Survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 128–145, Apr. 2021, doi: 10.1109/TAI.2021.3076974.
- [20] Q. Xu, W. Zheng, Y. Song, C. Zhang, X. Yuan, and Y. Li, "Scene image and human skeleton-based dual-stream human action recognition," *Pattern Recognit Lett*, vol. 148, pp. 136–145, Aug. 2021, doi: 10.1016/j.patrec.2021.06.003.
- [21] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, no. 2, pp. 690–712, Feb. 2021, doi: 10.1007/s10489-020-01823-z.
- [22] H.-S. Fang *et al.*, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," Nov. 2022, [Online]. Available: <http://arxiv.org/abs/2211.03375>

- [23] Y. Chen, L. Wang, D. Hu, and H. Cheng, "Multi-view graph convolution network for the recognition of human action with spatial and temporal occlusion problems," *J Vis Commun Image Represent*, vol. 97, Dec. 2023, doi: 10.1016/j.jvcir.2023.103957.
- [24] W. Hu, C. Zhang, F. Zhan, L. Zhang, and T. T. Wong, "Conditional Directed Graph Convolution for 3D Human Pose Estimation," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 602–611. doi: 10.1145/3474085.3475219.
- [25] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," 2019. [Online]. Available: <https://github.com/lshiwjx/2s-AGCN>
- [26] M. S. Alsawadi, E. S. M. El-Kenawy, and M. Rio, "Using BlazePose on Spatial Temporal Graph Convolutional Networks for Action Recognition," *Computers, Materials and Continua*, vol. 74, no. 1, pp. 19–36, 2023, doi: 10.32604/cmc.2023.032499.
- [27] M. Terreran, L. Barcellona, and S. Ghidoni, "A general skeleton-based action and gesture recognition framework for human–robot collaboration," *Rob Auton Syst*, vol. 170, Dec. 2023, doi: 10.1016/j.robot.2023.104523.
- [28] C. Pang, X. Lu, and L. Lyu, "Skeleton-based Action Recognition through Contrasting Two-Stream Spatial-Temporal Networks," Jan. 2023, [Online]. Available: <http://arxiv.org/abs/2301.11495>
- [29] Y. Jiang, Z. Sun, S. Yu, S. Wang, and Y. Song, "A Graph Skeleton Transformer Network for Action Recognition," *Symmetry (Basel)*, vol. 14, no. 8, Aug. 2022, doi: 10.3390/sym14081547.
- [30] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001, doi: 10.1006/cviu.2000.0897.
- [31] H. B. Zhang *et al.*, "A comprehensive survey of vision-based human action recognition methods," Mar. 01, 2019, *MDPI AG*. doi: 10.3390/s19051005.
- [32] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation." [Online]. Available: <https://github.com/leoxiaobin/>
- [33] K. Gedamu, Y. Ji, L. L. Gao, Y. Yang, and H. T. Shen, "Relation-mining self-attention network for skeleton-based human action recognition," *Pattern Recognit*, vol. 139, Jul. 2023, doi: 10.1016/j.patcog.2023.109455.
- [34] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," Dec. 2018, [Online]. Available: <http://arxiv.org/abs/1812.08008>

- [35] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human Action Recognition Systems: A Review of the Trends and State-of-the-Art," *IEEE Access*, vol. 12, pp. 36372–36390, 2024, doi: 10.1109/ACCESS.2024.3373199.
- [36] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," May 31, 2021, *Association for Computing Machinery*. doi: 10.1145/3447744.
- [37] M. Li, K. Chen, Y. Bai, and J. Pei, "Skeleton action recognition via graph convolutional network with self-attention module," *Electronic Research Archive*, vol. 32, no. 4, pp. 2848–2864, 2024, doi: 10.3934/ERA.2024129.
- [38] H. Li et al., "Hierarchical Graph Networks for 3D Human Pose Estimation," Apr. 2023, [Online]. Available: <http://arxiv.org/abs/2111.11927>
- [39] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition," Apr. 2019, [Online]. Available: <http://arxiv.org/abs/1904.01189>
- [40] R. Sun, Z. Lin, S. Leng, A. Wang, and L. Zhao, "An In-Depth Analysis of 2D and 3D Pose Estimation Techniques in Deep Learning: Methodologies and Advances," Apr. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/electronics14071307.
- [41] S. Salisu, K. U. Danyaro, M. Nasser, I. M. Hayder, and H. A. Younis, "Review of models for estimating 3D human pose using deep learning," *PeerJ Comput Sci*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2574.
- [42] A. Dittadi, S. Dziadzio, D. Cosker, B. Lundell, T. Cashman, and J. Shotton, "Full-Body Motion from a Single Head-Mounted Device: Generating SMPL Poses from Partial Observations."
- [43] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting Skeleton-based Action Recognition," 2022. [Online]. Available: <https://github.com/kennymckormick/pyskl>.
- [44] G. Yao, T. Lei, and J. Zhong, "A review of Convolutional-Neural-Network-based action recognition," *Pattern Recognit Lett*, vol. 118, pp. 14–22, Feb. 2019, doi: 10.1016/j.patrec.2018.05.018.
- [45] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Jun. 2014, [Online]. Available: <http://arxiv.org/abs/1406.2199>
- [46] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3.^a ed. Pearson Prentice Hall, 2008.

- [47] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, "Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN," *Sci Afr*, vol. 21, Sep. 2023, doi: 10.1016/j.sciaf.2023.e01796.
- [48] S. Dill *et al.*, "Accuracy Evaluation of 3D Pose Reconstruction Algorithms Through Stereo Camera Information Fusion for Physical Exercises with MediaPipe Pose," *Sensors*, vol. 24, no. 23, Dec. 2024, doi: 10.3390/s24237772.
- [49] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: a survey," *Neural Comput Appl*, vol. 35, no. 5, pp. 4145–4182, Feb. 2023, doi: 10.1007/s00521-022-07937-4.
- [50] Y. Cao and T. Li, "SVM action recognition model based on skeletal key point analysis with posture sensors to help sports training," *BMC Sports Sci Med Rehabil*, vol. 17, no. 1, Dec. 2025, doi: 10.1186/s13102-025-01260-w.
- [51] N. Manouchehri and N. Bouguila, "Human Activity Recognition with an HMM-Based Generative Model," *Sensors*, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031390.
- [52] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," 2015, *Frontiers Media S.A.* doi: 10.3389/frobt.2015.00028.
- [53] M. Nan and A. M. Florea, "Fast Temporal Graph Convolutional Model for Skeleton-Based Action Recognition," *Sensors*, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197117.
- [54] F. Gu, M. H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A Survey on Deep Learning for Human Activity Recognition," Nov. 30, 2022, *Association for Computing Machinery*. doi: 10.1145/3472290.
- [55] S. Singh Sengar, A. Kumar, and O. Singh, "Efficient Human Pose Estimation: Leveraging Advanced Techniques with MediaPipe," 2024. [Online]. Available: https://github.com/avhixd/Human_pose_estimation
- [56] G. V. Urquiza, "Detección de anomalías en el comportamiento de una persona," Tesis de Maestría, Tecnológico Nacional de México – Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Morelos, 2022.
- [57] S. García, "Identificación de situaciones de interés centradas en el humano.," Tesis de Maestría, Tecnológico Nacional de México – Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Morelos, 2024.
- [58] L. P. Arriaga, "Estimación de la posición del cuerpo humano en secuencia de imágenes.," Tesis de Maestría, Tecnológico Nacional de México – Centro

- Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Morelos, 2024.
- [59] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," 2012. [Online]. Available: <http://crcv.ucf.edu/data/UCF101.php>
- [60] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, IEEE, Nov. 2011, pp. 2556–2563. doi: 10.1109/ICCV.2011.6126543.
- [61] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, "A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2093705.
- [62] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," Jun. 10, 2020, *MDPI*. doi: 10.3390/jimaging6060046.
- [63] E. Saund, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Models," in *Proceedings CVPR '89: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Comput. Soc. Press, Jun. 1989, pp. 70–78. [Online]. Available: <https://ieeexplore.ieee.org/document/37830>
- [64] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2013, pp. 3551–3558. doi: 10.1109/ICCV.2013.441.
- [65] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection." [Online]. Available: <http://lear.inrialpes.fr>
- [66] M. H. Siddiqi, M. Alruwaili, A. Ali, S. Alanazi, and F. Zeshan, "Human Activity Recognition Using Gaussian Mixture Hidden Conditional Random Fields," *Comput Intell Neurosci*, vol. 2019, 2019, doi: 10.1155/2019/8590560.
- [67] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Beijing, China: IEEE, Oct. 2005, pp. 1395–1402 Vol. 2. doi: 10.1109/ICCV.2005.28.
- [68] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, British Machine Vision Association, BMVA, 2008. doi: 10.5244/C.22.99.

- [69] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, New York, NY, USA: ACM, Sep. 2007, pp. 357–360. doi: 10.1145/1291233.1291311.
- [70] S. Singh, C. Arora, and C. V. Jawahar, "Trajectory Aligned Features For First Person Action Recognition," Apr. 2016, [Online]. Available: <http://arxiv.org/abs/1604.02115>
- [71] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 698–703. doi: 10.1016/j.procs.2019.08.100.
- [72] H. Le, C. K. Lu, C. C. Hsu, and S. K. Huang, "Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network," *Applied Intelligence*, vol. 55, no. 5, Apr. 2025, doi: 10.1007/s10489-024-06082-w.
- [73] M. Elnady and H. E. Abdelmunim, "A novel YOLO LSTM approach for enhanced human action recognition in video sequences," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01898-z.
- [74] R. Vrskova, P. Kamencay, R. Hudec, and P. Sykora, "A New Deep-Learning Method for Human Activity Recognition," *Sensors*, vol. 23, no. 5, Mar. 2023, doi: 10.3390/s23052816.
- [75] C. I. Orozco, E. Xamena, M. E. Buemi, and J. J. Berlles, "Human Action Recognition in Videos using a Robust CNN LSTM Approach," *Ciencia y Tecnología*, pp. 21–34, Dec. 2020, doi: 10.18682/cyt.vi0.3288.
- [76] M. Joudaki, M. Imani, and H. R. Arabnia, "A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection," *Technologies (Basel)*, vol. 13, no. 2, Feb. 2025, doi: 10.3390/technologies13020053.
- [77] A. Mansouri, T. Bakir, and A. Elzaar, "Improved semantic-guided network for skeleton-based action recognition," *J Vis Commun Image Represent*, vol. 104, Oct. 2024, doi: 10.1016/j.jvcir.2024.104281.
- [78] H. C. Nguyen, T. H. Nguyen, R. Scherer, and V. H. Le, "Deep Learning for Human Activity Recognition on 3D Human Skeleton: Survey and Comparative Study," Jun. 01, 2023, *MDPI*. doi: 10.3390/s23115121.
- [79] A. Ali, E. Pinyoanuntapong, P. Wang, and M. Dorodchi, "Skeleton-based Human Action Recognition via Convolutional Neural Networks (CNN)," Jan. 2023, [Online]. Available: <http://arxiv.org/abs/2301.13360>
- [80] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks," *IEEE*

- Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020, doi: 10.1109/TIP.2020.3028207.
- [81] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 1227–1236. doi: 10.1109/CVPR.2019.00132.
- [82] J. Wang, E. Bergeret, and I. Falih, “Skeleton-Based Action Recognition with Spatial-Structural Graph Convolution,” Jul. 2024, [Online]. Available: <http://arxiv.org/abs/2407.21525>
- [83] J. Do and M. Kim, “SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition,” Jul. 2024, [Online]. Available: <http://arxiv.org/abs/2403.09508>
- [84] L. Wang and P. Koniusz, “3Mformer: Multi-order Multi-mode Transformer for Skeletal Action Recognition,” Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.14474>
- [85] N. H. B. Long, “STEP CATFormer: Spatial-Temporal Effective Body-Part Cross Attention Transformer for Skeleton-based Action Recognition,” Dec. 2023, [Online]. Available: <http://arxiv.org/abs/2312.03288>
- [86] X. Zhu, P.-Y. Huang, J. Liang, C. M. de Melo, and A. Hauptmann, “STMT: A Spatial-Temporal Mesh Transformer for MoCap-Based Action Recognition.” [Online]. Available: <https://github.com/zgzxy001/>
- [87] D. Ahn, S. Kim, H. Hong, and B. C. Ko, “STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition,” Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2210.07503>
- [88] W. Wu, C. Zheng, Z. Yang, C. Chen, S. Das, and A. Lu, “Frequency Guidance Matters: Skeletal Action Recognition by Frequency-Aware Mixed Transformer,” in *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2024, pp. 4660–4669. doi: 10.1145/3664647.3681009.
- [89] Y. Wen, B. Ding, M. Liu, Z. Tang, and Y. Pang, “Interactive Spatiotemporal Token Attention Network for Skeleton-based General Interactive Action Recognition,” 2024, doi: 10.48550/arXiv.2307.07469.
- [90] Q. Wang, J. Peng, S. Shi, T. Liu, J. He, and R. Weng, “IIP-Transformer: Intra-Inter-Part Transformer for Skeleton-Based Action Recognition,” Oct. 2021, doi: 10.1109/BigData59044.2023.10386970.
- [91] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, “Human Action Recognition From Various Data Modalities: A Review,” Mar. 01, 2023, *IEEE Computer Society*. doi: 10.1109/TPAMI.2022.3183112.

- [92] V. T. Le, K. Tran-Trung, and V. T. Hoang, "A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition," 2022, *Hindawi Limited*. doi: 10.1155/2022/8323962.
- [93] L. Wu, C. Zhang, and Y. Zou, "SpatioTemporal focus for skeleton-based action recognition," *Pattern Recognit*, vol. 136, Apr. 2023, doi: 10.1016/j.patcog.2022.109231.
- [94] L. Weng, W. Lou, X. Shen, and F. Gao, "A 3D graph convolutional networks model for 2D skeleton-based human action recognition," *IET Image Process*, vol. 17, no. 3, pp. 773–783, Feb. 2023, doi: 10.1049/ipr2.12671.
- [95] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition," 2019.
- [96] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Part-Level Graph Convolutional Network for Skeleton-Based Action Recognition," 2020. [Online]. Available: www.aaai.org
- [97] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition," 2018.
- [98] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," in *AAAI Conference on Artificial Intelligence*, 2018.
- [99] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," 2018. [Online]. Available: www.aaai.org
- [100] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," Feb. 2017, [Online]. Available: <http://arxiv.org/abs/1609.02907>
- [101] M. Cui, W. Wang, K. Zhang, Z. Sun, and L. Wang, "Pose-Appearance Relational Modeling for Video Action Recognition," *IEEE Transactions on Image Processing*, vol. 32, pp. 295–308, 2023, doi: 10.1109/TIP.2022.3228156.
- [102] C. I. Orozco, M. E. Buemi, and J. J. Berlles, "Towards an attention mechanism LSTM framework for human action recognition in videos," in *2020 IEEE Congreso Bienal de Argentina, ARGENCON 2020 - 2020 IEEE Biennial Congress of Argentina, ARGENCON 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ARGENCON49523.2020.9505440.
- [103] J. Qiu and L. Wang, "Evolving Skeletons: Motion Dynamics in Action Recognition," vol. 22, no. 25, 2024, doi: 10.1145/3701716.3717739.

- [104] A. Snoun, N. Jlidi, T. Bouchrika, O. Jemai, and M. Zaied, "Towards a deep human activity recognition approach based on video to image transformation with skeleton data," *Multimed Tools Appl*, vol. 80, no. 19, pp. 29675–29698, Aug. 2021, doi: 10.1007/s11042-021-11188-1.
- [105] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," Mar. 2019, [Online]. Available: <http://arxiv.org/abs/1903.02428>
- [106] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial Temporal Transformer Network for Skeleton-Based Action Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 694–701. doi: 10.1007/978-3-030-68796-0_50.
- [107] J. Shin *et al.*, "Fall recognition using a three stream spatio temporal GCN model with adaptive feature aggregation," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-95508-7.
- [108] N. Varshney, B. Bakariya, A. K. S. Kushwaha, and M. Khare, "Rule-based multi-view human activity recognition system in real time using skeleton data from RGB-D sensor," *Soft comput*, vol. 27, no. 1, pp. 405–421, Jan. 2023, doi: 10.1007/s00500-021-05649-w.
- [109] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, and Q. Miao, "Transformer for Skeleton-based action recognition: A review of recent advances," *Neurocomputing*, vol. 537, pp. 164–186, Jun. 2023, doi: 10.1016/j.neucom.2023.03.001.
- [110] Y. Wang, P. Stykes, P. Pathak, and C. Muntean, "A Real-Time Human Action Recognition Model for Assisted Living," Mar. 2025, doi: doi.org/10.48550/arXiv.2503.18957.
- [111] C. Shi and S. Liu, "Human action Recognition with Transformer based on Convolutional Features," 2024, doi: 10.21203/rs.3.rs-3855469/v1.
- [112] S. Hurtado, J. García-Nieto, A. Popov, and I. Navas-Delgado, "Human Activity Recognition From Sensorised Patient's Data in Healthcare: A Streaming Deep Learning-Based Approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 23–37, 2023, doi: 10.9781/ijimai.2022.05.004.
- [113] A. W. Muhamada and A. A. Mohammed, "Review on recent Computer Vision Methods for Human Action Recognition," 2021, *University of Salamanca*. doi: 10.14201/ADCAIJ2021104361379.
- [114] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: a narrative review," *Artif Intell Rev*, vol. 55, no. 6, pp. 4755–4808, Aug. 2022, doi: 10.1007/s10462-021-10116-x.

- [115] D. Kim, H. Kim, Y. Mok, and J. Paik, "Real-time surveillance system for analyzing abnormal behavior of pedestrians," *Applied Sciences (Switzerland)*, vol. 11, no. 13, Jul. 2021, doi: 10.3390/app11136153.
- [116] L. Xiao, X. Yang, T. Peng, H. Li, and R. Guo, "Skeleton-Based Activity Recognition for Process-Based Quality Control of Concealed Work via Spatial–Temporal Graph Convolutional Networks," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041220.
- [117] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, "Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model," *Applied Sciences (Switzerland)*, vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042700.
- [118] H. Boudlal, M. Serrhini, and A. Tahiri, "A Monitoring System for Elderly People Using WiFi Sensing with Channel State Information," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 12, pp. 112–131, 2023, doi: 10.3991/ijim.v17i12.36613.
- [119] S. M. Kang and R. P. Wildes, "Review of Action Recognition and Detection Methods," Oct. 2016, [Online]. Available: <http://arxiv.org/abs/1610.06906>
- [120] D. Silva, A. Manzo-Martínez, F. Gaxiola, L. Gonzalez-Gurrola, and G. Ramírez-Alonso, "Analysis of CNN Architectures for Human Action Recognition in Video," *Computacion y Sistemas*, vol. 26, no. 2, pp. 623–641, 2022, doi: 10.13053/CyS-26-2-4245.
- [121] Tomislav. Petković, Davor. Petrinović, and Sven. Loncaric, *ISPA 2021 : 12th International Symposium on Image and Signal Processing and Analysis: online/Zagreb, Croatia, September 13-15, 2021*. University of Zagreb, 2021.
- [122] A. M. Rekavandi, F. Boussaid, A. K. Seghouane, and M. Bennamoun, "B-Pose: Bayesian Deep Network for Camera 6-DoF Pose Estimation from RGB Images," *IEEE Robot Autom Lett*, vol. 8, no. 10, pp. 6747–6754, Oct. 2023, doi: 10.1109/LRA.2023.3313062.
- [123] M. H. Arshad, M. Bilal, and A. Gani, "Human Activity Recognition: Review, Taxonomy and Open Challenges," Sep. 01, 2022, *MDPI*. doi: 10.3390/s22176463.
- [124] Mr. Nagesh UB, Abhishek V Doddagoudra, Adarsh KM, Mayoore K Bhat, and Shreya L, "A Brief Study on Human Action Recognition," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 39–44, Jan. 2023, doi: 10.48175/ijarsct-7832.
- [125] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition," 2023. [Online]. Available: <https://github.com/Jho-Yonsei/HD-GCN>.

- [126] T. Li, Z. Sun, and X. Chen, "Group-Skeleton-Based Human Action Recognition in Complex Events," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 4703–4707. doi: 10.1145/3394171.3416280.
- [127] P. Pareek and A. Thakkar, "A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications," *Artif Intell Rev*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.
- [128] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020, doi: 10.1109/TPAMI.2019.2916873.
- [129] D. K. Ghosh, A. Chakrabarty, H. Moon, and M. J. Piran, "A Spatio-Temporal Graph Convolutional Network Model for Internet of Medical Things (IoMT)," *Sensors*, vol. 22, no. 21, Nov. 2022, doi: 10.3390/s22218438.
- [130] J. Yang, Y. He, J. Zhu, Z. Lv, and W. Jin, "Fall Detection Method for Infrared Videos Based on Spatial-Temporal Graph Convolutional Network," *Sensors*, vol. 24, no. 14, Jul. 2024, doi: 10.3390/s24144647.
- [131] N. Chinchor, "MUC-4 Evaluation Metrics," in *Proceedings of the 4th Message Understanding Conference (MUC-4)*, McLean, VA, USA: Science Applications International Corporation (SAIC), 1992. [Online]. Available: <https://aclanthology.org/M92-1002>
- [132] A. Reinke et al., "Understanding metric-related pitfalls in image analysis validation," Feb. 2024, doi: 10.1038/s41592-023-02150-0.
- [133] D. Martin Ward Powers, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," *Article in Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011, doi: 10.9735/2229-3981.
- [134] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Dec. 2019, [Online]. Available: <http://arxiv.org/abs/1912.01703>
- [135] J. Glover, V. Lazzarini, and J. Timoney, "Python For Audio Signal Processing," 2011. [Online]. Available: <http://docs.scipy.org>,
- [136] P. L. Shopbell et al., "matplotlib-A Portable Python Plotting Package," in *Astronomical Data Analysis Software and Systems XIV (ADASS XIV)*, ASP Conference Series, San Francisco, CA, USA: Astronomical Society of the Pacific, 2005, p. 91. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2005ASPC..347...91B>

- [137] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jan. 2012, [Online]. Available: <http://arxiv.org/abs/1201.0490>

Anexo A



Figura A. 1 Ponencia: Descripción automática de acciones en secuencias de imágenes- Coloquio IA 2024, TecNM Chilpancingo.



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

OTORGA LA PRESENTE

CONSTANCIA

A

Diego Alejandro Reyes Solorio

Por su valiosa participación como Personal de Apoyo en la organización y desarrollo del
3er Congreso Internacional de Tecnología y Ciencia Aplicada (CITCA)
Llevado a cabo del 19 al 22 de noviembre de 2024 en las instalaciones del Tecnológico Nacional de
México campus Centro nacional de Investigación y Desarrollo Tecnológico (TecNM/CENIDET)

Cuernavaca, Morelos, 22 de noviembre de 2024



D080624

<http://constancias.cenidet.tecnm.mx>

LORENZO O. HERNÁNDEZ
DIRECTOR DEL CENTRO NACIONAL DE
INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

ALICIA MARTÍNEZ REBOLLAR
COORDINADORA GENERAL
III CITCA

Sello Digital:

ACfO9FERTQ+nSyhLc1tWK+DIkEDKxBuQ3hdcBbVp9geMuat2TryGX1fXCuRJu110J+zxKJCxa6azmQcTCNSYL1
+E+ATewfnnIjXhpoFkDPQagv5xAuy1pH1zE1oO2/EYcIH9ENq3YypYHMTZ0H88Iz1BcZ4+82bk4901oYbJX3
cQEOwXfjO5MLjYb+beghFdb4RD13Nr6JC1gAaUKvV/BDCLjsZ0GSQ6MwoLob0btEglttbgWFMorbwOjjgyp1Or
Dszo+ndfIMkRr3IAS1b0QycS1VeRyQ4Vx3dJ3dMZ4W1Rj5wTuZAReGs7GkzeMkQbzQqjhr75bhKYGOSzDsfQ==

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico

CITCA
3er. CONGRESO INTERNACIONAL
DE TECNOLOGÍA Y CIENCIA APLICADA



Figura A. 2 Participación como personal de apoyo en la organización y desarrollo del 3er CITCA.



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

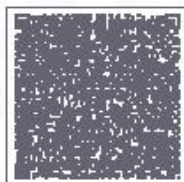
DIEGO ALEJANDRO REYES SOLORIO | RAÚL PINTO ELÍAS | ANDREA
MAGADAN SALAZAR | NIMROD GONZÁLEZ FRANCO

Por su destacada participación con la Contribución: ARTÍCULO PARA REVISTA, en al
área: CIENCIAS DE LA COMPUTACIÓN, con el Título:
DESCRIPCIÓN AUTOMÁTICA DE ACCIONES EN SECUENCIAS DE IMÁGENES
presentada en el marco del 4.º Congreso Internacional de Tecnología y Ciencia
Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y Desarrollo
Tecnológico del 21 al 23 de mayo de 2025.

Cuernavaca, Morelos, 23 de mayo de 2025

LORENZO O. HERNÁNDEZ
Director del CENIDET

MANUEL ADAM MEDINA
Coordinador General
4.º CITCA



Sello Digital:

1PVoUhmT2OyrIsm1YEEFLI4laL0e9XNaqWpnzOPkOSybRIV67/QAN06zPtFcY44XUBRrHwqf
0fM9joFhGKXvVR5jZNVjqcQ0Hgw70ada7E7+kisfIaw+pwu/j0YuFGMSBwq+5dYi9bnPOa
32IB2M87NH8pYr6EkSm1qzh7Fb8tIfgtw14COVCB301vxxwDHy0Wn+gNg+IW6IKzpyJdwSGN+
oE+f1xRa3ekA11yOudo3Y/th52VHy6UsLfw6gXIBRNvxJJrKL4cCVKnM2HYNAU/NVnFpW4
sh8K59Y3kgW1K/jNchsaElPrTbsiqHOFW1mF07zN2Kf8Zc3XsnNktg==

CITCA117A25 <http://constancias.cenidet.tecnm.mx>

Figura A. 3 Artículo presentado en el Congreso CITCA 2025 (presencial).