

Real-Time Gesture Recognition with Virtual Glove Markers

Finlay McKinnon


School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

Pedro Machado 

School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom
pedro.machado@ntu.ac.uk

David Ada Adama 

School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom
david.adama@ntu.ac.uk

Isibor Kennedy Ihianle 

School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom
isibor.ihianle@ntu.ac.uk




ABSTRACT

Due to the universal non-verbal natural communication approach that allows for effective communication between humans, gesture recognition technology has been steadily developing over the previous few decades. Many different strategies have been presented in research articles based on gesture recognition to try to create an effective system to send non-verbal natural communication information to computers, using both physical sensors and computer vision. Hyper accurate real-time systems, on the other hand, have only recently begun to occupy the study field, with each adopting a range of methodologies due to past limits such as usability, cost, speed, and accuracy. A real-time computer vision-based human-computer interaction tool for gesture recognition applications that acts as a natural user interface is proposed. Virtual glove markers on users hands will be created and used as input to a deep learning model for the real-time recognition of gestures. The results obtained show that the proposed system would be effective in real-time applications including social interaction through telepresence and rehabilitation.

KEYWORDS

Hand Gesture Recognition, Glove Markers, Computer Vision, Hand Rehabilitation

ACM Reference Format:

Finlay McKinnon, David Ada Adama , Pedro Machado , and Isibor Kennedy Ihianle . 2022. Real-Time Gesture Recognition with Virtual Glove Markers. In *The 15th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '22)*, June 29–July 1, 2022, Corfu, Greece. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3529190.3534749>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PETRA '22, June 29–July 1, 2022, Corfu, Greece

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9631-8/22/06...\$15.00

<https://doi.org/10.1145/3529190.3534749>

1 INTRODUCTION

Hand gesture recognition is an important area of development and a focus point for Human-Computer Interaction (HCI) that can be applied to a variety of areas, including sign language recognition, virtual and augmented reality, assisted living technology, and industrial use [1, 6, 16, 17].

Gesture recognition may be viewed as a promising research topic when discussing HCI and computer vision, due to the innate properties of gestures, being natural and decisive forms of communication. In HCI research, it is important to consider the amount of usability friction that the human has when interacting with the computer. A device with many physical components would have a considerably high usability friction, as layers of complexity are being added to the interaction from the user's perspective. The use of alternative approaches are desired for less complexity. For example, in applications involving gesture recognition, the use of a vision-based device would have an incredibly low usability friction, as the movements are natural to the user, and there are no additional layers.

Over the years, gesture HCI has been improved upon to a great amount, starting with unsatisfactory results using bulky physical sensors [20], to accurate recognition using everyday objects [13]. However, the development of optical recognition technology solutions has been slow, due to inconsistencies in the nature of optical sensing and how it is affected by background and environmental lighting, occlusions, processing times against image resolution and frame rates all of which makes gesture recognition performance sub-optimal [9].

With regard to the wide range of hand gesture recognition applications, the technology has been applied to many domains. Hand gesture recognition can be used for a safer driving experience to minimise driver distraction [8], aiding the visually impaired via smart-watch gesture recognition [13], and construction worker safety training [1]. Such technology has also been suggested for highly precise situations, as being used for augmented reality Graphical User Interfaces (GUI) dance in robotic surgery [18], as well as social interactions with its suggested use being included in an autonomous telepresence robot for remote conferencing [4]. With these approaches, it can be understood that the technology holds promise in a variety of fields, aiding in vastly different areas with multiple differing methods to yield beneficial results.

According to Donchysts et al. [5], in order to truly capitalise on the extent of gesture recognition, the user’s experience should be as fluid as possible, which can be achieved with low usability friction that computer-vision based gesture recognition can offer as capable through natural user interfaces, which traditional means of interaction cannot. Traditional methods of HCI, command lines, GUI, keyboards and mice are all inconvenient and unnatural [9], and may even be unusable if the user has some form of physical impairment.

A form of human computer interaction which is a Natural User Interface (NUI) through a real-time computer vision-based hand gesture recognition system is proposed. It takes advantage of the benefits of hand gesture recognition as a form of HCI over the traditional methods highlighted above. The research uses a means of image processing to extract regions of a user’s hand which are used as virtual glove markers for detecting specific points of the hand.

The remainder of the paper is structured as follows; Section 2 presents a review of related work. In Section 3, the methodology of the proposed real-time gesture recognition with virtual glove markers system is discussed. Section 4 describes the experimental setup and results obtained. Section 5 concludes the paper and gives directions for future work.

2 RELATED WORK

Research into gesture recognition has been an important research area due to its wide-ranging areas of application, some of which includes solutions to minimise distractions and dangers while driving [8] to comfort and efficiency [14]. A brief background research about gesture recognition is presented in this section.

Recent advances have shown the different approaches to hand gesture recognition. Stergiopolou and Papamarkos [15] proposed colour segmentation and a neural gas neural network to identify finger positioning and gestures utilising histograms. Shangchen et al. [6] proposed a robust system utilising a neural network for hand detection accompanied by hand key-point location estimations. Yin and Xie [19] suggests applying an RCE neural network for hand segmentation, 2D and 3D feature extraction to reconstruct hand movements and gestures for HCI. Feature extraction based on colour segmentation, depth information, or deep learning, as well as feature segmentation, key point estimation, and histograms to estimate hand position and orientation, are frequently proposed approaches for developing a vision-based hand gesture recognition system; however, these are only a few of the many proposed methods. Recognition based on skin colour, appearance based on background subtraction, motion-based, depth-based, and deep-learning based, as well as a variety of additional techniques are covered in [11].

More recently, in 2020, driver distraction was responsible for roughly 10% of motor vehicle deaths [10], though with the research proposed by Molchanov [8], driver distraction could be lowered through changing the HCI method within cars to a gesture based system. Furthermore, research done by Reifinger [14] shows that gesture based HCI can be more intuitive, more comfortable, and up to 60% quicker than the aforementioned traditional forms of HCI.

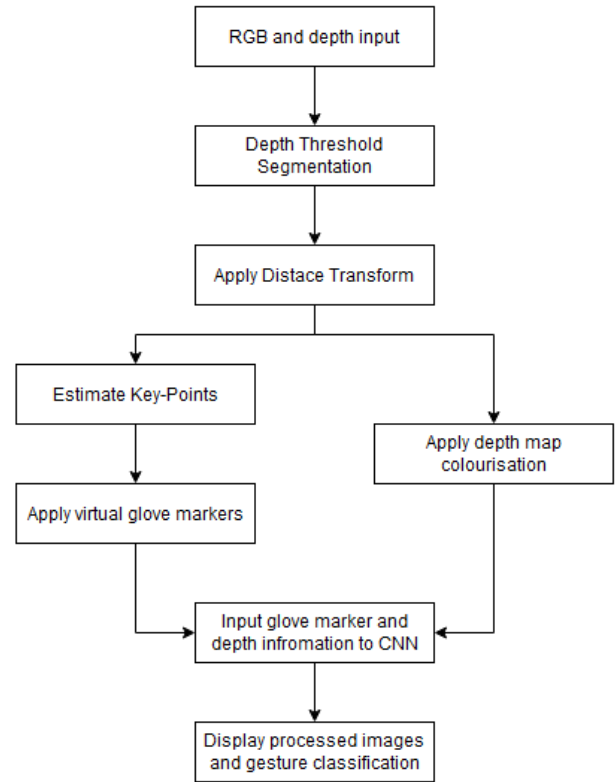


Figure 1: Overview of real-time gesture recognition system using virtual glove markers.

The cost of physical sensors, complexity of gesture recognition and hand tracking algorithms, presents challenges and limitations to the accessibility of the aforementioned forms NUI to users; however, with the approach proposed in this paper, it is possible to create relatively accurate recognition with hardware cheap and accessible to potential users.

The approach utilising a virtual glove was inspired by the classical physical glove marker approach demonstrated by [17]. The simplicity of the core idea, utilising colour markers to indicate sections of a hand and machine learning methods to approximate the hand’s positioning. This approach adapts the improved machine learning algorithm proposed by Wang et al. [17] which allows for efficient and accurate algorithms to assist in applying a virtual glove to the user’s hand in real time.

3 METHODOLOGY

The aim of this paper is to provide insight into an approach to creating an NUI through the use of computer vision, specifically utilising depth information, segmentation and a Convolutional Neural Network (CNN). This will be accomplished by the creation of virtual glove markers, as opposed to the standard physical markers as proposed by Wang et al. [14, 17]. The research provides a mixture of techniques not commonly used together to attempt to create a system in which static hand gestures may be accurately understood by the system as a new promising form of HCI in the form of a NUI

using glove markers. Glove markers allow a vision-based tracking system to extract the orientation and position of the hand. However, the method requires physical gloves to be worn as a medium for the system to work. An alternative method that includes replacing the physical glove marker technique into virtual glove markers via segmenting the hand after key-point estimation is proposed in this article.

An overview of the methodology for the real-time gesture recognition system using virtual glove markers presented in this paper is given in Figure 1. The key steps in the methodology are summarised as follows:

Red, Green, Blue and depth (RGBd) input: Firstly, the research makes use of RGBd information of the hand, obtained using a depth sensor. An example of both Red, Green and Blue (RGB) and depth input frames can be seen in Figures 2(a) and 2(b).

Depth Thresholding: in the system is the act of segmenting the Region of Interest (RoI), being the user's hand in this case, from the rest of an image. This will be utilised as form of reducing the computational workload when the image is passed to the recognition model, as there is less nonessential data in each image to be processed. From the collected data, the RGB frame will undergo a depth-threshold transformation via the depth data obtained in the respective depth frame. This depth-threshold transformation will effectively remove all RGB data from a frame if the information is past a given distance, being the threshold. Figure 2(b) shows an example of image after depth thresholding has been applied. The threshold chosen for the research is 500mm from the camera, used in an environment where there are no objects closer than 500mm other than the user's hand.

Distance Transform: the RGB image processed in the RoI segmentation stage, is converted to binary image by a transformation of the RGB image matrix, converting the image to a 2D binary image, with a value of 0 representing background and value 1 representing the hand as shown in Figure 3(a). A distance transform is then applied to the binary image of the hand, being used to calculate a representation of the most central point in the hand. This is possible by calculating the relative distance of each hand pixel element from a background pixel element as shown in Figure 3(b).

Key-Point (Finger Point) Estimation: involves the use of a Machine Learning (ML) algorithm known as MediaPipe¹ to estimate the fingertips and knuckles in the given frame. MediaPipe creates key points on the hand known as landmarks, these landmarks including the fingertips and knuckle points, allow the system to efficiently estimate their location. MediaPipe has been used in this system as a tool for proof of concept, due to computational constraints. The fingertip key point coordinates of the frame and then saved for later use.

Centre Palm Point Estimation: as the machine learning algorithm utilised for finger key-points does not calculate a central palm point, the system utilises the distance transform to calculate the most central point of the hand. This has been done through searching for the highest value in the distance transform, representing the pixel furthest from a background pixel, which on a hand is generally the central point of the palm. The radius of the

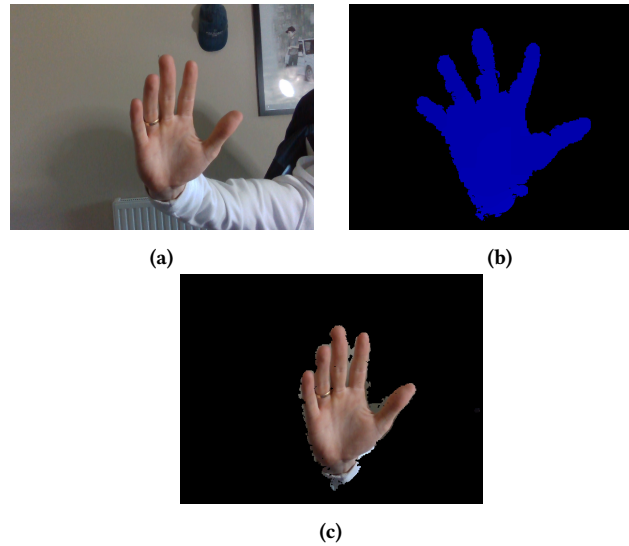


Figure 2: Depth threshold filter applied to an RGB image. (a) RGB image input. (b) Colourised depth input from sensor. (c) ROI segmentation through depth threshold.

palm can be calculated by the value of the pixel element, as the value increments by one per pixel, this will give an accurate relative distance to the edges of the palm. This estimated palm area is then visually displayed on the virtual glove, using the central point and calculated radius of the palm. A moving average, of the previous 5 frames, is used to smooth the jitter.

The virtual glove is then drawn onto the hand, atop the processed image with the estimated palm drawn to it. The proposed glove has been applied through plotting lines using the Open Computer Vision (OpenCV) library² to points gathered from the key-point estimation process, creating palm to knuckle then knuckle to fingertip links as shown in Figure 4. The processed image displaying each portion of the hand in a distinct colour will act as the virtual glove to be used in the model for gesture categorisation.

Gesture Categorisation using CNN: the categorisation of each gesture shall be done by passing each frame to a CNN, categorising each frame based on previous training of specific images of predefined gestures. The CNN will be trained on both the depth information and virtual glove, the data being given categories to recognise different gestures as.

4 EXPERIMENTAL RESULTS

The experiments conducted to validate the proposed approach to real-time gesture recognition using virtual glove markers utilises an Intel RealSense D435i depth camera³. Using this device, both RGB and depth information are obtained. Five different gestures are used for the experiments conducted in this paper. These gestures can be seen in Figure 5 and comprise of a 'one finger', 'thumbs-up', 'Ok', 'two fingers' and 'shaka' gestures.

¹Available online, <https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html>, last accessed: 21/03/2022

²Available online, <https://opencv.org/>, last accessed 21/03/2022

³Available online, <https://store.intelrealsense.com/buy-intel-realsense-depth-camera-d435i.html>, last accessed: 21/03/2022



Figure 3: Distance transform applied to segmented binary image. (a) Binary conversion of the RoI segmented image. (b) Binary image after distance transform.

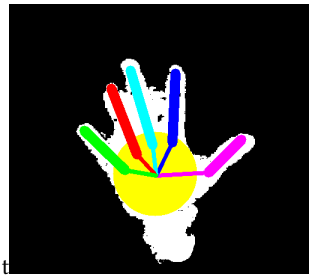


Figure 4: Processed hand depth image with virtual glove marker.

Data was collected for all five gestures. A total of 1500 frames were collected which corresponds to 300 frames for each gesture. This data was then pre-processed and fed into the CNN model following the methodology discussed in Section 3. Table 1 shows the performance of the proposed system on the dataset. This shows the correctly classified gestures, the total number of each gesture and percentage accuracy.

Information in the accuracy table displays the overall accuracy of the current model’s recognition ability, scoring an overall accuracy of 95%. The accuracy of each gesture show some inconsistency, this possibly being due to overfitting of the CNN in some aspects; this can be viewed through the difference in accuracy between gestures such as ‘OK’, scoring 99.3% accuracy, and ‘Thumb’, scoring only 90% accuracy.

The proposed system’s average processing speed clocks in at 135ms, which is significantly less than the average human’s ability to detect visual stimuli at between 180ms and 200ms [7]. The results presented in Table 1 yielding total of 95% real time accuracy shows the potential of the proposed system for use in varying applications. This supporting the aim to create a system that is easier for users to interact with, being a NUI and innately being a more natural form of HCI, based on the results.

To evaluate the system proposed in this paper, a comparison is made with other similar works. This comparison is reported in Table 2. The accuracy of the proposed system demonstrates the functionality of a real-time gesture recognition, which research in this field often lag behind; moreover, statistical comparisons between the proposed system and other systems of similar design proposing non-real time testing results can be made. The system

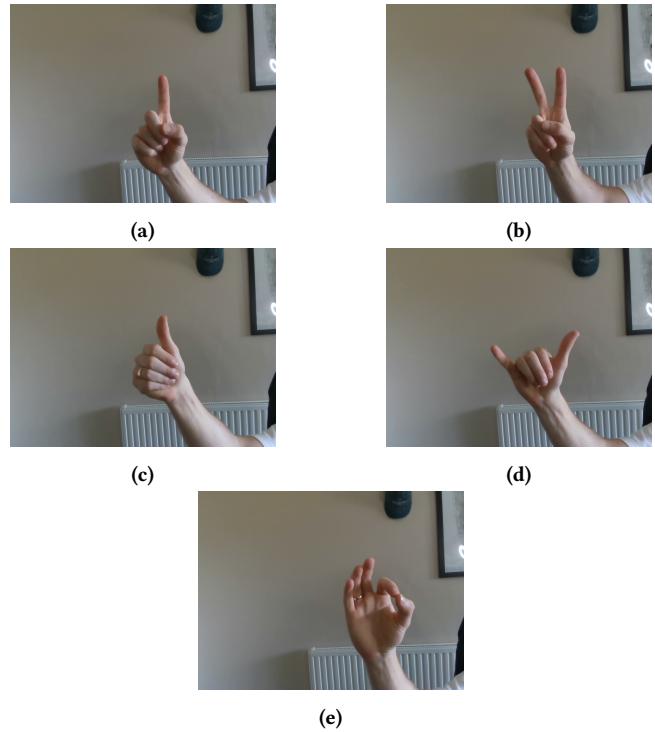


Figure 5: RGB image of recognised gestures used in this paper. (a) ‘One Finger’ Gesture. (b) ‘Two Fingers’ Gesture. (c) ‘Thumb’ Gesture. (d) ‘Shaka’ Gesture. (e) ‘OK’ Gesture.

proposed in [3] presents a similar system to the work in this paper, achieving a 99.9% accuracy in training and a 95% accuracy in testing, which is very similar to the results obtained in this paper of 98.1% and 95% respectively. However, the proposed system allows for multiple more use cases in dynamic and cluttered environments, as displayed by the accuracy of a more similar system proposed in [2] scoring 85% in dynamic background testing.

Also, the work in [12] proposed another system similar to what has been proposed in this paper, utilising estimated skeletal points to infer human gestures. The research reported information about real-time complex background testing, scoring a 90% accuracy within these tests which is significantly lower than the results achieved in this paper.

5 CONCLUSION AND FUTURE WORK

This paper proposed a system which is relatively unique in its design, utilising virtual glove markers and RGBd data fed to a CNN model for real-time gesture recognition. As stated in the introduction, this process has yielded promising results for the computer vision gesture recognition field, though future could be done to improve on the systems’ robustness to more gestures.

The proposed system could be improved to make the approach more robust and offset the limitations of the current system. Firstly, there are limitations within the proposed system, being that the depth-threshold algorithm is not entirely accurate, capturing some

Table 1: Real-time accuracy results.

-	One Finger	Two Finger	Thumb	Shaka	OK	Total
Correct	290	271	270	296	298	1425
Attempted	300	300	300	300	300	1500
Accuracy	96.67%	90.34%	90.0%	93.3%	99.3%	95%

Table 2: Comparison of the proposed system with other related works. BG - Background

-	Training	Validation	Static BG	Dynamic BG
Proposed System	98.09%	98.42%	95%	95%
Parelli [12]	-	-	90%	90%
Chung [3]	99.9%	98.1%	-	-
Bao [2]	-	-	85%	85%

background around the hand. This may be improved by adding filters to the algorithm to lessen the unwanted sections of the RoI captured. Furthermore, lighting must be considered with the system as the key-point estimation ML algorithm utilises feature detection, the hand must be well enough illuminated to allow the algorithm to accurately find these points.

Due to processing speeds the user may not have immediate recognition of their gesture and must hold the gesture for about 135ms, this may also mean that gestures done quickly may be missed.

REFERENCES

- [1] Srikanth Sagar Bangaru, Chao Wang, Xu Zhou, Hyun Jeon, and Yulong Li. 2020. Gesture Recognition-Based Smart Training Assistant System for Construction Worker Earplug-Wearing Training. *Journal of Construction Engineering and Management* 146 (12 2020), 04020144. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001941](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001941)
- [2] Peijun Bao, Ana I. Maqueda, Carlos R. del Blanco, and N. García. 2017. Tiny hand gesture recognition without localization via a deep convolutional network. *IEEE Transactions on Consumer Electronics* 63 (2017), 251–257.
- [3] Hung-Yuan Chung, Yao-Liang Chung, and Wei-Feng Tsai. 2019. An Efficient Hand Gesture Recognition System Based on Deep CNN. In *2019 IEEE International Conference on Industrial Technology (ICIT)*. 853–858. <https://doi.org/10.1109/ICIT.2019.8755038>
- [4] Ha M. Do, Craig J. Mouser, Ye Gu, Weihua Sheng, Sam Honarvar, and Tingting Chen. 2013. An open platform telepresence robot with natural human interface. In *2013 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems*. 81–86. <https://doi.org/10.1109/CYBER.2013.6705424>
- [5] Gennadiy Donchyts, Fedor Baart, Arthur Dam, and Bert Jagers. 2014. Benefits of the use of natural user interfaces in water simulations.
- [6] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. 2020. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality. *ACM Trans. Graph.* 39, 4, Article 87, 13 pages. <https://doi.org/10.1145/3386569.3392452>
- [7] Aditya Jain, Ramta Bansal, Avnish Kumar, and K Singh. 2015. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research* 5, 122. <https://doi.org/10.4103/2229-516X.157168>
- [8] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. 2015. Multi-sensor system for driver's hand-gesture recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–8. <https://doi.org/10.1109/FG.2015.7163132>
- [9] G. R. S. Murthy and R. S. Jadon. 2009. A Review of Vision Based Hand Gestures Recognition. *International Journal. Of Information Technology and Knowledge* 2, 2, 405–410.
- [10] National Highway Traffic Safety Administration (NHTSA). 2020. Traffic safety facts," Tech. Rep.
- [11] Munir Oudah, Ali Al-Naji, and Javaan Chahl. 2020. Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging* 6, 8. <https://www.mdpi.com/2313-433X/6/8/73>
- [12] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. 2020. Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 249–263.
- [13] Lorenzo Porzi, Stefano Messelodi, Carla Modena, and Elisa Ricci. 2013. A smart watch-based gesture recognition system for assisting people with visual impairments. 19–24. <https://doi.org/10.1145/2505483.2505487>
- [14] Stefan Reifinger, Frank Wallhoff, Markus Ablassmeier, Tony Poitschke, and Gerhard Rigoll. 2007. Static and Dynamic Hand-Gesture Recognition for Augmented Reality Applications. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 728–737.
- [15] E. Stergiopoulou and N. Papamarkos. 2006. A New Technique for Hand Gesture Recognition. In *2006 International Conference on Image Processing*. 2657–2660. <https://doi.org/10.1109/ICIP.2006.313056>
- [16] Kollipara Varun, I. Puneeth, and Prem Jacob. 2019. Hand Gesture Recognition and Implementation for Disables using CNN'S. 0592–0595. <https://doi.org/10.1109/ICCCSP.2019.8697980>
- [17] Robert Wang and Jovan Popovic. 2009. Real-time hand-tracking with a color glove. *ACM Trans. Graph.* 28. <https://doi.org/10.1145/1576246.1531369>
- [18] Rong Wen, Liangjing Yang, Chee-Kong Chui, Kah-Bin Lim, and Sha Chang. 2010. Intraoperative Visual Guidance and Control Interface for Augmented Reality Robotic Surgery. In *2010 8th IEEE International Conference on Control and Automation, ICCA 2010*. 947 – 952.
- [19] Xiaoming Yin and Ming Xie. 2007. Hand Posture Segmentation, Recognition and Application for Human-Robot Interaction. <https://doi.org/10.5772/6097>
- [20] Thomas Zimmerman, Jaron Lanier, Chuck Blanchard, Steve Bryson, and Young Harvill. 1986. A hand gesture interface device. *ACM Sigchi Bulletin* 17, 189–192. <https://doi.org/10.1145/30851.275628>