

Hand and Gestures Tracking with AI & OpenCV

Riveros Pérez O. L.¹ and Vargas Gutiérrez Y. E.²

¹oriverosp@unal.edu.co

²yvargasgu@unal.edu.co

Abstract Hand Tracking, is a constantly evolving technology that allows detecting and tracking the movement of hands in real time. This capability not only facilitates the control of interfaces, but also opens up a range of possibilities from sign language translation to the creation of digital art through intuitive gestures. In a first approach, by means of segmentation, morphology, convolution and filtering techniques, the ideal way to characterize hand gestures is identified.

Keywords

Hand tracking, Artificial Intelligence, OpenCV, Threshold segmentation, Color space filters.

Introducción

El Hand Tracking, o seguimiento de manos, es una tecnología en constante evolución que permite detectar y rastrear el movimiento de las manos en tiempo real. Esta capacidad no solo facilita el control de interfaces, sino que también abre un abanico de posibilidades que van desde la traducción del lenguaje de señas [1], [2], [3], [4] hasta la creación de arte digital mediante gestos intuitivos [5], [6].

En este contexto, se han explorado diversos enfoques para identificar patrones de manera óptima a partir de las poses de la mano y sus significados. Entre estos enfoques, destaca el uso de la API Mediapipe de Google [1], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] que se ha convertido en una herramienta popular. Mediante esta API, se obtiene un modelo de puntos de referencia que permite estimar con precisión la pose de la mano. Sin embargo, también se han desarrollado otros métodos que emplean técnicas avanzadas de Machine Learning y Deep Learning [2], [3], [15], [16], [17], [18], [19], [20] con el objetivo de estimar directamente los patrones de la mano o incluso de obtener modelos de marcas de referencia de manera similar a la API Mediapipe [21].

Aunque los enfoques previamente mencionados han logrado avances significativos, es crucial tener en cuenta que todos ellos se basan en técnicas avanzadas que, con frecuencia, conllevan desafíos adicionales en cuanto a rendimiento y eficiencia. Esto puede dificultar la adecuada evaluación del problema de extracción de características de la mano. Por esta razón, también se exploran enfoques más simples basados en técnicas de segmentación y morfología para desarrollar modelos de mano más simples y eficaces [22], [23], [24], [25], [26].

El artículo propone un ejercicio académico para implementar y comparar varios enfoques del reconocimiento de la mano, con el objetivo de evaluar su rendimiento computacional, precisión y limitaciones. Se comenzará con una revisión del estado del arte en la siguiente sección, seguida de la presentación de diversas arquitecturas para el reconocimiento de la mano. Posteriormente, se analizarán los resultados obtenidos de cada arquitectura y se proporcionarán conclusiones correspondientes.

Antecedentes

El Hand tracking, como se ha mencionado anteriormente, tiene una amplia gama de aplicaciones, y una de las más notables es la detección del lenguaje de señas en diferentes idiomas. Este desafío particular implica tanto poses estáticas como dinámicas de las manos. En un estudio reciente [1], los autores proponen el uso de la API Mediapipe de Google para extraer características de la imagen. Estas características se someten a procesamientos geométricos

y temporales para clasificar gestos estáticos y dinámicos utilizando algoritmos de Support Vector Machine (SVM). Como resultado de este enfoque, logran una precisión del 97.20% en la detección de gestos estáticos y dinámicos en el lenguaje de señas japonés. Este avance representa un paso significativo hacia la comprensión y aplicación efectiva del lenguaje de señas en entornos tecnológicos.

En otro estudio sobre el reconocimiento del lenguaje de señas [3], los autores proponen un enfoque innovador que combina la segmentación de la imagen con el uso de Redes Neuronales Convolucionales (CNN). En particular, emplearon MobileNetV3 para adaptarlo a dispositivos móviles. Este enfoque les permitió lograr una precisión del 75.38% en la detección de gestos estáticos en el lenguaje de señas indonesio. Este método destaca por su capacidad para ser implementado en dispositivos móviles, lo que lo hace accesible y práctico para aplicaciones cotidianas.

Otro aspecto importante del Hand Tracking es el reconocimiento de gestos estáticos para la interacción con el entorno. En un estudio reciente [20], los autores presentan una Red Neuronal Convolutacional (CNN) llamada RGRNet, diseñada específicamente para imágenes desenfocadas. Este enfoque logra una precisión del 78.2%; no obstante, se destaca que requiere cierta capacidad computacional para su ejecución.

Un artículo reciente [21] destaca un enfoque investigativo innovador en el cual los autores proponen el uso de una Red Neuronal Convolutacional (CNN) para reconocer marcas de referencia de manera similar a la API MediaPipe. Su propuesta incluye el uso de imágenes con información de profundidad y un Detector de Disparo Único de Redes Neuronales Convolucionales (SSD-CNN) para extraer los puntos de referencia de la mano. Posteriormente, emplean un algoritmo de Máquinas de Vectores de Soporte (SVM) para clasificar los gestos. Los resultados obtenidos muestran una precisión del 83.45% en la identificación de gestos.

Los métodos actuales comúnmente emplean algoritmos de machine learning; sin embargo, es útil explorar enfoques que estimen las características de las manos a partir de la geometría y operaciones básicas. En este sentido, el enfoque descrito en el artículo [22] propone la estimación de la pose de la mano utilizando operaciones típicas de procesamiento de imágenes. Este método comienza obteniendo con precisión el contorno de la mano mediante varios filtros. Luego, se realiza un proceso de convolución utilizando una matriz invariante a la rotación, seguido de una umbralización de la imagen. Posteriormente, se aplica una operación de dilatación para obtener el perfil de los dedos de la mano. A partir de esta información, se puede estimar el ángulo de los dedos con respecto a la muñeca. Este enfoque simple proporciona información suficiente para estimar poses básicas de

la mano.

Técnicas

En un primer acercamiento al hand tracking y sus antecedentes, se evidencia la constante discusión de la técnica más adecuada para detectar la piel, por destacar algunos autores [22] utiliza el espacio de color YCrCb, [26] HSV y en [27] se demuestra que es más relevante el número de bins en los que se normaliza el histograma de la imagen. Por ello, antes del desarrollo completo del algoritmo de caracterización de la mano, se han de verificar estas teorías.

- Filtros

Transformaciones a espacios de color desde BGR hacia

- YCrCb, reconocimiento de piel
- HSV, reconocimiento de piel
- GRAY, algoritmos de búsqueda y segmentación por umbrales

Búsqueda de fronteras, contornos y bordes

- Contornos por kernel
- Canny, fronteras internas

Aplicación de máscaras a imágenes originales mediante operación bitWise AND

- Morfología

Eliminación de ruido y aumento de zonas de interés

- Dilatación, aumento la intensidad de las áreas de interés
- Erosión, eliminación de líneas y ruido perjudicial

- Segmentación

En su totalidad estrategias de umbralización

- Binary, reconocimiento de piel
- ToZero, reconocimiento de sombras
- Otsu, búsqueda de umbrales automáticamente

- Convolución

Aplicación de matrices invariantes a la rotación

- Histogramas

- Normalización
- Suavizado del histograma, compresión de máximos locales
- Reducción de bins, reconocimiento de piel

Datos utilizados

Un total de 4 imágenes, donde 1a, 1b y 1c se utilizan para seleccionar el mejor método de reconocimiento de piel.



(a) Mano en interiores



(b) Mano sobre un patrón



(c) Mano en un ambiente con ruido



(d) Gesto en interior

Figura 1. Imágenes estáticas

Una matriz invariante a la rotación, con las siguientes características; Posee 3 radios, donde $r_3 = 3 \cdot r_1$, $r_2 = 2 \cdot r_1$ y $r_1 = N$, es decir.

$$r_1 = r_2 - r_1 = r_3 - r_2 = N$$

Donde $N = 0,5 \cdot d$, y d es la mitad del grosor promedio de los dedos, por lo que N es un cuarto del grosor promedio de los dedos. En la sección de resultados se da a entender el porqué de estas dimensiones.

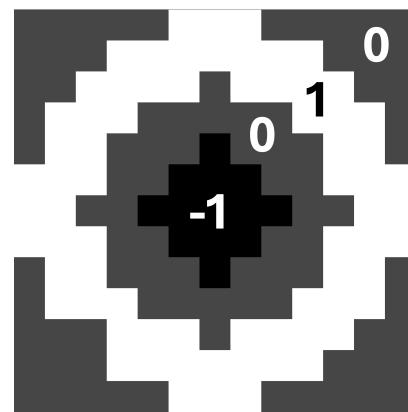


Figura 2. Matriz invariante a la rotación

Procedimiento y resultados

Métodos de reconocimiento de piel

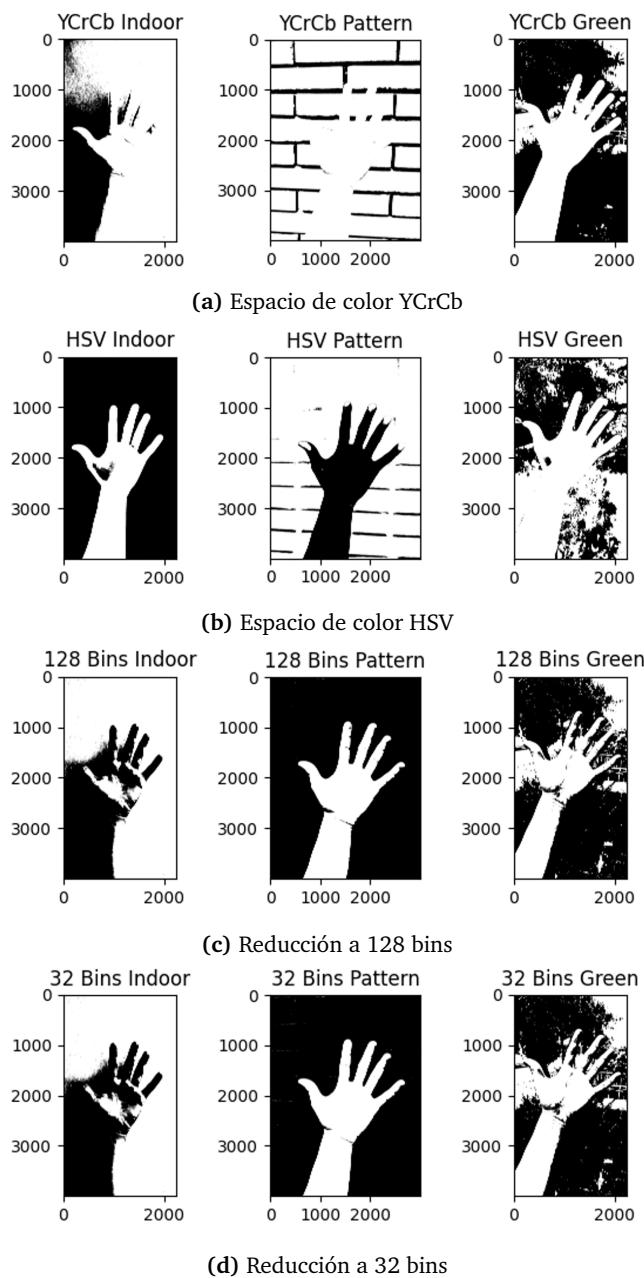


Figura 3. Resultados, métodos de reconocimiento de piel

Aunque hay ventajas en interiores con en el espacio de color HSV y en exteriores con YCrCb con respecto a la reducción en el número de bins, esta última mantiene un balance en todos los casos, como se observa en las Figuras 3c y 3d. Evidenciando lo descrito en [27] en un estudio sobre detección de piel en imágenes. Sin embargo, debido al ambiente de exposición, el espacio de color HSV es la mejor opción para interiores.

Obtención de la silueta completa de la mano

Aplicando la máscara obtenida de la umbralización del espacio de color HSV, se segmenta y aísla la mano Figura 4a.

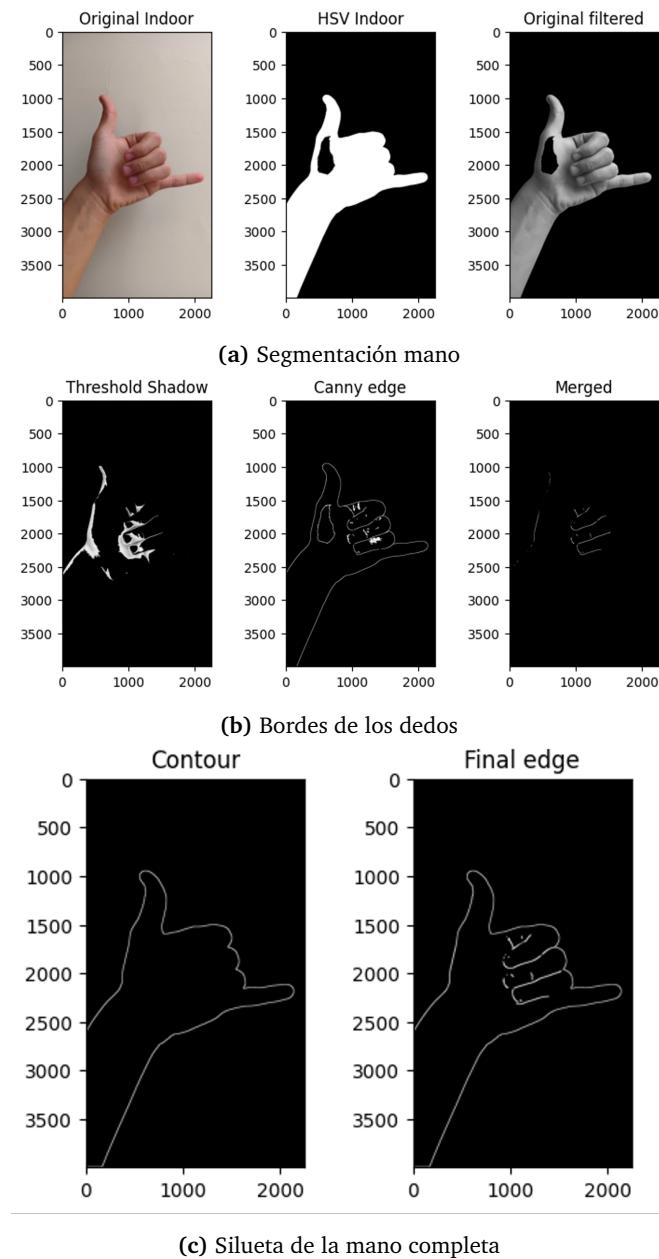


Figura 4. Resultados, silueta de gesto

Para encontrar los bordes asociados a los dedos, en primer lugar se hallan las sombras que estos producen sobre la palma de la mano al estar flexionados, de la intersección de sombras y el filtro Canny para bordes, se obtienen los bordes de los dedos flexionados Figura 4b.

Aplicando los kernes necesarios para la obtención del contorno, mediante la operación bitwise OR se unen el contorno exterior y el interior, correspondiente a los dedos. Obteniendo así la silueta completa de la mano.

Identificación de dedos

En este punto se retoma la matriz invariante a la rotación Figura 2, la cual está diseñada para identificar el área entre cada dedo sin perturbar el resto de la imagen, por ello la distribución de sus radios y los valores asignados a cada área entre ellos, donde solo se representa la sección intermedia, correspondiente al rango $(2N, 3N]$. Es una matriz cuadrada de dimensión $1 + 6 \cdot N$, impar debido a la normalización de los kernes, para su correcto funcionamiento es necesario desarrollar una función que escala la matriz al N asignado, para el caso de estudio $N = 50$, se denota esta matriz como G .

Seguidamente, se hace la convolución entre el kernel almacenado en la matriz G y la silueta de la mano completa, seguida de una erosión para evitar intersección de áreas no relacionadas, se obtiene la primera imagen de la izquierda en la Figura 5.

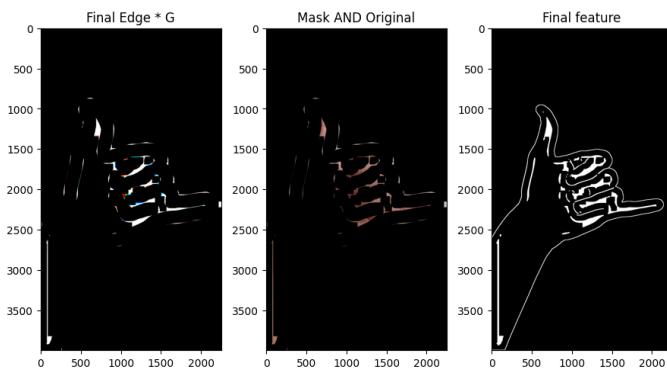


Figura 5. Identificación de los dedos en la pose

Este resultado se utiliza como máscara para la imagen original mediante la operación bitwise AND. Al ser la imagen original y conociendo el umbral correspondiente a la piel en el espacio de color HSV, para este caso de estudio $Thres = 88$ con un máximo de 255 (Este umbral se obtuvo mediante la técnica Otsu, en los métodos de reconocimiento de piel), las líneas aún visibles al exterior son eliminadas, ya que no hacen parte de la mano.

Finalmente, se dilata la segmentación obtenida para hacer más notable las secciones pertenecientes a los dedos, y este último se une mediante la operación bitwise OR, para obtener la caracterización final del gesto.

Con esta última imagen hay vía libre para hallar la orientación de los dedos respecto a la muñeca y lograr así estimar poses en tiempo real.

Conclusiones

- Ante el planteamiento de cualquier problema de caracterización de objetos es necesario hacer uso de una gran cantidad de estrategias y técnicas para lograr los resultados idóneos, como sucede en este caso de estudio.
- En un principio, la teoría expuesta en [27] acerca del reconocimiento de la piel en imágenes puede prestarse para interpretaciones erróneas, si bien el desempeño de la reducción de bins es sobresaliente, en casos de estudio donde no es relevante obtener toda la información de la piel, como en hand tracking donde solo se necesita la silueta del objeto, este desempeño es innecesario. Pasando por alto las ventajas que tienen los espacios de color, YCrCb tiene un desempeño excelente en ambientes con mucho ruido y HSV de igual manera en interiores.
- Los kernes son una parte esencial del procesamiento de imágenes y visión por computadora, ya que permiten realizar filtros y transformaciones en las imágenes, para resaltar características específicas, mediante la operación de convolución. Operación por la cual es posible la identificación de los dedos en gestos con manos.
- Con los resultados actuales, para un segundo avance en el proyecto, se espera la caracterización en vivo de gestos, partiendo de la orientación de los dedos respecto a la muñeca. Avance que puede desembocar en estrategias más avanzadas de inteligencia artificial.

Referencias

- [1] M. Kakizaki, A.S.M. Miah, K. Hirooka, and J. Shin. Dynamic japanese sign language recognition throw hand pose estimation using effective feature extraction and classification approach. *Sensors*, 24, 2024. doi: 10.3390/s24030826.
- [2] H. Alsolai, L. Alsolai, F.N. Al-Wesabi, M. Othman, M. Rizwanullah, and A.A. Abdelmageed. Automated sign language detection and classification using reptile search algorithm with hybrid deep learning. *Heliyon*, 10, 2024. doi: 10.1016/j.heliyon.2023.e23252.
- [3] R. Sutjiadi. Android-based application for real-time indonesian sign language recognition using convolutional neural network. *TEM Journal*, 12:1541–1549, 2023. doi: 10.18421/TEM123-35.
- [4] B.A. Dabwan, M.E. Jadhav, H.A. Abosaq, F.A. Olayah, M. Al Yami, and Y.A. Abdelrahman. Real-time system for translating american sign language to text using robust techniques. 2023. ISBN 9798350306927. doi: 10.1109/ICRASET59632.2023.10420110.
- [5] T. Watanabe, M. Maniruzzaman, M.A.M. Hasan, H.-S. Lee, S.-W. Jang, and J. Shin. 2d camera-based air-writing recognition using hand pose estimation and hybrid deep learning

- model. *Electronics (Switzerland)*, 12, 2023. doi: 10.3390/electronics12040995.
- [6] P. Shukla and P. Das. Enhancing human-computer interaction: Hand detection for air writing utilizing numpy and opencv. pages 517–521, 2023. ISBN 9798350342338. doi: 10.1109/ICTACS59847.2023.10390179.
- [7] S.K. Baruah, U. Konwar, R. Mahanta, A. Boruah, and D. Sarma. A contactless control mechanism for computerized systems using hand gestures. pages 246–250, 2023. ISBN 9798350325881. doi: 10.1109/ASPCON59071.2023.10396173.
- [8] G. Csonka, M. Khalid, H. Rafiq, and Y. Ali. Ai-based hand gesture recognition through camera on robot. pages 256–261, 2023. ISBN 9798350395785. doi: 10.1109/FIT60620.2023.00054.
- [9] S.S. Sugantha Mallika, M. Priyadharsini, S. Samritha, C. Sowmiya, and B. Nikitha. Hand gesture recognition using convolutional neural networks. pages 249–255, 2023. ISBN 9798350340235. doi: 10.1109/ICACRS58579.2023.10404885.
- [10] Q. Wang and Z. Xie. Arias: An ar-based interactive advertising system. *PLoS ONE*, 18, 2023. doi: 10.1371/journal.pone.0285838.
- [11] E. Aksoy, A.D. Çakir, B.A. Erol, and A. Gumus. Real time computer vision based robotic arm controller with ros and gazebo simulation environment. 2023. ISBN 9798350360493. doi: 10.1109/ELECO60389.2023.10416078.
- [12] N. Kumar, H. Dalal, A. Ojha, A. Verma, and M. Kaur. Real-time hand gesture recognition for device control: An opencv-based approach to shape-based element identification and interaction. pages 1537–1541, 2023. ISBN 9798350342338. doi: 10.1109/ICTACS59847.2023.10390298.
- [13] D. Wu, J. Huang, M. Zheng, and Y. Li. Virtual model interaction based on single rgb camera. pages 298–301, 2023. ISBN 9798350380859. doi: 10.1109/NTCI60157.2023.10403685.
- [14] V.L. Adluri, P. Kadiyala, S. Gopu, and S. Jangiti. Virtual mouse with hand gestures using machine learning. pages 1564–1568, 2023. ISBN 9798350313987. doi: 10.1109/ICSCNA58489.2023.10370381.
- [15] N. Zerrouki, F. Harrou, A. Houacine, R. Bouarroudj, M.Y. Cheffifi, A.A. Zouina, and Y. Sun. Deep learning for hand gesture recognition in virtual museum using wearable vision sensors*. *IEEE Sensors Journal*, 2024. doi: 10.1109/JSEN.2024.3354784.
- [16] A. Das, K. Maitra, S. Roy, B. Ganguly, M. Sengupta, and S. Biswas. Development of a real time vision-based hand gesture recognition system for human-computer interaction. pages 294–299, 2023. ISBN 9798350325881. doi: 10.1109/ASPCON59071.2023.10396583.
- [17] A. Yildiz, N.G. Adar, and A. Mert. Convolutional neural network based hand gesture recognition in sophisticated background for humanoid robot control. *International Arab Journal of Information Technology*, 20:368–375, 2023. doi: 10.34028/ijait/20/3/9.
- [18] R. Özakar and E. Gedikli. Evaluation of hand washing procedure using vision-based frame level and spatio-temporal level data models. *Electronics (Switzerland)*, 12, 2023. doi: 10.3390/electronics12092024.
- [19] S. Karegoudra and R.K. Veerasha. Hand gestures and machine learning: A novel path to number prediction. 2023. ISBN 9798350306927. doi: 10.1109/ICRASET59632.2023.10419929.
- [20] G. Huang, S.N. Tran, Q. Bai, and J. Alty. Real-time automated detection of older adults' hand gestures in home and clinical settings. *Neural Computing and Applications*, 35:8143–8156, 2023. doi: 10.1007/s00521-022-08090-8.
- [21] F.A. Farid, N. Hashim, J.B. Abdullah, M.R. Bhuiyan, M. Kairanbay, Z. Yusoff, H.A. Karim, S. Mansor, M.D.T. Sarker, and G. Ramasamy. Single shot detector cnn and deep dilated masks for vision-based hand gesture recognition from video sequences. *IEEE Access*, 12:28564–28574, 2024. doi: 10.1109/ACCESS.2024.3360857.
- [22] Y. Zhou, G. Jiang, and Y. Lin. A novel finger and hand pose estimation technique for real-time hand gesture recognition. *Pattern Recognition*, 49:102–114, 2016. doi: 10.1016/j.patcog.2015.07.014.
- [23] R. Zahra, A. Shehzadi, M.I. Sharif, A. Karim, S. Azam, F. De Boer, M. Jonkman, and M. Mehmood. Camera-based interactive wall display using hand gesture recognition. *Intelligent Systems with Applications*, 19, 2023. doi: 10.1016/j.iswa.2023.200262.
- [24] H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. *IEEE Transactions on Multimedia*, 16:1241–1253, 2014. doi: 10.1109/TMM.2014.2306177.
- [25] R.M. Gurav and P.K. Kadbe. Real time finger tracking and contour detection for gesture recognition using opencv. pages 974–977, 2015. ISBN 9781479971657. doi: 10.1109/IIC.2015.7150886.
- [26] Z.-H. Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan. Real-time hand gesture recognition using finger segmentation. *Scientific World Journal*, 2014, 2014. doi: 10.1155/2014/267872.
- [27] M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 274–280 Vol. 1, 1999. doi: 10.1109/CVPR.1999.786951.