INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 5, No 2 (2024)
ISSN 2695-6578

Reconocimiento de Gestos para HoloLens2 con Puntos 3D de las Manos Bajo Limitaciones de Datos

Gesture Recognition for HoloLens2 using 3D Hand Points under Data Limitations

Mario Andreu Villar
Instituto Tecnológico de Informática
Valencia, España
mandreu@iti.es

Carlos-D. Martínez-Hinarejos
Universitat Politècnica de València
Valencia, España
cmartine@dsic.upv.es

Patricia Pons
Instituto Tecnológico de Informática
Valencia, España
ppons@iti.es

Jose Luis Soler-Domínguez
Instituto Tecnológico de Informática
Valencia, España
jlsoler@iti.es

Samuel Navas-Medrano
Instituto Tecnológico de Informática
Valencia, España
snavas@iti.es

Vicent Ortiz Castelló
Instituto Tecnológico de Informática
Valencia, España
vortiz@iti.es

Marta García-Ballesteros
Instituto Tecnológico de Informática
Valencia, España
martagarcia@iti.es

Recibido: 04.10.2024 | Aceptado: 17.12.2024

Palabras Clave

Aprendizaje automático
Realidad mixta
Reconocimiento de gestos de la mano
Interacción Persona-Ordenador
Microsoft HoloLens 2
Inteligencia Artificial
Aprendizaje con pocos Ejemplos

Resumen

En la actualidad, la interacción persona-ordenador busca ser cada vez más intuitiva. En el caso de la realidad mixta, el uso de gestos emerge como una solución factible para lograr interacciones más naturales y fluidas. En este trabajo, implementamos un sistema completo de reconocimiento de gestos de las manos para las Microsoft HoloLens 2, basado en un par de clasificadores en cascada. Para el entrenamiento de los modelos, utilizamos la parte disponible públicamente del conjunto de datos SHREC22, que cuenta con un número limitado de muestras, convirtiendo esta tarea en un problema de aprendizaje con pocos ejemplos, ya que solo disponemos de 36 muestras por clase. Exploramos diversas arquitecturas de redes neuronales para identificar la más adecuada en este contexto. Al evaluar el sistema en su conjunto, logramos una tasa de error de gestos (GER) del 9.6%, lo que demuestra el potencial del enfoque propuesto, si bien su rendimiento podría optimizarse con futuros ajustes y más datos de entrenamiento.

Keywords

Machine Learning
Mixed Reality
Hand Gesture Recognition
Human-Computer Interaction
Microsoft HoloLens 2
Artificial Intelligence
Few-shot learning

Abstract

Nowadays, human-computer interaction is seeking to be more intuitive. In the context of mixed reality, hand gestures emerge as a feasible solution to achieve more natural and seamless interactions. In this work, we implement a complete hand gesture recognition system for Microsoft HoloLens 2, based on a pair of cascaded classifiers. For training the models, we use the publicly available part of the SHREC22 dataset, which has a limited number of samples, turning this task into a few-shot learning problem, since only 36 samples per class are available. We explore various neural network architectures to identify the most suitable one in this context. When evaluating the system as a whole, we achieved a gesture error rate (GER) of 9.6%, which highlights the potential of the proposed approach, although its performance could be optimized with future tuning and more training data.

1. Introducción

Desde la aparición de los ordenadores para el público general, la investigación en interacción persona-ordenador (*Human-Computer Interaction*, HCI) ha estado marcada por la búsqueda constante de formas más intuitivas y sencillas de interacción. A lo largo de las décadas, se han explorado diversas tecnologías con el objetivo de acercar la experiencia de uso a algo más natural y cercano a la interacción humana. Desde la introducción de pantallas táctiles, que permiten manipular directamente los objetos en pantalla, hasta los sistemas de reconocimiento de voz, que nos facilitan emitir comandos mediante el habla, la búsqueda de interfaces más accesibles no ha dejado de aumentar. Estas innovaciones reflejan un objetivo común: hacer que el uso de las interfaces sea cada vez más sencillo y natural de forma que nos resulte casi innato usarlas, eliminando barreras entre el usuario y la tecnología (Bannon, 2011).

Desde hace algunos años, un nuevo paradigma ha ganado relevancia en el ámbito de la interacción persona-ordenador: la realidad mixta (*Mixed Reality*, MR) (Milgram & Kishino, 1994). Esta tecnología se diferencia de la realidad virtual (*Virtual Reality*, VR) y de la realidad aumentada (*Augmented Reality*, AR) en su capacidad para integrar elementos virtuales con el entorno físico, en lugar de simplemente sumergir al usuario en un mundo digital (como en VR) o superponer objetos digitales al mundo real (como en AR) (Milgram et al., 1995; Ribeiro et al., 2021). La MR permite que los objetos virtuales interactúen dinámicamente con el entorno físico, ofreciendo una experiencia mucho más inmersiva y coherente. Para comprender mejor las diferencias entre estas tecnologías dentro del continuo realidad-virtualidad, véase la Figura 1. De esta manera, la MR habilita nuevas posibilidades para aplicaciones educativas, industriales, de entretenimiento y más, y ofrece un potencial significativo para crear interacciones más naturales, dinámicas e inmersivas, donde el usuario puede manipular objetos virtuales como si estuvieran presentes en el mundo real, ampliando así las capacidades de las interfaces tradicionales (Dipesh Gyawali, 2023).

Sin embargo, en el mundo de la MR persiste un desafío fundamental: en muchos casos, las interfaces utilizadas para ejecutar acciones replican las mismas que encontramos en las pantallas táctiles, como las de los dispositivos móviles. Este enfoque, aunque funcional, carece de la naturalidad que se busca en un entorno inmersivo (Asadi & Hemadi, 2024; Rokhsaritalemi et al., 2020). Interactuar con botones o menús flotantes en el aire no refleja las acciones que realizamos en la vida diaria. Por ejemplo, en una carrera en el mundo real, el inicio se señala con el descenso de una bandera, un gesto claro y reconocible, no con un contador digital o un botón que deba ser presionado en el aire. Este contraste pone en evidencia que, aunque la MR ha avanzado considerablemente, aún queda mucho trabajo por hacer para que las interfaces sean verdaderamente intuitivas y se alineen con los comportamientos naturales del ser humano.

Entiéndase por interacción natural la capacidad de los usuarios para interactuar con el entorno virtual de manera fluida y coherente, utilizando gestos y comportamientos que imitan los de la vida cotidiana. Por otro lado, interacción intuitiva hace referencia a la facilidad con la que los usuarios comprenden y utilizan el sistema sin necesidad de instrucciones complejas, basándose en su experiencia previa y en el uso común de gestos y movimientos. Estos conceptos son clave para lograr una experiencia inmersiva auténtica en MR (Spittle et al., 2022).

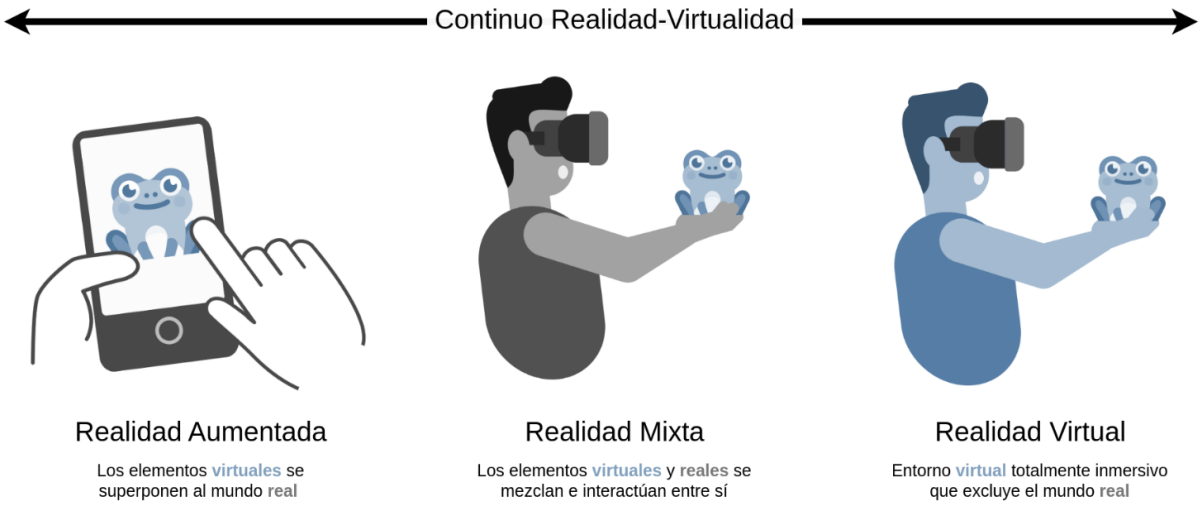

Figura 1: Diferencias entre las tecnologías AR, MR y VR dentro del continuo realidad-virtualidad. Adaptado de Soler-Dominguez et al., 2024

Entre las soluciones propuestas para lograr una interacción más natural e intuitiva en la MR, el uso de gestos destaca como una alternativa prometedora (Guo et al., 2021). La detección de las manos ha experimentado avances significativos gracias a los desarrollos en inteligencia artificial (*Artificial Intelligence*, AI) y aprendizaje automático (*Machine Learning*, ML) (Khan & Ibraheem, 2012; Yang et al., 2019; Mehran Maghoumi & LaViola, 2019), así como a la incorporación de sensores avanzados en los dispositivos de MR (Balakrishnan & Guo, 2024). Estos avances han permitido un seguimiento más preciso y en tiempo real de los movimientos de las manos, superando muchas de las limitaciones previas en la captura de gestos. Con este primer paso ya consolidado, el siguiente desafío es el diseño e implementación de sistemas de reconocimiento de gestos de las manos (*Hand Gesture Recognition*, HGR) que sean tanto rápidos como efectivos (Emporio et al., 2022; Yang et al., 2019). La creación de estos sistemas implica no solo la detección precisa de cuándo se ha realizado un gesto, sino también la capacidad de interpretarlo y responder de manera eficiente dentro del entorno virtual, garantizando una experiencia de usuario fluida y natural.

En este trabajo, nos centramos en el estudio de las aproximaciones actuales al HGR en el contexto de MR. Exploramos y proponemos mejoras de los algoritmos de ML más recientes que facilitan la clasificación de gestos, tanto estáticos como dinámicos. Además, proponemos un sistema integral de detección de gestos específicamente diseñado para su implementación en uno de los dispositivos más utilizados en el mercado, las Microsoft HoloLens 2[1]. El resto del trabajo se estructura de la siguiente manera: en la sección 2, revisamos el estado del arte en HGR y las técnicas de ML aplicadas. La sección 3 detalla el sistema de HGR propuesto. En la sección 4, presentamos el conjunto de datos a utilizar. La experimentación realizada se presenta en la sección 5. Finalmente, en las secciones 6 y 7, discutimos los resultados y las conclusiones, así como las posibles direcciones futuras para la mejora y expansión del sistema.

2. Estado del arte

La IA generativa (GenAI) puede crear diversos tipos de contenido como textos, imágenes, audios y videos, utilizando herramientas como ChatGPT, Copilot, Gemini o Midjourney. Asimismo, en los últimos años han surgido diversos modelos de lenguaje grande (LLM, Large Language Models) sobre los que se pueden desarrollar nuevas aplicaciones que acerquen la IA generativa a los usuarios para resolver problemas de diversa índole. En el contexto educativo, su implementación permite transformar no solo cómo se crea y personaliza el contenido para el aprendizaje, sino también los procesos de enseñanza-

El HGR es un área de investigación prominente dentro del campo de visión por ordenador (*Computer Vision*, CV) y la HCI (Khan & Ibraheem, 2012). Consiste en la interpretación y comprensión automatizadas de los movimientos y configuraciones de las manos para inferir gestos y órdenes del usuario. El HGR ha cobrado gran importancia debido a su potencial para revolucionar diversas aplicaciones (Benedict et al., 2019), como la MR y VR (Emporio et al., 2021), el reconocimiento del lenguaje de signos (Parcheta & Martínez-Hinarejos, 2017) y la interacción persona-robot (Guo et al., 2021).

2.1 Enfoques para el HGR

El HGR se ha abordado históricamente de dos formas principales. La primera es mediante técnicas de CV, donde se emplean imágenes o secuencias de vídeo capturadas por cámaras para identificar gestos a partir de la apariencia visual de la mano. En este enfoque, los algoritmos analizan características como contornos, texturas y movimientos dentro del campo visual. Aunque esta aproximación ha sido ampliamente explorada y puede funcionar eficazmente para ciertos gestos (Köpüklü et al., 2019), puede encontrar dificultades con la segmentación del fondo y el primer plano, las variaciones de iluminación y la limitada información de profundidad (Chakraborty et al., 2018). La segunda técnica se basa en el análisis de coordenadas espaciales, observando puntos clave de la mano, conocidos como "*joints*" o articulaciones, obtenidos por sensores especializados que rastrean su posición en un espacio tridimensional (Emporio et al., 2022). Este método proporciona información más precisa sobre la posición, orientación y profundidad de la mano, facilitando la interpretación de gestos complejos (Guo et al., 2021).

El enfoque basado en CV presenta la ventaja de requerir hardware relativamente sencillo para la captura de datos, ya que cualquier cámara estándar es suficiente para obtener imágenes o vídeo de la mano. En contraste, el segundo enfoque, que se basa en el análisis de coordenadas espaciales, necesita hardware más complejo y costoso, como sensores de profundidad o sistemas de seguimiento específicos (Guo et al., 2021). En nuestro caso, dado que este proyecto se centra en el desarrollo de un sistema de HGR para las Microsoft HoloLens 2, que ya están equipadas con sensores avanzados para el seguimiento de manos en 3D, hemos optado por utilizar esta segunda aproximación.

2.2 HGR analizando coordenadas 3D

Las primeras aproximaciones al HGR mediante ML analizando coordenadas 3D se basaron en métodos tradicionales como las máquinas de vectores de soporte (*Support Vector Machines*, SVM), *random forests* y clasificadores basados en disimilitud (Marin et al., 2015; Caputo et al., 2020). Estos enfoques mostraron resultados iniciales prometedores, pero con la evolución del campo, las redes neuronales se han impuesto

[1] https://www.microsoft.com/es-es/hololens

como la técnica predominante hoy en día. Por un lado, las redes neuronales recurrentes (*Recurrent Neural Networks*, RNN), y específicamente las *Long Short-Term Memory* (LSTM), han sido ampliamente utilizadas para procesar datos secuenciales, como la variación a lo largo del tiempo de las posiciones de las manos (Avola et al., 2019). Además, utilidades como *Deep Gesture Recognition*, que emplea *Gated Recurrent Units* (GRU) apiladas junto con un modelo de atención global, han demostrado ser eficientes y efectivas en la clasificación de gestos (Mehran Maghoumi & LaViola, 2019).

En otros enfoques, se ha observado que redes neuronales convolucionales unidimensionales (1D-CNN) con módulos de resumen de movimientos pueden lograr resultados competitivos con una menor complejidad computacional (Yang et al., 2019). En este contexto, destaca la arquitectura STRONGER (*Simple Trajectory-based Online Gesture Recognizer*) (Emporio et al., 2021), basada en una versión modificada de la arquitectura DDNet (*Double-feature Double-motion Network*) (Yang et al., 2019). DDNet es una red neuronal profunda diseñada específicamente para la clasificación de señales direccionales. La red procesa y combina las señales de entrada en varias direcciones utilizando filtros direccionales, seguidos de capas convolucionales y de *pooling* que reducen la dimensionalidad de las características aprendidas. La ventaja de la DDNet es su capacidad para capturar patrones direccionales específicos que resultan difíciles de identificar para las redes neuronales convencionales. Dado que los gestos se pueden ver como datos direccionales que evolucionan en el tiempo, tanto la DDNet como la STRONGER son adecuadas para su clasificación.

2.3 Conjuntos de datos disponibles

La disponibilidad de *datasets* que incluyan coordenadas tridimensionales de las articulaciones de la mano para el HGR es relativamente escasa. Hasta donde tenemos conocimiento, únicamente hemos identificado tres conjuntos de datos que cumplen este requisito. Por otro lado, vale la pena mencionar que no hemos encontrado modelos neuronales preentrenados diseñados específicamente para esta tarea. Un desafío significativo en la utilización de estos conjuntos de datos es la falta de interoperabilidad entre ellos, ya que cada estudio utiliza diferentes dispositivos de captura y, por ende, cada uno detecta distintos puntos clave de la mano. Esta situación complica la generalización de los resultados y la comparación entre los diferentes estudios. A continuación, presentamos las características de los tres *datasets* principales.

2.3.1 SHREC22[2]

El conjunto de datos SHREC22 incluye 288 secuencias de gestos de la mano, cada una de las cuales contiene un número variable de gestos que oscila entre 3 y 5 por secuencia. Este *dataset* abarca un total de 16 clases diferentes de gestos, con una distribución equilibrada de muestras entre las clases. El conjunto de datos se divide en un conjunto de entrenamiento (144 muestras), con anotaciones, y un conjunto de prueba (144 muestras), que carece de anotaciones debido a su uso en una competición. Los datos se grabaron con las Microsoft HoloLens 2, capturando 26 puntos de interés en las articulaciones de la mano (Emporio et al., 2022).

2.3.2 Dynamic Hand Gesture 14/28[3]

Este conjunto de datos consta de 1.400 secuencias con 14 clases de gestos realizados de dos formas (2.800 en total): con un dedo y con toda la mano. Cada gesto es ejecutado cinco veces por 20 participantes diestros. Las secuencias incluyen imágenes de profundidad y 22 coordenadas de articulaciones tanto en el espacio 2D de la imagen de profundidad como en el espacio 3D del mundo, formando un esqueleto completo de la mano. Los datos se capturan con una cámara de profundidad de corto alcance Intel RealSense, a 30 fotogramas por segundo, con una resolución de 640x480 para las imágenes de profundidad. La duración de los gestos oscila entre 20 y 50 fotogramas y se capturan 22 puntos de interés en las articulaciones de la mano (De Smedt et al., 2016).

2.3.3 SHREC21[4]

La colección de datos SHREC21 comprende 180 secuencias de gestos de la mano, cuidadosamente planificadas para incluir de 3 a 5 gestos por secuencia, complementadas con movimientos de la mano semialeatorios etiquetados como no gestos. El diccionario original contiene 18 gestos, clasificados en gestos estáticos, caracterizados por una postura fija de la mano, y gestos dinámicos, caracterizados por trayectorias de la mano y las articulaciones. Sin embargo, posteriormente se eliminó uno de los gestos debido a posibles conflictos, dejando un total de 17 clases de gestos y manteniendo una distribución equilibrada de muestras entre todas las clases. El conjunto de datos se divide en dos partes: un conjunto de entrenamiento, que incluye 108 secuencias con aproximadamente 24 gestos por clase, y un conjunto de test (anotado), que incluye 72 secuencias con aproximadamente 16 gestos por clase. Las trayectorias de los gestos se tomaron con sensores LeapMotion a 50 FPS, capturando 20 puntos de interés con coordenadas de posición y cuaterniones (Caputo et al., 2021).

[2] https://univr-vips.github.io/Shrec22/#dataset

[3] http://www-rech.telecom-lille.fr/DHGdataset

[4] https://univr-vips.github.io/Shrec21/#revision

2.4 Desafíos y oportunidades

En la actualidad, la mayoría de los avances en HGR se han concentrado en enfoques basados en CV, mientras que la aproximación que utiliza coordenadas 3D de las articulaciones de la mano ha sido menos estudiada, a pesar de sus ventajas, como una representación compacta y un procesado potencialmente eficiente (Yang et al., 2019). Las redes neuronales utilizadas para la detección de gestos en este contexto aún requieren mayor investigación. Aunque se han logrado resultados prometedores, existe un amplio margen de mejora (Emporio et al., 2022), particularmente si se busca una detección en tiempo real. De nada sirve una detección extremadamente precisa si esta introduce demoras de varios segundos, lo que afectaría la experiencia del usuario. Por tanto, es fundamental seguir investigando nuevas arquitecturas de redes y optimizaciones en las actuales para mejorar tanto la precisión como los tiempos de reconocimiento (Emporio et al., 2022).

Otro desafío significativo es la escasez de *datasets* disponibles para el entrenamiento de modelos basados en coordenadas 3D. Como vemos en la sección 2.3, los conjuntos de datos existentes son limitados en número y tamaño, lo que restringe el desarrollo y evaluación de algoritmos más robustos. En nuestro caso particular, trabajando con las Microsoft HoloLens 2, de los tres conjuntos de datos mencionados, solo podemos utilizar el SHREC22, que presenta únicamente 144 secuencias utilizables debido a la falta de anotaciones del conjunto de prueba. Esto pone de manifiesto la necesidad urgente de desarrollar nuevos conjuntos de datos más amplios y específicos que permitan la creación de modelos más generalizables y eficientes.

3. Sistema propuesto

Como discutimos en la sección 2.1, nuestra estrategia de HGR se basa en el uso de información posicional tridimensional de las articulaciones de la mano capturada por el dispositivo Microsoft HoloLens 2. Nuestro enfoque se centra en el diseño de un sistema robusto de clasificación de gestos: dada una ventana temporal compuesta por múltiples *frames* (donde cada *frame* representa la posición tridimensional de los 26 *joints* de la mano en un instante de tiempo), nuestro clasificador principal está diseñado para identificar el gesto específico dentro de dicha ventana.

Si bien, como mencionamos en la sección 2.4, utilizamos el conjunto de datos SHREC22 por ser el único diseñado específicamente para las HoloLens 2, nuestro enfoque difiere ligeramente del planteamiento de la competición asociada al *dataset*. En dicha competición, el objetivo es simular un reconocimiento *online* de manera *offline*: se proporcionan secuencias de gestos mezcladas con ruido y se espera que el sistema identifique los gestos presentes en la secuencia e indique sus posiciones de inicio y fin. Nuestra aproximación, sin embargo, no se centra en localizar directamente las posiciones exactas de los gestos dentro de una secuencia. En su lugar, priorizamos el diseño de un sistema de reconocimiento que, dado un único gesto, sea capaz de clasificarlo correctamente. Para procesar secuencias completas, aplicaremos técnicas de ventana deslizante que nos permitirán identificar las regiones donde ocurre un gesto.

De este modo, nos centramos en el reconocimiento de gestos de forma aislada. Aunque ignorar el contexto puede reducir la precisión en algunos casos, esta simplificación presenta una ventaja clave: permite emplear redes neuronales más sencillas, lo que se traduce en un procesamiento más rápido y optimizado. Para garantizar la eficiencia y el análisis de gestos en tiempo real, proponemos un sistema basado en dos clasificadores en cascada. Además del clasificador principal, empleamos un clasificador binario preliminar, también conocido como detector, cuyo objetivo es determinar de manera rápida la presencia o ausencia de un gesto en una ventana temporal. Esto permite un mecanismo de filtrado eficiente, en el que el análisis más complejo se prioriza únicamente cuando se detecta un gesto, optimizando así la capacidad de respuesta global del sistema. En conjunto, el sistema propuesto sigue un proceso lógico similar al ilustrado en la Figura 2.

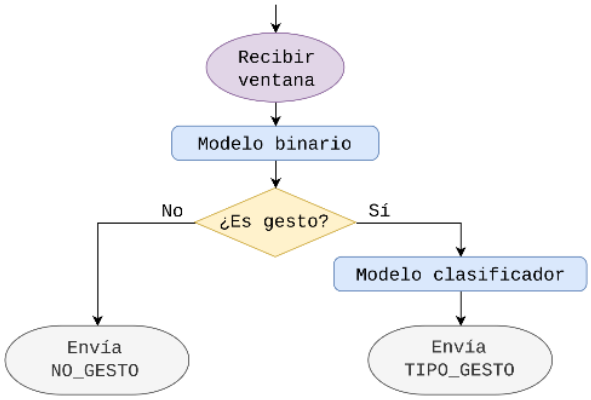

Figura 2: Proceso de toma de decisiones del sistema completo de reconocimiento de gestos

Respecto al despliegue del sistema, decidimos no realizar la inferencia directamente en el dispositivo. Aunque esta opción es técnicamente viable, presenta ciertos desafíos (Zaccardi et al., 2023). La principal razón de esta decisión radica en la preocupación por una posible disminución del rendimiento, ya que la saturación del dispositivo podría lastrar la fluidez de la interfaz MR, comprometiendo la experiencia del usuario. En su lugar, configuramos un servidor externo encargado de realizar las inferencias. Esta configuración reduce significativamente la carga computacional sobre las HoloLens 2, que únicamente se encargan de enviar la información de las articulaciones de la mano y recibir la respuesta con la etiqueta de la clase correspondiente.

En la práctica, la interacción se desarrolla de la siguiente manera: las HoloLens 2 acumulan *frames* con la información de los *joints* y, una vez completada una ventana temporal, envían estos datos al servidor, que realiza la inferencia jerárquica descrita en la Figura 2. La etiqueta resultante es enviada de vuelta a la aplicación, lo que permite respuestas en tiempo real, como la transición de escenas o la activación de elementos interactivos. Para facilitar esta comunicación, utilizamos los protocolos ZeroMQ (ZMQ) sobre paquetes TCP, diseñados específicamente para interacciones de baja latencia. La Figura 3 ilustra este concepto de interacción.

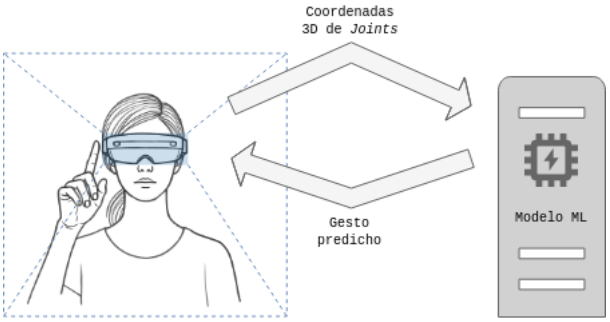

Figura 3: Esquema de comunicación del sistema completo de reconocimiento de gestos

Con esta estructura, la aplicación de MR puede aprovechar las capacidades del sistema combinado de reconocimiento de gestos sin comprometer la experiencia inmersiva. Además, el diseño modular permite realizar ajustes y ampliaciones fácilmente para adaptarse a diversos escenarios.

4. Descripción del conjunto de datos

En este trabajo, utilizamos el conjunto de datos SHREC22, ya que es el único disponible que ha sido específicamente diseñado para clasificar gestos de las manos utilizando las Microsoft HoloLens 2. Como mencionamos en la sección 2.3.1, este *dataset* incluye un total de 288 secuencias de gestos, cada una de ellas con un número variable de gestos (entre 3 y 5). Los datos se dividen en un conjunto de entrenamiento etiquetado con 144 secuencias y un conjunto de prueba con 144 secuencias sin anotaciones. Dado que no disponemos de las etiquetas del conjunto de prueba, utilizamos únicamente el conjunto de entrenamiento.

Los "*joints*", también conocidos como puntos de articulación, son puntos de la mano que representan posiciones importantes para el seguimiento del movimiento. Cada *joint* se caracteriza por tres valores que indican su posición en el espacio tridimensional (x, y, z). En el caso de las Microsoft HoloLens 2, estas coordenadas se expresan en un sistema de coordenadas cartesianas, donde el origen corresponde a la posición inicial de la cabeza del usuario al iniciar la aplicación de realidad mixta.

La Figura 4 muestra los diferentes puntos de articulación que se capturan en la mano.

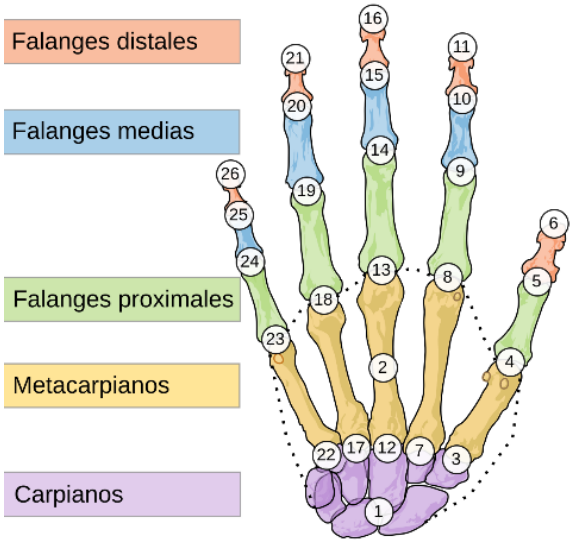

Figura 4: Visualización de las articulaciones capturadas por las HoloLens 2 en una representación esquelética de la mano

En la Figura 5 podemos ver el conjunto de gestos a reconocer en SHREC22. Está compuesto por 16 gestos, divididos en 4 categorías: gestos estáticos, caracterizados por una pose fija mantenida; gestos dinámicos, definidos por una trayectoria única de la mano; gestos dinámicos de alta precisión, caracterizados por movimientos detallados de los dedos; y gestos dinámico-periódicos, en los que un mismo patrón de movimiento de los dedos se repite varias veces. La Tabla 1 proporciona una descripción detallada de cada gesto.

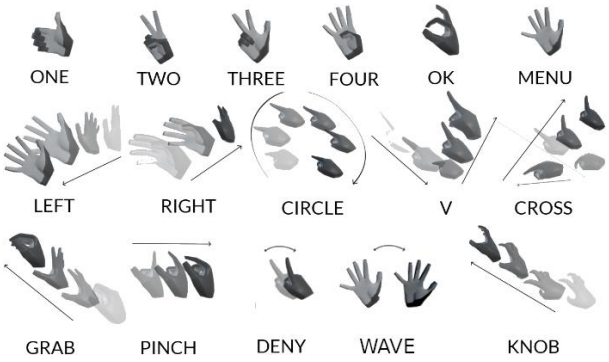

Figura 5: Gestos a reconocer del conjunto de datos SHREC22 (Emporio et al., 2022)

Tabla 2: Descripción de los gestos del dataset SHREC22

Gesto	ID	Descripción
ONE	1	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Independiente de la orientación.
TWO	2	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Independiente de la orientación.
THREE	3	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Independiente de la orientación.
FOUR	4	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Independiente de la orientación.
OK	5	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Independiente de la orientación.
MENU	6	El usuario mantiene las manos en la pose visible en la Figura 5 durante al menos 1 segundo. Dependiente de la orientación.
LEFT	7	El usuario mueve las manos de derecha a izquierda con los dedos abiertos, la mano plana y el pulgar hacia arriba.
RIGHT	8	El usuario mueve las manos de izquierda a derecha con los dedos abiertos, la mano plana y el pulgar hacia arriba.
CIRCLE	9	El usuario mantiene el índice apuntando y dibuja un círculo en el sentido de las agujas del reloj.
V	10	El usuario mantiene el índice apuntando y dibuja una forma de V de izquierda a derecha.
CROSS	11	El usuario mantiene el índice apuntando y dibuja una X empezando desde arriba a la izquierda.
GRAB	12	El usuario mueve la mano hacia delante y realiza un gesto de agarre cerrando todos los dedos hacia la palma, como si estuviera sujetando un objeto imaginario.
PINCH	13	El usuario mueve la mano hacia delante y realiza un gesto de pellizco, acercando el pulgar al índice, como si estuviera agarrando un objeto pequeño entre ambos dedos.
DENY	14	El usuario mantiene el índice apuntando hacia arriba y lo gira periódicamente alrededor de la muñeca en el clásico gesto de negar (2-4 veces).
WAVE	15	El usuario mantiene la mano abierta paralela a la cámara y apuntando hacia arriba y la gira periódicamente alrededor de la muñeca en el clásico gesto de saludo (2-4 veces).
KNOB	16	El usuario simula el agarre y giro de un pomo.

Las secuencias de entrenamiento se distribuyen junto con anotaciones que indican el inicio y el final de cada ejecución de gesto. Como discutimos en la sección 3, para construir nuestro *dataset*, en lugar de utilizar directamente las secuencias completas proporcionadas, hemos extraído los gestos individuales a partir de estas anotaciones, eliminando el contexto circundante. En total, hemos extraído 576 gestos (36 gestos de cada clase) a partir de las 144 secuencias.

Dado que el conjunto de entrenamiento presenta una limitación significativa en cuanto a la cantidad de datos disponibles, abordamos esta tarea como un problema de aprendizaje con pocos ejemplos (*few-shot learning*). Para evaluar el rendimiento del modelo, implementamos un esquema de validación cruzada con 5 iteraciones (*5-fold cross-validation*) sobre el conjunto de entrenamiento. Para complementar la evaluación, hemos grabado un conjunto de test propio que sigue la misma estructura que el *dataset* SHREC22, asegurando la consistencia en el tipo de captura. Estas muestras fueron registradas utilizando el mismo dispositivo y los mismos puntos de articulación que en el conjunto original. En total, hemos grabado 16 secuencias que contienen entre 3 y 5 gestos, resultando en un total de 65 gestos distintos. Estos gestos han sido aislados de manera similar a como se realizó en el conjunto original. La tabla 2 presenta las estadísticas de ambos conjuntos de datos.

Es importante señalar que hemos observado una diferencia significativa en la duración de los gestos grabados. Aunque la frecuencia de adquisición en SHREC22 se reporta como "relativamente baja y no perfectamente estable (aproximadamente 20 Hz)" (Emporio et al., 2022), mientras que nosotros grabamos a 30 Hz, que, hasta donde sabemos, es el estándar para las HoloLens 2 en el seguimiento de manos, esta diferencia en la frecuencia de muestreo no parece ser suficiente para explicar la discrepancia en la duración de los gestos.

Tabla 3: Resumen estadístico de la duración de las muestras (en frames) para los conjuntos de datos

Estadístico	Conjunto de entrenamiento	Conjunto de test (antes del submuestreo)
Cantidad de muestras	576	65
Media	37.6	176.0
Desviación estándar	12.7	56.8
Valor mínimo	12	110
Primer cuartil (25%)	30	136
Mediana (50%)	36	163
Tercer cuartil (75%)	44	194
Valor máximo	82	416

Una posible explicación es que los gestos en SHREC22 hayan sido estrictamente acotados durante las anotaciones, ya que los autores mencionan que las muestras fueron anotadas en posprocesamiento, mientras que, en nuestro caso, la marca de inicio y fin de los gestos se realizó simultáneamente con la grabación mediante un comando de voz, lo que pudo haber dejado algo de margen de error. Sin embargo, no creemos que esta sea la causa principal, ya que una anotación más precisa por sí sola no debería generar una variación tan grande. Otra posible explicación es que la velocidad de ejecución de los gestos pudo haber sido diferente: en nuestro caso, los usuarios pueden haber realizado los gestos más despacio, simulando un uso cotidiano más calmado, mientras que los gestos de SHREC22 podrían haberse realizado a una velocidad mayor.

Para adaptar la duración de nuestros gestos a la del conjunto SHREC22, aplicamos un submuestreo, en el que se realizan saltos de tamaño N, conservando 1 de cada N *frames*. Dado que la mediana de duración de nuestros gestos es de 163 *frames* y la de los gestos del conjunto SHREC22 es de 36 *frames*, calculamos $\mathrm{N} = \left\lceil \frac{163}{36} \right\rceil = 5$.

En ambos conjuntos, se ha realizado finalmente una normalización de las ventanas para adaptarlas a un tamaño de ventana estándar de 36 *frames*, que coincide con la mediana del *dataset*. En el caso de que una ventana de gesto fuese mayor a 36 *frames*, se descartaron los *frames* de los extremos, manteniendo los centrales. Si la ventana de gesto era menor a 36 *frames*, se aplicó un *padding* en los extremos para completar el tamaño necesario.

5. Experimentación

Para la experimentación, seguimos el enfoque planteado en la sección 3, que propone el uso de dos clasificadores: un clasificador binario rápido, encargado de determinar la presencia de un gesto, y un clasificador principal que identifica qué gesto específico se está realizando. A continuación, presentamos las diversas arquitecturas de red estudiadas.

Un componente clave en este estudio es el bloque de transformaciones (TRANSF). Definimos el bloque de transformaciones como el conjunto de ramas extractoras de características por las que pasa la entrada antes del bloque de concatenación en la arquitectura STRONGER o en las DDNet. Este bloque inicial realiza, en el caso de STRONGER, cinco transformaciones sobre los datos: JCD (*Joint Collection of Distances*), JPD (*Joint Pairs' Directions*), PO (*Palm Orientation*), Mfast y Mslow (movimiento a escala rápida y lenta). Estas transformaciones están diseñadas para extraer patrones relevantes en la información posicional y temporal de las articulaciones, realizando una especie de *embedding* del gesto a reconocer.

Para construir el clasificador principal, hemos diseñado e implementado cinco arquitecturas diferentes. Como *baseline*, utilizamos un perceptrón multicapa (MLP) (Murtagh, 1991). Las otras cuatro arquitecturas están basadas en 1D-CNNs que han demostrado ser efectivas en el estado del arte para tareas similares (Yang et al., 2019). Estas arquitecturas incluyen una versión de RESNET (He et al., 2016) adaptada a señales 1D, la red STRONGER (Emporio et al., 2021) con su bloque de transformaciones, y dos versiones híbridas que combinan el bloque de transformaciones con un MLP y con RESNET, respectivamente.

Un aspecto importante que buscamos explorar es si el bloque de transformaciones de la STRONGER puede integrarse satisfactoriamente en otras arquitecturas, como MLP o RESNET. De esta forma, evaluamos si este bloque puede mejorar el rendimiento de modelos tradicionales, aprovechando la generación del "resumen del gesto" que realiza. Las configuraciones TRANSF+MLP y TRANSF+RESNET nos permiten estudiar esta hipótesis y su impacto en la clasificación de gestos.

En términos de metodología, todos los clasificadores toman como entrada un tensor de dimensiones 36x78, donde 36 corresponde a la duración mediana de los gestos en el conjunto de datos, y 78 corresponde a las 26 articulaciones de la mano, cada una representada por sus tres coordenadas espaciales (x, y, z). Los experimentos se han realizado en una máquina con un procesador Intel i7-7700K a 4.199GHz, una tarjeta gráfica NVIDIA RTX 3060 de 12GB de memoria, 16GB de RAM y Ubuntu 22.04.

5.1 Resultados experimentales

A continuación, presentamos los resultados en cada conjunto de evaluación. Véase que tanto en la Figura 6 como en la Figura 7, cada violín representa la distribución de resultados de 25 ejecuciones independientes. En el caso concreto de la Figura 6 (validación cruzada), cada violín refleja los resultados obtenidos para cada uno de los 5 *folds* en cada una de las 25 ejecuciones independientes. Las rayas dentro de cada violín corresponden a los cuartiles de la distribución.

En la Figura 6 se muestran los resultados obtenidos en el conjunto de datos mediante validación cruzada con 5 *folds*. El clasificador MLP, que actúa como línea base, presenta la mayor variabilidad, con precisiones que oscilan entre el 70% y el 95%. Por su parte, RESNET alcanza una mediana cercana al 95%, aunque con cierta dispersión. Las arquitecturas con transformaciones STRONGER y TRANSF+MLP muestran precisiones que varían entre el 95% y el 100%. Finalmente, el modelo TRANSF+RESNET se aproxima al 100% de precisión en todas las ejecuciones, demostrando un rendimiento notablemente superior.

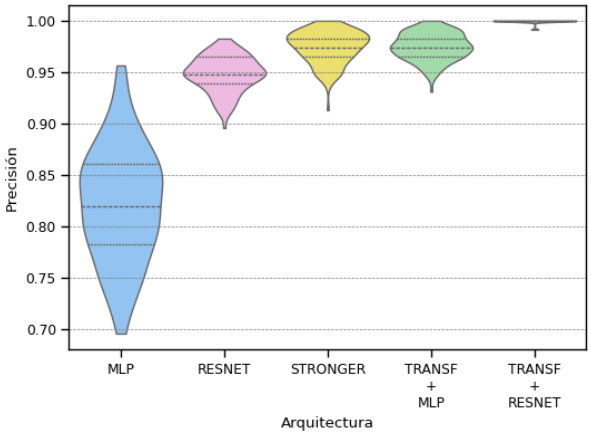

Figura 6: Resultados en el conjunto de datos SHREC22 con validación cruzada

Por otra parte, la Figura 7 muestra el rendimiento en el conjunto de test que grabamos. Se puede observar una tendencia similar a la de la figura anterior, con la única diferencia de que RESNET presenta un rendimiento inferior al de MLP. MLP logra una precisión mediana cercana al 48%, mientras que RESNET apenas alcanza el 34%. En cuanto a las arquitecturas que incluyen transformaciones, STRONGER logra una precisión mediana del 72%, TRANSF+MLP obtiene un 69%, y TRANSF+RESNET alcanza aproximadamente el 85%, siendo esta última la opción con los mejores resultados.

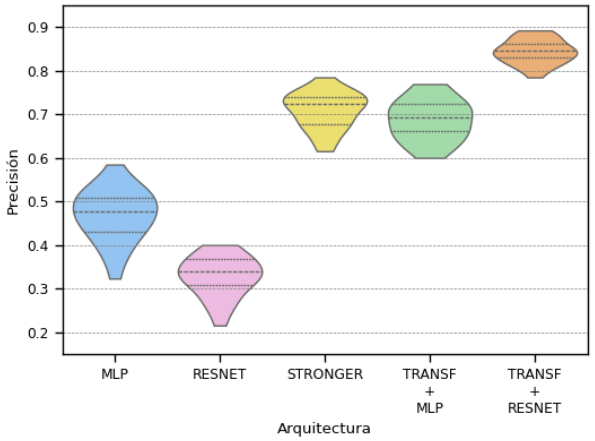

Figura 7: Resultados en el conjunto de test

5.2 Evaluación del sistema completo

Para determinar la presencia de gestos necesitamos crear un modelo binario. Para ello, hemos desarrollado un nuevo conjunto de datos con dos clases de muestras: gestos y no-gestos. Para ello, hemos extraído los espacios entre gestos de las primeras 120 secuencias del conjunto de entrenamiento de SHREC22, donde, según los propios autores, no debería haber gestos (Emporio et al., 2022). Se ha dejado un margen de 55 *frames* entre gestos, lo que nos ha permitido obtener 460 muestras de no-gesto (ventanas de tamaño 36x78). Paralelamente, hemos seleccionado los primeros 460 gestos de esas mismas 120 secuencias, etiquetándolos como gestos.

Con este conjunto de datos, hemos entrenado una red neuronal de arquitectura sencilla, un MLP, que en el mejor caso ha alcanzado una precisión del 91.3% en un conjunto de test diseñado de manera similar, utilizando las 24 secuencias restantes. Además, hemos entrenado el clasificador más efectivo del apartado anterior, TRANSF+RESNET, empleando únicamente los gestos extraídos de las primeras 120 secuencias. El sistema completo se ha estructurado de acuerdo con el diagrama presentado en la Figura 2, incorporando además un umbral al clasificador de gestos: si el detector indica que existe un gesto en la ventana, pero la confianza del clasificador principal en la inferencia es inferior a 0.95, se considerará que no hay gesto.

Finalmente, hemos utilizado las últimas 24 secuencias para evaluar el sistema completo. El método de procesado y evaluación se realiza mediante una ventana deslizante de tamaño 36, sin solape. Para cada ventana, se procesa a través del sistema completo y se compara el resultado con el etiquetado del *frame* central de la ventana. Para medir el rendimiento, hemos utilizado la métrica GER (*Gesture Error Rate*), calculada como la suma de falsos positivos, falsos negativos y errores de clasificación, dividida por el total de ventanas evaluadas. Tras procesar las 24 secuencias (414 ventanas) restantes no utilizadas en entrenamiento, se ha obtenido un GER del 9.6%.

6. Discusión

En la sección 1, defendemos que la interacción natural debe emular los comportamientos y gestos que los usuarios emplean en su vida cotidiana. Sin embargo, los gestos incluidos en el conjunto de datos que estudiamos son principalmente convenciones culturales simbólicas. Estos gestos, aunque eficaces para la comunicación en ciertos contextos, requieren que el usuario los aprenda previamente, lo que los convierte en "palabras clave" en lugar de movimientos instintivos y naturales. A pesar de esta limitación, es relevante señalar que su uso puede ser adecuado en diversos contextos y, de hecho, puede constituir un punto de partida valioso para el desarrollo de sistemas de interpretación de gestos en MR. De igual forma, queda patente que un aspecto fundamental en estos sistemas de HGR es definir cuidadosamente un conjunto de gestos a reconocer que sea tanto funcional como intuitivo para el usuario.

Los resultados presentados en la sección 5 indican que, pese a las limitaciones del tamaño del conjunto de datos, es posible desarrollar un sistema funcional con un rendimiento prometedor en la clasificación de gestos. Aunque evaluamos nuestro sistema con la parte etiquetada de SHREC22 y nuestro propio test, no pudimos compararlo directamente con los resultados de SHREC22 por la falta de acceso a su conjunto de prueba, lo que limita la comparación con otros métodos del estado del arte. No obstante, los resultados obtenidos proporcionan un marco de referencia para trabajos futuros.

Uno de los aspectos más destacados es la eficacia del bloque de transformaciones (TRANSF) en las arquitecturas de red. Este bloque ha demostrado ser un recurso para mejorar el reconocimiento de gestos, pudiendo ser incorporado en diversas arquitecturas. Por ejemplo, podemos observar como la adición de este bloque al MLP permite alcanzar resultados de precisión comparables a los obtenidos con arquitecturas más complejas como STRONGER.

No obstante, la tendencia observada en los resultados de los clasificadores durante la validación cruzada sugiere que la falta de muestras impacta de manera significativa en el rendimiento. El hecho de que algunos modelos saturen al 100% de precisión indica que no hay margen de mejora en el aprendizaje de la red, mientras que el rendimiento en el conjunto de test revela un considerable potencial de mejora. Esto resalta la necesidad de un conjunto de datos más grande y diverso, que permita al modelo aprender de una mayor variedad de ejemplos, lo que, a su vez, podría mejorar su capacidad de generalización y robustez.

La métrica GER nos proporciona una perspectiva sobre la eficacia del sistema en condiciones de operación real. Desgraciadamente, no podemos evaluar el sistema completo con nuestras propias muestras de test, ya que no podemos asegurar que lo grabado entre gestos corresponda a manos en el área de captura, ni que, por ejemplo, no se estuviera practicando un gesto antes de grabarlo como tal. Con un GER del 9.6% para las 24 secuencias que no se han visto en entrenamiento, el sistema muestra un nivel aceptable de precisión, aunque este porcentaje indica que hay margen para la mejora, especialmente considerando que estos resultados se obtuvieron con datos grabados de manera similar a los de entrenamiento.

7. Conclusiones y trabajo futuro

En conclusión, en este trabajo presentamos un sistema de reconocimiento de gestos de las manos utilizando datos tridimensionales capturados por las Microsoft HoloLens 2, donde utilizamos una arquitectura de clasificadores en cascada para optimizar el rendimiento del sistema en conjunto. A pesar de las limitaciones impuestas por el tamaño reducido del conjunto de datos, los resultados obtenidos muestran que las transformaciones aplicadas al inicio del modelo mejoran el reconocimiento de gestos, incluso en arquitecturas más simples como un MLP. Sin embargo, los resultados también evidencian la necesidad de un conjunto de datos más amplio para mejorar la generalización, ya que los clasificadores muestran un rendimiento excelente en el conjunto de entrenamiento con validación cruzada, pero aún presentan margen de mejora en el conjunto de test. Finalmente, el sistema completo obtiene un GER del 9.6%, lo cual es un buen punto de partida para futuras optimizaciones. Los resultados obtenidos ponen de manifiesto el potencial de este enfoque, aunque resalta la importancia de seguir explorando nuevas formas de enriquecer los datos y mejorar la robustez del sistema.

Como trabajo futuro, es crucial expandir significativamente el conjunto de datos para abordar el problema de sobreajuste que hemos detectado y mejorar la capacidad de generalización del sistema. También sería esencial desarrollar un *dataset* binario específico para mejorar el rendimiento del detector de gestos frente a no-gestos, permitiendo así un filtrado más preciso. Además, se podría explorar la integración de técnicas de *data augmentation* para aumentar artificialmente el número de muestras sin necesidad de nuevas capturas y mejorar la robustez del sistema. A nivel de arquitectura, es interesante investigar modelos más avanzados que incorporen bloques de transformaciones, así como experimentar con arquitecturas específicas para la clasificación binaria. Por otro lado, queda pendiente evaluar la eficiencia del sistema en cuanto a velocidad y rendimiento en tiempo real, así como evaluarlo en escenarios más cercanos a aplicaciones reales, incluyendo pruebas con gestos realizados de forma espontánea y en ambientes más ruidosos.

Referencias

Asadi, A. R., & Hemadi, R. (2024). Towards Mixed Reality as the Everyday Computing Paradigm: Challenges & Design Recommendations. https://doi.org/10.48550/arxiv.2402.15974

Avola, D., Bernardi, M., Cinque, L., Foresti, G. L., & Massaroni, C. (2018). Exploiting Recurrent Neural Networks and Leap Motion Controller for Sign Language and Semaphoric Gesture Recognition. IEEE Transactions on Multimedia, 21(1), 234-245. https://doi.org/10.1109/tmm.2018.2856094

Balakrishnan, P., & Guo, H.-J. (2024). HoloLens 2 Technical Evaluation as Mixed Reality Guide. Lecture Notes in Computer Science, 145-165. https://doi.org/10.1007/978-3-031-61041-7_10

Bannon, L. (2011). Reimagining HCI: toward a more human-centered perspective. Interactions, 18(4), 50. https://doi.org/10.1145/1978822.1978833

Benedict, J. D., Guliuzo, J. D., & Chaparro, B. S. (2019). The Intuitiveness of Gesture Control with a Mixed Reality Device. Proceedings of the Human Factors and Ergonomics Society, 63(1), 1435-1439. https://doi.org/10.1177/1071181319631403

Caputo, A., Giachetti, A., Giannini, F., Lupinetti, K., Monti, M., Pegoraro, M., & Ranieri, A. (2020). SFINGE 3D: A novel benchmark for online detection and recognition of heterogeneous hand gestures from 3D fingers' trajectories. Computers & Graphics, 91, 232-242. https://doi.org/10.1016/J.CAG.2020.07.014

Caputo, A., Giachetti, A., Soso, S., Pintani, D., D'Eusanio, A., Pini, S., Borghi, G., Simoni, A., Vezzani, R., Cucchiara, R., Ranieri, A., Giannini, F., Lupinetti, K., Monti, M., Maghoumi, M., LaViola, J. J., Le, M. Q., Nguyen, H. D., & Tran, M. T. (2021). SHREC 2021: Skeleton-based hand gesture recognition in the wild. Computers and Graphics (Pergamon), 99, 201-211. https://doi.org/10.1016/j.cag.2021.07.007

Chakraborty, B. K., Sarma, D., Bhuyan, M. K., & MacDorman, K. F. (2018). Review of constraints on vision-based gesture recognition for human-computer interaction. IET Computer Vision, 12(1), 3-15. https://doi.org/10.1049/IET-CVI.2017.0052

De Smedt, Q., Wannous, H., & Vandeborre, J. P. (2016). Skeleton-Based Dynamic Hand Gesture Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1206-1214. https://doi.org/10.1109/CVPRW.2016.153

Dipesh Gyawali. (2023). Mixed Reality: The Interface of the Future. https://doi.org/10.48550/arxiv.2309.00819

Emporio, M., Caputo, A., & Giachetti, A. (2021). STRONGER: Simple TRajectory-based ONline GEsture Recognizer. Eurographics Italian Chapter Proceedings - Smart Tools and Applications in Graphics, STAG, 109-117. https://doi.org/10.2312/stag.20211481

Emporio, M., Caputo, A., Giachetti, A., Cristani, M., Borghi, G., D'Eusanio, A., Le, M. Q., Nguyen, H. D., Tran, M. T., Ambellan, F., Hanik, M., Nava-Yazdani, E., & von Tycowicz, C. (2022). SHREC 2022 track on online detection of heterogeneous gestures. Computers and Graphics (Pergamon), 107, 241-251. https://doi.org/10.1016/j.cag.2022.07.015

Guo, L., Lu, Z., & Yao, L. (2021). Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. IEEE Transactions on Human-Machine Systems, 51(4), 300-309. https://doi.org/10.1109/THMS.2021.3086003

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770-778. https://doi.org/10.1109/CVPR.2016.90

Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2019). Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks. Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019. https://doi.org/10.1109/FG.2019.8756576

Maghoumi, M., & LaViola, J. J. (2018). DeepGRU: Deep Gesture Recognition Utility. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11844 LNCS, 16-31. https://doi.org/10.1007/978-3-030-33720-9_2

Marin, G., Dominio, F., & Zanuttigh, P. (2016). Hand gesture recognition with jointly calibrated Leap Motion and depth sensor. Multimedia Tools and Applications, 75(22), 14991-15015. https://doi.org/10.1007/S11042-015-2451-6

Milgram, P., & Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays. IEICE Transactions on Information and Systems, 77, 1321-1329.

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: a class of displays on the reality-virtuality continuum. Telemanipulator and Telepresence Technologies, 2351. https://doi.org/10.1117/12.197321

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. Neurocomputing, 2(5-6), 183-197. https://doi.org/10.1016/0925-2312(91)90023-5

Parcheta, Z., & Martínez-Hinarejos, C. D. (2017). Sign language gesture recognition using HMM. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10255 LNCS, 419-426. https://doi.org/10.1007/978-3-319-58838-4_46

Ribeiro, R., Ramos, J., Safadinho, D., Reis, A., Rabadão, C., Barroso, J., & Pereira, A. (2021). Web AR Solution for UAV Pilot Training and Usability Testing. Sensors, 21(4), 1456. https://doi.org/10.3390/s21041456

Rokhsaritalemi, S., Sadeghi-Niaraki, A., & Choi, S. -M. (2020). A Review on Mixed Reality: Current Trends, Challenges and Prospects. Applied Sciences, 10(2), 636. https://doi.org/10.3390/app10020636

Soler-Dominguez, J. L., Navas-Medrano, S., & Pons, P. (2024). ARCADIA: A Gamified Mixed Reality System for Emotional Regulation and Self-Compassion. https://doi.org/10.1145/3613904.3642123

Spittle, B., Frutos-Pascual, M., Creed, C., & Williams, I. (2022). A Review of Interaction Techniques for Immersive Environments. IEEE Transactions on Visualization and Computer Graphics, 1-1. https://doi.org/10.1109/TVCG.2022.3174805

Yang, F., Sakti, S., Wu, Y., & Nakamura, S. (2019). Make Skeleton-based Action Recognition Model Smaller, Faster and Better. http://arxiv.org/abs/1907.09658

Zaccardi, S., Frantz, T., Beckwée, D., Swinnen, E., & Jansen, B. (2023). On-Device Execution of Deep Learning Models on HoloLens2 for Real-Time Augmented Reality Medical Applications. Sensors 2023, Vol. 23, Page 8698, 23(21), 8698. https://doi.org/10.3390/S23218698

Zaman Khan, R., & Ibraheem, N. (2012). Hand Gesture Recognition: A Literature Review. International Journal of Artificial Intelligence & Applications, 3(4), 161-174. https://doi.org/10.5121/ijaia.2012.3412