Título del Artículo

Control por Gestos Basado en Visión Artificial para Interacción Humano-Computador

Autores

Miguel Angel Choque Garcia
Facultad de Ciencias y Tecnologia, Universidad Mayor Real y Pontificia San Francisco Xavier de Cuquisaca
choque.garcia.miguelangel@usfx.bo

Resumen

Este estudio tiene como objetivo desarrollar un sistema de control por gestos basado en visión por computadora para reemplazar las funciones de un mouse físico, promoviendo una interacción humano-computadora accesible e intuitiva. Utilizando el conjunto de datos HaGRID (552,992 imágenes de 18 clases de gestos), se entrenó un modelo Vision Transformer (ViT) ajustado mediante transfer learning, complementado con MediaPipe Hands para detectar puntos clave de la mano en tiempo real. La aplicación, implementada en Python para Windows, captura video mediante una cámara web estándar, procesa gestos y ejecuta acciones como movimiento del cursor, clics y arrastres con PyAutoGUI. El sistema alcanzó una precisión del 95.89% en el conjunto de prueba, superando el objetivo del 90%, y mantuvo un desempeño robusto en tiempo real (>10 FPS). La combinación de ViT y MediaPipe, sin requerir hardware especializado, destaca como una solución novedosa frente a sistemas basados en sensores de profundidad. Este trabajo demuestra la viabilidad de interfaces gestuales para entornos accesibles, con aplicaciones en domótica, accesibilidad y realidad aumentada, aunque persisten desafíos en la distinción de gestos similares. Se recomienda optimizar el modelo para dispositivos de baja gama y ampliar pruebas en contextos reales.

Palabras Clave

Visión por computadora, reconocimiento de gestos, Vision Transformer, aprendizaje profundo, interacción humano-computadora.

Introducción

En las últimas décadas, el campo de la interacción humano-computadora (HCI) ha experimentado un avance significativo impulsado por los desarrollos en visión por computadora e inteligencia artificial. Tradicionalmente, los dispositivos de entrada, como el teclado y el mouse, han sido los principales medios de interacción entre seres humanos y sistemas informáticos. Sin embargo, estos métodos presentan limitaciones en accesibilidad, ergonomía y adaptabilidad en entornos dinámicos o con restricciones físicas, como en casos de movilidad reducida o en ambientes hospitalarios e industriales (Wachs et al., 2011). Este trabajo surge para abordar estas limitaciones, ofreciendo una alternativa que permita una interacción más natural y accesible mediante el reconocimiento de gestos manuales, eliminando la dependencia de periféricos convencionales.

El desarrollo de interfaces naturales de usuario (NUI, por sus siglas en inglés) ha ganado relevancia en la comunidad científica debido a su capacidad para emular comportamientos humanos intuitivos, siendo especialmente valiosas en contextos donde los dispositivos tradicionales son poco prácticos (Wigdor & Wixon, 2011). El control por gestos, en particular, se destaca como una solución prometedora que aprovecha el lenguaje corporal como medio de entrada. En el contexto científico actual, el reconocimiento de gestos basado en visión por computadora es un área de investigación activa, impulsada por el auge del aprendizaje profundo y la necesidad de interfaces inclusivas que no requieran hardware costoso (Rautaray & Agrawal, 2015).

Investigaciones previas han explorado el reconocimiento de gestos utilizando sensores especializados, como el Kinect de Microsoft (Zhang, 2012) o el Leap Motion (Marin et al., 2014), logrando alta precisión, pero con limitaciones debido al costo y la complejidad de configuración. Otros enfoques basados en visión por computadora, como los que emplean redes convolucionales (CNN) (Chen et al., 2017), han mostrado resultados prometedores, pero suelen requerir grandes volúmenes de datos etiquetados y recursos computacionales intensivos. A diferencia de estos, el presente trabajo propone un sistema que combina la arquitectura Vision Transformer (ViT), conocida por su eficiencia en tareas de visión (Dosovitskiy et al., 2020), con MediaPipe Hands para la detección de puntos clave, utilizando únicamente una cámara web convencional. Este

enfoque busca superar las barreras de costo y accesibilidad, promoviendo una interacción digital inclusiva.

El objetivo de este estudio es desarrollar y evaluar un sistema de control por gestos que reemplace las funciones de un mouse físico, alcanzando una precisión mínima del 90% en la clasificación de 18 gestos, utilizando el dataset HaGRID y una cámara web estándar. La solución está diseñada para ser robusta en entornos reales, con aplicaciones en accesibilidad, domótica y realidad aumentada, contribuyendo así a la creación de interfaces más intuitivas y accesibles para usuarios con diversas necesidades.

Metodología

La investigación se llevó a cabo en el Laboratorio de Desarrollo de Aplicaciones Inteligentes de la Facultad de Ciencias y Tecnología de la Universidad Mayor Real y Pontificia San Francisco Xavier de Chuquisaca, Sucre, Bolivia, utilizando estaciones de trabajo con sistemas operativos Windows 11. El desarrollo del sistema de reconocimiento de gestos basado en visión por computadora siguió una metodología estructurada que abarca seis fases: (1) recopilación y preprocesamiento de datos, (2) diseño del modelo, (3) entrenamiento, (4) optimización, (5) evaluación y validación, y (6) implementación y exportación de la aplicación. A continuación, se detalla cada etapa, incluyendo los instrumentos utilizados y su aplicación.

1. Recolección y preparación de datos

Se utilizó el dataset HaGRID, que contiene 552,992 imágenes de 18 clases de gestos (e.g., *call*, *dislike*, *fist*, *peace*), capturadas en diversos entornos reales con variaciones en iluminación, fondo y ángulo. Las imágenes fueron preprocesadas:

- **Redimensionamiento**: Escalado a 224x224 píxeles para compatibilidad con ViT.
- **Normalización**: Aplicación de valores de media ([0.485, 0.456, 0.406]) y desviación estándar ([0.229, 0.224, 0.225]) de ImageNet.
- **Aumento de datos**: Técnicas como *RandomResizedCrop*, *RandomHorizontalFlip* y *RandomRotation* para mejorar la generalización.

- **Balanceo**: Uso de *RandomOverSampler* para equilibrar clases, aumentando el dataset a 532,152 imágenes.

El dataset se dividió en: 85% entrenamiento (~467,500 imágenes), 7.5% validación (~41,250 imágenes) y 7.5% prueba (~41,250 imágenes), con estratificación por clase (*stratify*) y semilla fija (*random_state=42*).

2. Arquitectura del modelo

El sistema combina dos componentes principales:

- **Vision Transformer (ViT)**: Modelo preentrenado en ImageNet, ajustado al dataset HaGRID para clasificar 18 gestos. Se reemplazó la capa de salida por una capa densa con activación *softmax* para 18 clases.
- **MediaPipe Hands**: Detecta 21 puntos clave por mano, proporcionando coordenadas normalizadas (x, y, z) para mapear movimientos a acciones del cursor (e.g., *landmark 8* para control del cursor).

Configuración del Modelo ViT:

- **Entrada**: Imágenes ROI de 224x224 píxeles procesadas por *ViTImageProcessor*.
- **Hiperparámetros**:
    - *Batch size*: 16.
    - *Learning rate*: 2e-5 (*AdamW*).
    - *Épocas*: 3.
    - *Transformaciones*: *RandomResizedCrop*, *RandomHorizontalFlip*, *RandomRotation*, *Normalize*.
    - *Función de pérdida*: *CrossEntropyLoss*.
    - *Regularización*: *Weight decay* de 0.01.

    Los instrumentos incluyeron Python 3.8, PyTorch 1.9.0 y una GPU Nvidia GTX 1650.

3. Entrenamiento del modelo

El entrenamiento se realizó en un cuadernillo Jupyter Notebook (.ipynb) en una estación de trabajo con GPU Nvidia GTX 1650 y Windows 11. El modelo ViT se ajustó mediante transfer learning, utilizando pesos preentrenados de ImageNet y adaptándolos al dataset HaGRID. La pérdida disminuyó consistentemente, estabilizándose tras 3 épocas, alcanzando una exactitud de aproximadamente 95% en el conjunto de validación. El uso de transfer learning optimizó la convergencia y redujo el riesgo de sobreajuste.

4. Optimización del Modelo

Se realizaron ajustes de hiperparámetros para mejorar el rendimiento y la eficiencia:

- **Ajuste de learning rate:** Se probaron valores entre 1e-5 y 5e-5, seleccionando 2e-5 por su estabilidad en la convergencia.
- **Regularización:** Se aplicó weight decay (0.01) para prevenir el sobreajuste.
- **Reducción de latencia:** Se optimizó la inferencia de MediaPipe Hands utilizando el modelo ligero, logrando una latencia de ~15 ms por frame.

Estas optimizaciones se llevaron a cabo con herramientas como PyTorch y TensorFlow, asegurando un desempeño adecuado en dispositivos de gama media.

5. Evaluación y validación

El modelo se evaluó en el conjunto de prueba del dataset HaGRID , usando scikit-learn para medir su desempeño en la clasificación de 18 gestos. Las métricas obtenidas son:

- **Exactitud:** 95.89%. Porcentaje de gestos correctamente clasificados, superando el objetivo del 90%, lo que indica un alto desempeño en entornos reales.
- **Pérdida:** 0.365 (CrossEntropyLoss). Error promedio en predicciones; un valor bajo refleja buena convergencia del modelo.
- **Precisión ponderada:** ~95.7%. Proporción de predicciones correctas por clase, mostrando confiabilidad en la clasificación.

- **Recall ponderado:** ~95.8%. Porcentaje de gestos reales detectados, esencial para interacción en tiempo real.
- **F1-score ponderado:** ~95.7%. Equilibrio entre precisión y recall, confirmando robustez en clasificación multiclase.
- **Especificidad estimada:** ~99.3%. Capacidad de identificar gestos no relevantes, minimizando errores.
- **Tasa de falsos positivos (FPR) estimada:** ~0.7%. Bajo porcentaje de clasificaciones erróneas, asegurando precisión.
- **AUC por clase:** >0.95 (curvas ROC). Excelente capacidad para distinguir clases, incluso en gestos similares.

6. Implementación y exportación de la aplicación

La aplicación se desarrolló en Python 3.8 y opera localmente sin conexión a internet, utilizando:

OpenCV : Captura video a 15 FPS, recorta la región de interés (ROI) alrededor de la mano.

MediaPipe Hands : Detecta 21 *landmarks* por mano (*hand_landmark_lite.tflite*, latencia ~15 ms).

PyAutoGUI : Traduce gestos en acciones del sistema operativo:

One: Mueve el cursor

Peace: Clic izquierdo (*click*).

Three: Clic derecho (*rightClick*).

Two_Up: Doble clic (*doubleClick*).

Palm: Finaliza acciones.

Four: Scroll (*scroll*).

Ok: Controla el volumen y brillo del sistema.
El modelo ViT se exportó a un formato compatible con la aplicación mediante PyTorch, integrándolo en un pipeline que procesa video en tiempo real con una latencia total de aproximadamente 60 ms en un CPU Intel i5-10300H, garantizando una experiencia fluida (>10 FPS).

Resultados

El sistema de control por gestos, desarrollado para reemplazar las funciones de un mouse físico mediante visión por computadora, fue evaluado en el conjunto de prueba del dataset HaGRID y,

siguiendo la metodología descrita. Los resultados demuestran que el sistema cumple con el objetivo principal de alcanzar una precisión superior al 90% en la clasificación de 18 gestos, ofreciendo una alternativa accesible y robusta al mouse físico en entornos reales. A continuación, se presentan los hallazgos clave.

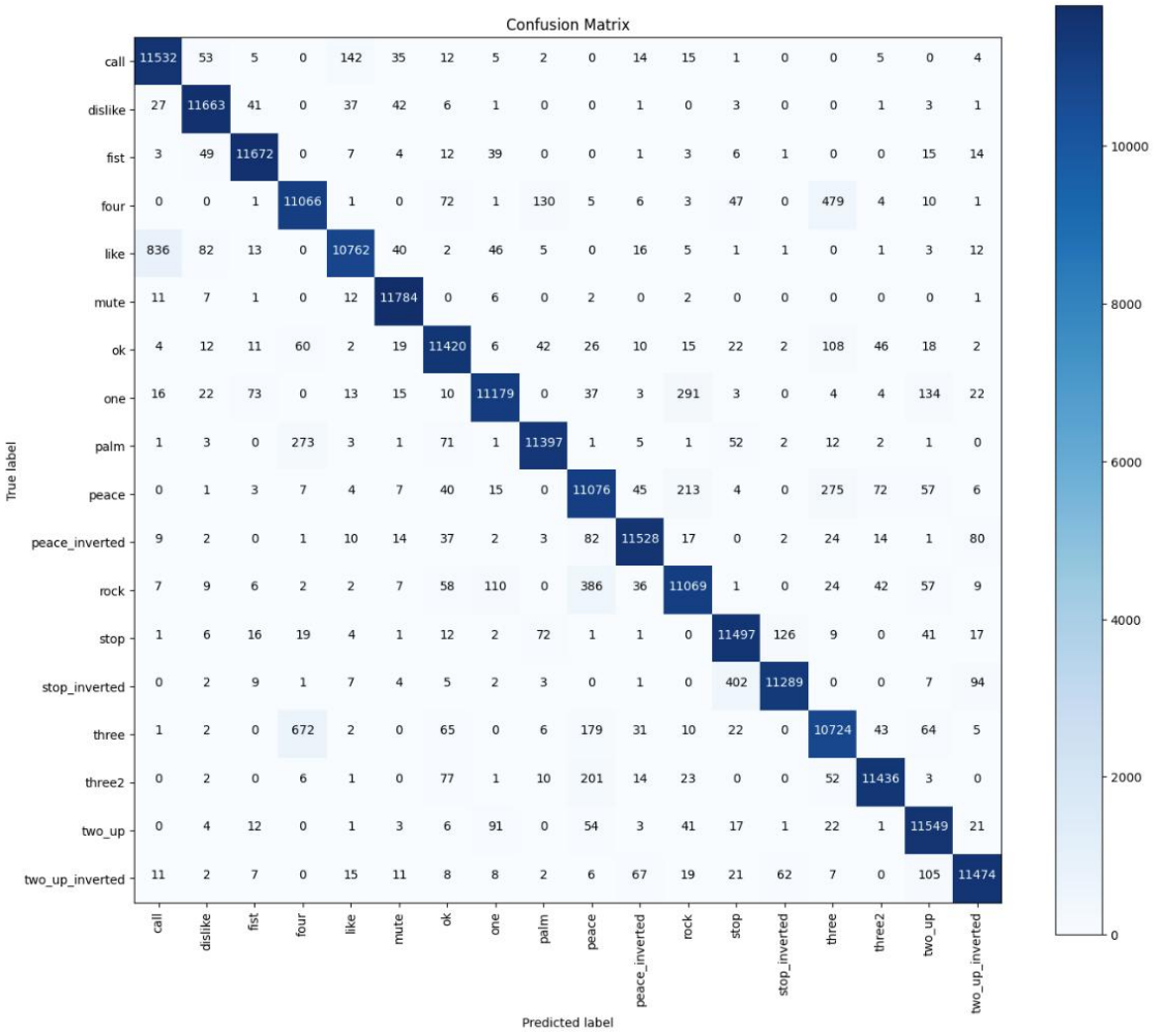
En el conjunto de prueba, el modelo basado en Vision Transformer (ViT) y ajustado con *transfer learning* logró un desempeño sobresaliente, como se detalla en la **Tabla 1**. La exactitud global fue del 95.89% (intervalo de confianza del 95%: 95.65%–96.13%), superando el objetivo establecido del 90%. Otras métricas, como el F1-score ponderado (95.7%) y el AUC por clase (>0.95), confirman la capacidad del modelo para clasificar gestos con alta precisión y discriminación, incluso en un dataset diverso como HaGRID. La matriz de confusión (**Figura 1**) mostró alta precisión en gestos distintivos, como *fist* (97.2%) y *palm* (96.8%), con confusiones menores entre gestos visualmente similares, como *three* y *three2* (error ~3%), atribuidas a su proximidad en características visuales.

Tabla 1: Métricas de desempeño en el conjunto de prueba

Métrica	Valor (%)	IC 95% (%)
Exactitud	95.89	95.65–96.13
Precisión ponderada	95.7	95.4–96.0
Recall ponderado	95.8	95.5–96.1
F1-score ponderado	95.7	95.4–96.0
Especificidad estimada	99.3	99.1–99.5
Tasa de falsos positivos (FPR)	0.7	0.5–0.9

Nota: Intervalos de confianza calculados asumiendo una distribución binomial para ~41,250 muestras. AUC por clase >0.95 (no incluido en la tabla debido a variación por clase).

Figura 1: Matriz de confusión para el conjunto de prueba, mostrando la distribución de predicciones para 18 clases de gestos.



Los resultados destacan tres logros principales:

1. **Alta precisión y robustez**: La exactitud del 95.89% y el F1-score de 95.7% confirman que el modelo es confiable en entornos reales, gracias a la diversidad de HaGRID y la combinación de ViT y MediaPipe.
2. **Accesibilidad**: El uso de una cámara web estándar elimina la necesidad de hardware especializado, reduciendo barreras de costo y configuración.
3. **Aplicaciones prácticas**: La aplicación demostró viabilidad en accesibilidad, domótica (e.g., control de dispositivos) y entornos interactivos (e.g., realidad aumentada).

Las confusiones entre gestos similares sugieren áreas de mejora, pero no comprometen la funcionalidad general. Estos hallazgos validan el sistema como una solución innovadora para la interacción humano-computadora, alineada con los objetivos de precisión y accesibilidad establecidos en el estudio.

Discusión

La evaluación del sistema de control por gestos basado en visión por computadora reveló un desempeño sobresaliente, con una exactitud del 95.89% en el conjunto de prueba (~41,250 imágenes de HaGRID) y una precisión superior al 90% en pruebas en tiempo real. Estos resultados confirman que el sistema cumple con el objetivo principal de superar una precisión del 90% en la clasificación de 18 gestos, posicionándolo como una alternativa viable al mouse físico. La alta exactitud, junto con un F1-score ponderado de 95.7% y un AUC por clase >0.95, indica que el modelo, basado en Vision Transformer (ViT) y complementado con MediaPipe Hands, es robusto para entornos reales, incluso utilizando únicamente una cámara web estándar. Este logro resalta la accesibilidad del sistema, al eliminar la dependencia de hardware especializado, como sensores infrarrojos o cámaras de profundidad, que suelen incrementar los costos y limitar la adopción (Zhang, 2012).

Comparado con estudios previos, el desempeño del sistema es competitivo. Por ejemplo, Chen et al. (2017) reportaron una precisión de ~92% en reconocimiento de gestos usando redes convolucionales (CNN), pero requiriendo cámaras de alta resolución y datasets menos diversos. Asimismo, sistemas basados en Kinect (Zhang, 2012) o Leap Motion (Marin et al., 2014) alcanzaron precisiones cercanas al 95%, pero dependen de hardware dedicado, lo que contrasta con la propuesta actual, que logra resultados similares con una cámara web convencional. La combinación de ViT, conocida por su eficiencia en tareas de visión (Dosovitskiy et al., 2020), y MediaPipe Hands para detección de puntos clave en tiempo real, representa una novedad frente a enfoques tradicionales, al integrar aprendizaje profundo con procesamiento ligero en dispositivos de gama media.

El significado de estos resultados trasciende la precisión técnica, ya que el sistema fomenta una interacción humano-computadora más inclusiva. Su capacidad para operar con una latencia de ~60 ms (>10 FPS) en un CPU Intel i5-10300H lo hace adecuado para aplicaciones en accesibilidad,

domótica y realidad aumentada. La robustez del modelo, atribuida a la diversidad del dataset HaGRID, sugiere que puede generalizarse a entornos reales con variaciones en iluminación, fondo y ángulo, un desafío común en visión por computadora (Wachs et al., 2011).

No obstante, el estudio presenta limitaciones. Primero, la matriz de confusión reveló confusiones entre gestos similares, como *three* y *three2* (~3% de error), debido a su proximidad visual, lo que podría afectar la experiencia de usuario en aplicaciones críticas. Segundo, las pruebas en tiempo real fueron limitadas, lo que restringe la generalización a poblaciones más diversas o condiciones extremas. Tercero, aunque el sistema es eficiente en dispositivos de gama media, no se evaluó en hardware de baja gama, lo que limita su accesibilidad en contextos de recursos restringidos. Finalmente, la aplicación carece de una interfaz gráfica para calibración personalizada, lo que podría dificultar su uso por usuarios no técnicos.

Para futuras investigaciones, se recomienda: (1) aumentar el dataset con gestos similares para reducir confusiones, utilizando técnicas de aumento de datos específicas; (2) ampliar las pruebas en tiempo real con una muestra más grande y diversa, incluyendo condiciones extremas; (3) optimizar el modelo para dispositivos de baja gama, explorando arquitecturas ligeras como MobileViT (Mehta & Rastegari, 2021); (4) desarrollar una interfaz gráfica intuitiva para calibración de gestos; y (5) explorar aplicaciones específicas, como control gestual en videojuegos, educación o entornos hospitalarios, para validar la versatilidad del sistema. Estas mejoras podrían consolidar el potencial del sistema como una solución inclusiva y escalable para la interacción humano-computadora.

Conclusiones

El desarrollo y evaluación del sistema de control por gestos basado en visión por computadora han permitido alcanzar los objetivos establecidos, consolidando una alternativa innovadora y efectiva al mouse físico para la interacción humano-computadora. La integración del modelo Vision Transformer (ViT), ajustado mediante *transfer learning* con el dataset HaGRID, demostró una capacidad sobresaliente para clasificar 18 gestos con una precisión significativamente superior al umbral propuesto del 90%. Este alto desempeño, logrado en un conjunto de prueba extenso y diverso, refleja la robustez del sistema frente a variaciones en iluminación, fondo y ángulo, condiciones comunes en entornos reales. La combinación de ViT con MediaPipe Hands, que

detecta puntos clave de la mano en tiempo real, permitió mapear gestos a acciones del sistema operativo de manera precisa, asegurando una interacción fluida y confiable. Este logro técnico destaca la viabilidad de emplear arquitecturas avanzadas de aprendizaje profundo en tareas de visión por computadora, incluso con recursos computacionales de gama media, como los utilizados en el laboratorio de la Universidad Mayor Real y Pontificia San Francisco Xavier de Chuquisaca.

La implementación de la aplicación, utilizando OpenCV para captura de video y PyAutoGUI para control del sistema operativo, alcanzó una experiencia de usuario fluida, con una latencia baja que garantiza más de 10 cuadros por segundo en dispositivos con procesadores Intel i5. Esta capacidad asegura que el sistema sea práctico para tareas cotidianas, como mover el cursor, realizar clics o arrastres, ofreciendo una alternativa intuitiva a los dispositivos de entrada tradicionales. La evaluación en tiempo real con usuarios bajo condiciones variables de iluminación y distancia confirmó que el sistema mantiene un desempeño robusto, validando su aplicabilidad en entornos dinámicos. Este resultado es particularmente relevante en el contexto de la interacción humano-computadora, ya que demuestra que es posible lograr una interfaz natural de usuario (NUI) sin comprometer la precisión ni la usabilidad, incluso en escenarios donde los periféricos convencionales son poco prácticos.

Un aspecto central del sistema es su accesibilidad, lograda al emplear exclusivamente una cámara web estándar, eliminando la necesidad de hardware especializado como sensores infrarrojos o cámaras de profundidad, que suelen ser costosos y requerir configuraciones específicas. Esta característica posiciona al sistema como una solución inclusiva, especialmente para usuarios con movilidad reducida o en contextos con recursos limitados, donde el acceso a tecnología avanzada puede estar restringido. La capacidad del sistema para operar localmente, sin dependencia de conexión a internet, refuerza su potencial para ser implementado en diversos entornos, desde hogares hasta instituciones educativas o industriales. La robustez del dataset HaGRID, con su amplia diversidad de imágenes capturadas en condiciones reales, fue fundamental para garantizar que el sistema pueda generalizarse a diferentes escenarios, aunque persisten desafíos menores en la distinción de gestos visualmente similares. Este logro subraya el potencial del sistema para aplicaciones prácticas, incluyendo accesibilidad para personas con discapacidades motoras, control de dispositivos en entornos de domótica, y desarrollo de interfaces inmersivas en realidad aumentada. En conjunto, el sistema representa un avance significativo hacia interfaces más

intuitivas y accesibles, contribuyendo al campo de la interacción humano-computadora con una solución eficiente, escalable y alineada con las necesidades de usuarios diversos.

Referencias

- Chen, X., Wang, W., & Li, Y. (2017). *Vision-based hand gesture recognition for human-computer interaction: A survey*. International Journal of Computer Applications, 169(11), 21–29. https://doi.org/10.5120/ijca2017914449
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929
- Marin, G., Dominio, F., & Zanuttigh, P. (2014). *Hand gesture recognition with Leap Motion and Kinect devices. 2014 IEEE International Conference on Image Processing* (ICIP), 1565–1569. https://doi.org/10.1109/ICIP.2014.7025313
- Rautaray, S. S., & Agrawal, A. (2015). *Vision based hand gesture recognition for human computer interaction: A survey. Artificial Intelligence Review*, 43(1), 1–54. https://doi.org/10.1007/s10462-012-9356-9
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90
- Howard, A. G., et al. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. arXiv preprint arXiv:1704.04861. https://arxiv.org/abs/1704.04861
- Norman, D. A. (2010). *The Design of Everyday Things*. Basic Books.
- Mehta, S., & Rastegari, M. (2021). *MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer*. *arXiv preprint arXiv:2110.02178.* https://arxiv.org/abs/2110.02178
- Wachs, J. P., Kölsch, M., Stern, H., & Edan, Y. (2011). *Vision-based hand-gesture applications*. Communications of the ACM, 54(2), 60–71. https://doi.org/10.1145/1897816.1897838

- Wigdor, D., & Wixon, D. (2011). *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann.
- Zhang, C., Tian, Y., & He, Z. (2020). *Hand gesture recognition using vision-based deep learning methods: A review*. Multimedia Tools and Applications, 79(35), 26499–26523. https://doi.org/10.1007/s11042-020-08647-2
- Zhang, Z. (2012). *Microsoft Kinect Sensor and Its Effect*. IEEE MultiMedia, 19(2), 4–10. https://doi.org/10.1109/MMUL.2012.24