International Journal of Information Science and Technological Applications-UAS
IJISTA
ISSN (en trámite)
https://revistas.uas.edu.mx/index.php/IJISTA

Mayo 2025
Vol.1.
Número:1

Manipulación gestual de robots mediante visión artificial

Gesture-based robots manipulation using computer vision

Adriana Lizbeth González Sarabia[1], Claudio Manuel Domínguez Ulloa[1]
[1]Facultad de Informática Culiacán / Posgrado en Ciencias de la Información, Universidad Autónoma de Sinaloa, México.

Autor de Correspondencia: Adriana Sarabia, adrianagonzalez.fim@uas.edu.mx, https://orcid.org/0009-0001-4749-9908

Recibido: Enero 2025, **Aceptado:** Marzo 2025, **Publicado**: Mayo 2025

Resumen:

Este artículo aborda la manipulación gestual de robots articulados mediante visión artificial, con el objetivo de permitir su control remoto para la realización de tareas específicas. La manipulación gestual se ha convertido en una alternativa intuitiva y eficiente dentro de los sistemas de teleoperación, ya que permite una interacción natural entre el ser humano y el robot. Se presenta una breve taxonomía de los sistemas de control gestual, clasificándolos en dos enfoques principales: aquellos basados en sensores físicos (como guantes o dispositivos inerciales) y los que utilizan visión artificial. Este estudio analiza el segundo enfoque, explorando distintas metodologías que permiten reconocer gestos humanos a partir de imágenes o secuencias de video, procesadas por algoritmos de visión por computadora. Se analizan propuestas existentes que permiten mapear dichos gestos en comandos para el control de robots articulados, considerando aspectos como la precisión del reconocimiento, latencia y la adaptabilidad del sistema. A través del análisis comparativo de estas metodologías, se identifican ventajas, limitaciones y posibles áreas de mejora. Los resultados evidencian el potencial de los sistemas basados en visión artificial para ofrecer una interacción más libre y natural, sin necesidad de equipo adicional, lo que puede traducirse en soluciones más accesibles y versátiles.

Palabras Clave: Manipulación-gestual, teleoperación, robots.

Abstract:

This article addresses the gestural manipulation of articulated robots through computer vision, with the aim of enabling their remote control for the execution of specific tasks. Gestural manipulation has become an intuitive and efficient alternative within teleoperation systems, as it allows for natural interaction between humans and robots. A brief taxonomy of gestural control systems is presented, classifying them into two main approaches: those based on physical sensors (such as gloves of inertial devices) and those that use computer vision. This study analyzes the second approach, exploring various methodologies that enable the recognition of human gestures from images or video sequences processed by computer vision algorithms. Existing proposals that map these gestures into commands for the control of articulated robots are analyzed, considering aspects such as recognition accuracy, latency, and system adaptability. Through a comparative analysis of these methodologies, advantages, limitations, and potential areas for improvement are identified. The results highlight the potential of computer vision-based systems to offer freedom and more natural interaction, without the need for additional equipment, which can lead to more accessible and versatile solutions.

key words: Gestual control, computer vision, teleoperation, human-robot interation.

Agradecimientos a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti) por el apoyo recibido para la realización de esta investigación, CVU: 1298047. Al Posgrado en Ciencias de la Información a la Facultad de Informática de Culiacán de la Universidad Autónoma de Sinaloa, por recibirme en el programa de Maestría en Ciencias de la Información

IJISTA, Vol. 1, Núm. 1, Mayo 2025, ISSN (en trámite)

1. Introducción

La manipulación gestual es la forma más natural de interacción con un robot a través de un ambiente virtual, es por eso que reconocer e interpretar los movimientos del usuario, mediante algún sensor o dispositivo de visión, permite recolectar, analizar y codificar la información para ser transmitida y reproducida por un manipulador [1] [2].

Según [3], existe una taxonomía de los gestos de la mano para este tipo de control de sistema, dividida en dos grupos; los que son basados en sensores y los que son basados en visión. Los que son basados en sensores se pueden agrupar en tres categorías; mediante guantes de datos, mediante señales EMG y mediante WI-FI, por otro lado, los que están basados en visión, son las cámaras monoculares, binoculares y de profundidad como la RGB-D de Intel y los dispositivos Kinect de Microsoft, por mencionar algunos.

En este artículo se analizarán las propuestas metodológicas que se han presentado en el estado del arte, donde se hace el uso de la manipulación gestual mediante la visión artificial, ósea que el operador no necesita cargar en sus manos algún dispositivo físico para control el robot, el cual lo hace a distancia por medio de un sistema de realidad virtual, en el que la escena real está simulada.

Entonces, los entornos de realidad virtual, según las investigaciones recientes, se definen como una herramienta indispensable para el seguimiento del progreso en la manipulación gestual de robots, permitiendo una evaluación precisa y una retroalimentación en tiempo real de los indicadores de desempeño del sistema de control [4].

Actualmente existen diferentes formas de capturar los movimientos para poder obtener información del estado actual del cuerpo humano, de sus extremidades o articulaciones [5]. Los sistemas MoCap (Motion capture o captura en movimiento), permiten la captura de los movimientos lineales y coordenadas angulares, así como, velocidades y aceleraciones [6]. Sistemas que tienen incorporadas técnicas con marcadores y otras que no lo requieren como Kinect, estos tipos de control de sistemas, se pueden controlar a través de sensores o a través de la visión [3].

2. Trabajos Relacionados

En [7], se desarrolla un sistema de interacción humano-robot (HRI) que utiliza gestos dinámicos de la mano para controlar robots cuadrúpedos equipados con un brazo robótico. El sistema emplea un marco Depth-MediaPipe para la extracción precisa de coordenadas tridimensionales (3D) de 21 puntos clave de la mano. Posteriormente, se implementa un modelo Semantic-Pose to Motion (SPM) que interpreta tanto la pose como la semántica de los gestos para traducirlos en acciones mecánicas en tiempo real, incluyendo la locomoción del robot y el seguimiento del efector final del brazo robótico.

En [8], se diseñó un sistema de control gestual para un brazo robótico en líneas de ensamblaje pequeñas. Utilizando visión por computadora, el sistema reconoce gestos específicos de la mano para controlar los movimientos del brazo robótico, mejorando la eficiencia y reduciendo la necesidad de interfaces físicas tradicionales.

En [9], se desarrolló un algoritmo para la teleoperación de robots móviles basado en el reconocimiento de gestos utilizando un sensor LeapMotion. El sistema implementa un filtro Gasussiano para suavizar y eliminar el ruido de los datos de gestos recogidos, mejorando la robustez y estabilidad del movimiento del robot. El control se realiza a través de una estructura cliente/servidor, permitiendo la asociación de gestos con comandos específicos del robot.

En [10], presentaron AutoNav, un sistema de teleoperación para múltiples robots que utiliza el reconocimiento de la palma de la mano como interfaz de control. Integrando el marco MediaPipe con el Sistema Operativo de Robots (ROS), el sistema permite la navegación autónoma y comandos gestuales interpretados a través de visión por computadora. Los resultados mostraron una reducción del 50% en el tiempo de ejecución y una disminución en las colisiones durante la teleoperación de los robots TurtleBot3.

En [11], se propone un sistema de reconocimiento de gestos de largo alcance mediante una cámara web convencional, orientado a la interacción humano-robot (HRI). El enfoque combina técnicas de visión artificial con aprendizaje automático para detectar gestos a distancias mayores a las tradicionales donde se requería cierta proximidad del usuario dependiendo del sensor, de esta manera se facilita la interacción sin necesidad de sensores especializados. A pesar de su bajo costo, la precisión disminuye conforme aumenta la distancia entre el usuario y la cámara.

En [12], se introduce un modelo de aprendizaje continuo de gestos de la mano para HRI, que permite que el sistema incorpore nuevos gestos a lo largo del tiempo sin olvidar los anteriores. Se utiliza una arquitectura basada en redes neuronales que se actualiza dinámicamente con nuevos datos, promoviendo una adaptación constante a diferentes usuarios y entornos. Esta característica lo hace adecuado para entornos cambiantes, aunque depende de la calidad de los datos previos para su rendimiento.

En [13], se presenta un modelo híbrido profundo basado en una red neuronal convolucional (CNN) y memoria a largo plazo y corto plazo (LSTM) para el reconocimiento de gestos dinámicos en interfaces de

interacción humano-computadora. El sistema extrae características espaciales y temporales de secuencia de video, permitiendo una detección precisa de gestos en tiempo real. Este enfoque es eficaz para comandos continuos, aunque requiere recursos computacionales elevados para su entrenamiento y ejecución.

En [14], se realiza una revisión exhaustiva sobre los sistemas de percepción en entornos industriales para HRI, analizando diferentes enfoques sensoriales, incluyendo cámaras RGB, sensores de profundidad y fusión de datos. El estudio destaca los desafíos de operar en ambientes con ruido visual y restricciones espaciales, proporcionando soluciones mediante sistemas híbridos y algoritmos robustos. Se identifican áreas clave de mejora en la integración entre visión artificial y control robótico.

En [15], se desarrolla un marco basado en aprendizaje profundo para reconocer gestos tanto estáticos como dinámicos utilizando cámaras RGB. El sistema emplea atención espacial-temporal para capturar secuencia de movimientos complejas mejorando la precisión en entornos no estructurados. Este enfoque es útil para aplicaciones donde los gestos varían ampliamente, aunque requiere una fase intensiva de entrenamiento previo.

En [16], se introduce GestLLM, un sistema innovador que aprovecha modelos de lenguaje de gran escala (LLM) junto con visión por computadora para interpretar gestos manuales en interacciones humano-robot. El sistema combina información visual con inferencia semántica, permitiendo la comprensión de intenciones más allá de gestos predefinidos. Su principal fortaleza es la flexibilidad interpretativa, aunque depende de la calibración precisa entre los modelos visuales y lingüísticos.

3. Metodología.

Esta sección presenta una revisión detallada de las metodologías actuales en la manipulación gestual de robots mediante visión artificial, enfocándose en sistemas que no requieren dispositivos físicos adicionales para el reconocimiento de gestos.

Para la presente revisión metodológica no se aplicó un criterio sistemático o filtrado especializado en la selección de los artículos analizados. En su lugar, se optó por incluir únicamente publicaciones de acceso abierto que abordaran directamente la manipulación gestual de robots mediante visión artificial. La elección de estos trabajos se fundamentó en la disponibilidad pública de los contenidos y en su pertinencia con respecto al tema central de esta investigación. A partir de esta base, se realiza un análisis exploratorio de cada propuesta seleccionada, desglosando su funcionamiento desde la etapa de captura o entrada de datos, pasando por el reconocimiento de gestos y el tratamiento de la información, hasta llegar a la ejecución final de la tarea asignada al robot. Este enfoque permite identificar las características distintivas de cada metodología y establecer comparaciones claras entre ellas.

Tabla 1. Comparación de las Metodologías.

Metodología	Captura	Gestos	Aplicación	Ventajas	Limite
Xie et al. (2025)	Cámara RGB-D MediaPipe	Dinámicos	Robot cuadrúpedo con brazo	Control preciso en 3D	Configuración compleja
Angelidis y Bampis (2025)	Cámara RGB	Estáticos	Brazo en ensamblaje	Implementación sencilla	Limitados a gestos definidos
Chen et al. (2024)	Sensor LeapMotion	Ambos	Robot móvil	Suavizado de movimientos. Cliente/servidor	Latencia de 140-200ms
Zick et al. (2024)	Cámara RGB con MediaPipe	Estáticos	Múltiples robots	Menos colisiones	Sensible a la luz
Banami et al. (2024)	Webcam (larga distancia)	Estáticos	Interacción H-R de largo alcance	Bajo costo	Precisión reducida con la distancia
Cucurull & Garrell (2023)	Cámara RGB	Estáticos	Aprendizaje continuo	Adaptación a nuevos gestos	Dependencia de datos previos
Ramalingam & Angappan. (2023)	Cámara RGB + CNN.LTSM	Dinámicos	Interacción hombre-máquina	Reconocimiento preciso en tiempo real	Requiere alto poder computacional
Bonci et al. (2021)	Variada (RGB, sensores)	Ambos	Entornos industriales	Integración sensorial robusta	Complejidad del entorno
Mazhar et al. (2021)	Cámara RGB	Estáticos y dinámicos	Interfaces H-R en visión artificial	Modelo híbrido con atención especial	Alto costo computacional para entrenamiento
Kobzarev et al. (2025)	Cámara RGB + MediaPipe + LLM	Estáticos y dinámicos	Amplia interacción H-R en leguaje natural	Interpretación de flexible de gestos	Requiere ajuste del modelo base

3.1. Técnica de captura de gestos

La captura de gestos es fundamental para la interacción humano-robot (HRI) y se basa en el uso de sensores ópticos, cámaras RGB-D o dispositivos portátiles. En [7], emplean una cámara de profundidad para capturar movimientos dinámicos de la mano, utilizando el algoritmo *MediaPipe Hands* para detectar puntos clave. La representación matemática de la posición de la mano en el espacio 3D se cómo:

$$\boldsymbol{P}(t) = [x(t), y(t), z(t)]^T \qquad (1)$$

Donde $\boldsymbol{P}(t)$ es el vector de posición en el tiempo t . Bamani et al. [11] proponen un sistema basado en cámaras web estándar, aplicando filtros de preprocesamiento como la normalización del histograma:

$$I_{norm}(u, v) = \frac{I(u,v)-\mu}{\sigma} \qquad (2)$$

donde μ y σ son la media y desviación estándar de la imagen I. Por otro lado, Kobzarev et al. [16] integran modelos de lenguaje grande *(LLMs)* para interpretar gestos complejos mediante embeddings semánticos.

3.2. Procesamiento y reconocimiento de gestos

El reconocimiento de gestos se realiza mediante técnicas de aprendizaje automático. Mazhar et al. [15] utilizan redes neuronales convolucionales (CNN) para gestos estáticos y redes LSTM para dinámicos. La función de pérdida empleada es la entropía cruzada:

$$L = -\sum_{i-1}^{N} \boldsymbol{y}_i \, log(\hat{y}_i) \quad (3)$$

Donde $\boldsymbol{y}_i$ y $\hat{y}_i$ son las etiquetas reales y predichas. Chen et al. [9] proponen *GestureMoRo,* un algoritmo que combina *Random Forests* con transformadas de Fourier para caracterizar gestos:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)_{e^{-j\omega t}} \, dt \quad (4)$$

Ramalingam y Angappan [13] introducen un modelo híbrido CNN-SVM, donde las características extraídas por la CNN se clasifican con una máquina de vectores soporte (SVM). Cucurull y Garrell [12] abordan el aprendizaje continuo (*continual learning*) para adaptarse a nuevos gestos sin olvidar los anteriores, usando una función de regularización:

$$L_{total} = L_{new} + \lambda \parallel \theta - \theta_{prev} \parallel^2 \quad (5)$$

3.3. Control robótico y ejecución

La traducción de gestos a acciones robóticas requiere mapeo espacio-temporal. Angelidis y Bampis [8] definen una matriz de transformación homogénea para mover un brazo robótico:

$$T = \begin{bmatrix} R & d \\ 0 & 1 \end{bmatrix} \quad (6)$$

donde **R** es la matriz de rotación y **d** el vector de traslación. Zick et al. [10] implementan un sistema de teleoperación para múltiples robots usando gestos, con retroalimentación haptica basada en la ley de control proporcional-derivativa (PD):

$$\boldsymbol{u}(t) = K_p \boldsymbol{e}(t) + K_d \frac{\boldsymbol{de}(t)}{dt} \quad (7)$$

Bonci et al. [14] destacan la fusión de datos multimodales (ej. visión e IMUs) para mejorar la robustez en entornos industriales. Finalmente, Xie et al. [7] asignan gestos a comandos de alto nivel para robots cuadrúpedos, como:

$$COMANDO = \begin{cases} Moverse\ adelante & si\ \Delta y > \tau \\ Girar & si\ \Delta\theta > \gamma \end{cases} \quad (8)$$

Estas ecuaciones y métodos ilustran la integración de percepción, decisión y acción en sistemas HRI modernos.

3.4. Evaluación comparativa

Los resultados comparativos (Tabla 4) destacan las diferencias clave entre los enfoques analizados. Los sistemas basados en cámaras RGB-D [7] y modelos híbridos [13] logran una precisión superior al 94%, pero requieren recursos computacionales elevados, mientras que las soluciones con cámaras web [11] ofrecen accesibilidad a costa de menor robustez. La latencia se mantiene por debajo de los 200 ms en la mayoría de los casos, siendo crítica para aplicaciones en tiempo real. La Figura 1 ilustra la relación entre precisión y latencia, evidenciando el equilibrio necesario entre rendimiento y eficiencia.

Tabla 4. Métricas clave.

Métrica	**[7]**	**[11]**	**[9]**	**[13]**
Precisión (%)	98.2	72.5	99.1	94.3
Latencia (ms)	120	210	140	180
Requerimiento de GPU	SÍ	No	No	Sí

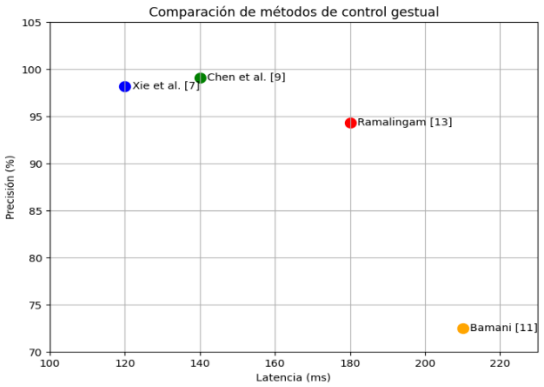

Fig. 1. Comparación de Métodos

4. Resultados

Los sistemas de control gestual analizados muestran un espectro de desempeños según su enfoque tecnológico. Las soluciones con cámaras RGB-D como la propuesta por Xie et al. [7] alcanzan el 98.2% de precisión mediante el algoritmo MediaPipe Hands, superando significativamente a los sistemas basados en webcam convencional [11] que logran 72.5% de precisión pero con limitaciones en rangos extendidos. Los modelos híbridos CNN-LSTM de Ramalingam y Angappan [13] demuestran versatilidad con 94.3% de precisión en gestos dinámicos, aunque demandan mayor capacidad computacional. En eficiencia temporal, mientras Xie et al. [7] registran 120ms de latencia usando ROS, los sistemas con LeapMotion [9] presentan retardos de 140-200ms debido a su filtrado Gaussiano. La integración multimodal de Bonci et al. [14] y la adaptabilidad de GestLLM [16] mediante LLMs emergen como enfoques prometedores para entornos complejos, aunque con desafíos en escalabilidad.

5. Análisis de Resultados

El análisis comparativo revela que la elección tecnológica implica compromisos clave: las cámaras RGB-D [7]

ofrecen precisión excepcional, pero con requerimientos técnicos que pueden limitar su adopción masiva, contrastando con las soluciones accesibles de Bamani et al. [11] pero menos robustas. Los resultados de Chen et al. [9] y Zick et al. [10] confirman que latencias bajo 200ms son viables para teleoperación, aunque aplicaciones críticas necesitarían las optimizaciones de edge computing sugeridas por Angelidis y Bampis [8]. El aprendizaje continuo de Cucurull y Garrell [12] y los modelos híbridos de Mazhar et al. [15] representan avances significativos en adaptabilidad, pero como señala Kobzarev et al. [16], la interpretación semántica de gestos sigue dependiendo de grandes volúmenes de entrenamiento. La revisión de Bonci et al. [14] enfatiza que la integración sensorial robusta sigue siendo el mayor reto para entornos industriales reales, completando así el panorama tecnológico actual en control gestual por visión artificial.

6. Conclusiones

El presente artículo versa sobre la manipulación gestual de robots mediante visión artificial, presenta una revisión exhaustiva de las metodologías actuales en este campo, así como los trabajos relacionados más relevantes al respecto. En lo que respecta a la visión artificial, estos sistemas ofrecen una interacción más libre y natural entre humanos y robots, sin necesidad de equipo adicional.

Así mismo, se presentan diversos enfoques metodológicos, incluyendo cámaras RGB-D, cámaras web convencionales, modelos híbridos CNN-LSTM y aprendizaje continuo, cada uno con sus ventajas y limitantes.

En la precisión y latencia, los sistemas basados en cámaras RGB-D y modelos híbridos logran precisiones superiores al 94%, pero requieren recursos computacionales elevados. La latencia se mantiene por debajo de los 200 ms en la mayoría de los casos. Derivado de este análisis exhaustivo, se tienen algunos desafíos, así como áreas de mejora, como la integración sensorial robusta en entornos industriales, la interpretación semántica de gestos y la escalabilidad de los sistemas son áreas clave para futuras investigaciones.

Los sistemas de control gestual pueden ser utilizados en diversas aplicaciones, como la teleoperación de robots en entornos industriales o la interacción humano-robot en entornos de servicio.

El análisis presentado en este artículo puede servir como base para futuras investigaciones en el campo de la manipulación gestual de robots mediante visión artificial.

Actualmente se trabaja en una propuesta de un sistema de control de robots basado en reconocimiento de gestos mediante visión artificial, con el objetivo de lograr una interacción natural y sin contacto. Esta propuesta permitirá una manipulación eficiente en tiempo real, con alta precisión y baja latencia. Las pruebas experimentales demuestran su viabilidad y comparabilidad con trabajos recientes del estado del arte [7]-[16].

Entre sus principales ventajas se encuentran la facilidad de uso, accesibilidad y la eliminación de dispositivos portables. No obstante, hasta el momento se trabaja en las limitaciones en ambientes no controlados y ante gestos ambiguos. Como propuesta de trabajo se propone incorporar modelos de aprendizaje continuo, aumentar el repertorio de gestos y evaluar su uso en aplicaciones reales de teleoperación o manufactura colaborativa.

7. Referencias

[26]Tsarouchi, P., Athanasatos, A., Makris, S., Chatzigeorgiou, X., & Chryssolouris, G. (2016). High level robot programming using body and hand gestures. *Procedia Cirp*, *55*, 1-5.

[27]Padilla, A. F., Peña, C. A., & Moreno-Contreras, G. G. (2020, November). Advances in industrial robots programming applying gestural guidance techniques. In *Journal of Physics: Conference Series* (Vol. 1704, No. 1, p. 012001). IOP Publishing.

[28]Qi, J., Ma, L., Cui, Z., & Yu, Y. (2024). Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, *10*(1), 1581-1606.

[29]Almansour, A. M. (2024). The Effectiveness of Virtual Reality in Rehabilitation of Athletes: A Systematic Review and Meta-Analysis. Journal of Pioneering Medical Sciences, 13, 147-154.

[30]Yavuz, E., Şenol, Y., Özçelik, M., & Aydın, H. (2021). Design of a String Encoder-and-IMU-Based 6D Pose Measurement System for a Teaching Tool and Its Application in Teleoperation of a Robot Manipulator. *Journal of Sensors*, *2021*(1), 6678673.

[31]Gómez Echeverry, L. L., Jaramillo Henao, A. M., Ruiz Molina, M. A., Velásquez Restrepo, S. M., Páramo Velásquez, C. A., & Silva Bolívar, G. J. (2018). Sistemas de captura y análisis de movimiento cinemático humano: Una revisión sistemática. *Prospectiva*, *16*(2), 24-34.

[32]Xie, J., Xu, Z., Zeng, J., Gao, Y., & Hashimoto, K. (2025). Human–Robot Interaction Using Dynamic Hand Gesture for Teleoperation of Quadruped Robots with a Robotic Arm. *Electronics*, *14*(5), 860.

[33]Angelidis, G., & Bampis, L. (2025). Gesture-Controlled Robotic Arm for Small Assembly Lines. *Machines*, *13*(3), 182.

[34]Chen, L., Li, C., Fahmy, A., & Sienz, J. (2024). GestureMoRo: an algorithm for autonomous mobile robot teleoperation based on gesture recognition. *Scientific Reports*, *14*(1), 6199.

[35]Zick, L. A., Martinelli, D., Schneider de Oliveira, A., & Cremer Kalempa, V. (2024). Teleoperation system

for multiple robots with intuitive hand recognition interface. *Scientific Reports*, *14*(1), 1-11.

[36]Bamani, E., Nissinman, E., Meir, I., Koenigsberg, L., & Sintov, A. (2024). Ultra-range gesture recognition using a web-camera in human–robot interaction. *Engineering Applications of Artificial Intelligence*, *132*, 108443.

[37]Cucurull, X., & Garrell, A. (2023). Continual Learning of Hand Gestures for Human-Robot Interaction. *arXiv preprint arXiv:2304.06319*.

[38]Ramalingam, B., & Angappan, G. (2023). A deep hybrid model for human-computer interaction using dynamic hand gesture recognition. *Computer Assisted Methods in Engineering and Science*, *30*(3), 263-276.

[39]Bonci, A., Cen Cheng, P. D., Indri, M., Nabissi, G., & Sibona, F. (2021). Human-robot perception in industrial environments: A survey. *Sensors*, *21*(5), 1571.

[40]Mazhar, O., Ramdani, S., & Cherubini, A. (2021). A deep learning framework for recognizing both static and dynamic gestures. *Sensors*, *21*(6), 2227.

[41]Kobzarev, O., Lykov, A., & Tsetserukou, D. (2025). GestLLM: Advanced Hand Gesture Interpretation via Large Language Models for Human-Robot Interaction. *arXiv preprint arXiv:2501.07295*.