




A Comparative Analysis of Methods for Hand Pose Detection in 3D Environments

Jorge G. Iglesias¹ ^a, Luis Montesinos^{1,2} ^b and David Balderas^{1,2} ^c

¹*School of Engineering and Sciences, Tecnológico de Monterrey, Mexico City, Mexico*

²*Institute of Advanced Materials for Sustainable Manufacturing, Tecnológico de Monterrey, Mexico City, Mexico*
{A01653261, lmontesinos, dc.balderassilva}@tec.mx

Keywords: Virtual Reality, Depth Maps, Stereoscopic Vision, Human-Computer Interaction.


Abstract: The ability to discern the pose and gesture of the human hand is of big importance in the field of human-computer interaction, particularly in the context of sign language interpretation, gesture-based control and augmented reality applications. Some models employ different methodologies to estimate the position of the hand. However, few have provided a comprehensive and objective comparison, resulting in a limited understanding of the approaches among researchers. The present study assesses the efficacy of three-dimensional (3D) hand pose estimation techniques, with a particular focus on those that derive the hand pose directly from depth maps or stereo images. The evaluation of the models considers endpoint pixel error as a principal metric for comparison between methods, with the aim of identifying the most effective approach. The objective is to identify a method that is suitable for virtual reality training considering memory usage, speed, accuracy, adaptability, and robustness. Furthermore, this study can help other researchers understand the construction of such models and develop their own models.


1 INTRODUCTION


Hand pose detection is the process of identifying and tracking the positions and movements of a hand and its individual fingers (Buran Basha et al., 2020). Computer vision algorithms and deep learning models can be used to achieve it (Buran Basha et al., 2020). This process involves extracting features from video frames or images of hands and using regression algorithms to identify specific gestures or behaviors. This technology has diverse applications, particularly in human-computer interaction for extended reality devices, where precise hand tracking is essential, as well as gesture recognition and accurate hand tracking (Zhang et al., 2023; Haji Mohd et al., 2023). The application in education has been found to check how students move their hands during laboratory procedures (Liu et al., 2023). Moreover, within the health-care sector, hand pose detection models contribute to the automatic identification of abnormal hand gestures, facilitating the early diagnosis and treatment of nerve injuries (Gu et al., 2022).

The detection of hand pose faces various challenges. Specifically, reliable detection among clutter and occlusions remains problematic (Alinezhad Noghre et al., 2023), compounded by variations in poses and background clutter (Haji Mohd et al., 2023). The current two-dimensional models for the detection of hand pose have been greatly developed in terms of the precision of the position and pose of each finger (Liu et al., 2023), gesture classification (Guan et al., 2023), and improvement in physical procedures involving hands (Zhang et al., 2023), among others. However, they lack applicability in 3D contexts, despite the potential for testing from its 2D approaches. Additionally, even if there are existing open source solutions (see section 3), each approach possesses unique qualities that may be absent in others, resulting in incomplete solutions as a whole. In addition, variability in the construction of the hand skeleton among different models adds a layer of complexity, as there is no standard format that defines the structure uniformly.

The objective of this study is to identify an appropriate three-dimensional hand estimation technique for virtual environments by comparing two stereoscopic models. One obtains depth maps, and the other generates stereo images. This was done to compare

^a  <https://orcid.org/0009-0008-4042-8251>

^b  <https://orcid.org/0000-0003-3976-4190>

^c  <https://orcid.org/0000-0001-7630-8608>

the two methods, with the aim of determining which is the most effective for locating the hand, determining its position, and analysing its pose. The study considers metrics such as the error distance being Endpoint Pixel Error (EPE) which compares the error distance between the predicted and real landmark coordinates. EPE metric is used because this is a standard keypoint detection metric, which has been used mainly in the evaluation of hand estimation (Chatzis et al., 2020; Sharma and Huang, 2021). This paper aims to serve as a guide for other researchers in selecting the best-suited methods for solving problems related to hand pose detection.

Sections 2 and 3 provide background and related work on hand pose detection, comprehensively analysing the existing models, databases, and metrics used. Section 4 describes the methods, Section 5 its results and discussion, and Section 6 the limitations encountered. Finally, sections 7 and 8 present future work and conclusions, respectively.

2 BACKGROUND

This section provides a context for understanding the research. The first subsection 2.1 focusses on the construction of hand poses and then section 2.2 shows the theory behind the metrics used to evaluate the models.

2.1 Hand Landmarks Construction

There are several ways of constructing virtual skeleton hands to identify their anatomical landmarks; which makes it difficult to make hand pose models since the joint landmarks vary. Many models are constructed using the skeletal joints of the hand, starting at the wrist or palm and then going through all the joints of the fingers. Each finger has 3 joints: the distal interphalangeal joint (DIP), the proximal interphalangeal joint (PIP) and the metacarpophalangeal joint (MCP joint). The thumb has different joint names: the interphalangeal joint (IP), the metacarpophalangeal joint (MCP), and the carpometacarpal joint (CMC) (American Society for Surgery of the Hand, 2024). From there, the hand can be easily constructed for computer algorithms, also by adding the tip for each finger to visualise it completely (called TIP, of course). An example of this is how the Media Pipe model is constructed, as shown in Figure 1. There are 21 knuckle coordinates in the captured hand regions (google, 2024).

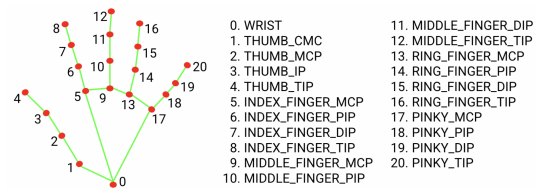


Figure 1: The 21 hand-knuckle coordinates (google, 2024).

2.2 Metrics

When evaluating 3D hand pose estimation methods, common metrics include the Endpoint Pixel Error (EPE) and the Percentage of Correct Keypoints (PCK). The EPE, as shown in Equation 1, is the average Euclidean distance between the predicted and reference joints, while the PCK measures the mean percentage of predicted joint locations that fall within certain error thresholds compared to correct poses. In short, the mean Euclidean distance between landmarks and predictions is used to calculate the EPE, while the PCK considers a prediction correct if the EPE between a pair of landmarks and its prediction is within a given threshold. To properly assess the performance of the model and compare different models, it is important to plot the PCK over different thresholds and calculate the Area Under the Curve (AUC) (Chatzis et al., 2020; Sharma and Huang, 2021).

$$EPE = \sqrt{\sum_{n=0}^k (g_n - p_n)^2} \quad (1)$$

3 RELATED WORK

The review of related work discusses previous research in hand pose models, highlighting trends and common approaches. Section 3.1 explores various methods, particularly Convolutional Neural Networks (CNNs), for detecting hand poses in three dimensions (3D). CNNs have proven to be effective in learning the features of images to accurately estimate hand positions. Additionally, Section 3.2 examines alternative models such as Generative Adversarial Networks (GANs) and PointNet, offering novel perspectives for 3D hand pose detection. While CNNs focus on visual feature learning, GANs can generate realistic 3D hand images, and PointNet excels at processing point clouds, crucial for scenarios requiring 3D information. Furthermore, the review discusses the specialised 3D hand pose databases in Section 3.3, which have improved model training and evaluation, leading to improved accuracy and generalisation in 3D hand pose detection systems.

3.1 Convolutional Neural Networks

The work of Malik and collaborators introduces WHSP-Net, a novel weakly supervised learning method to extract three-dimensional hand shape and pose from single depth images (Malik et al., 2019). Their approach integrates a CNN for joint position generation (in other words, the hand skeleton is created and positioned in accordance with the desired specifications), a shape decoder for dense mesh reconstruction, and a depth synthesiser for image reconstruction, learning from both real and synthetic data to compensate for the lack of ground truth. The score obtained achieves a mean 3D joint position error of 9.24 mm. In contrast, Sharma et al. (Sharma and Huang, 2021) propose an end-to-end framework for unconstrained 3D hand pose estimation from monocular RGB images. Their ConvNet model predicts hand information and infers pose solely from keypoint annotations, incorporating biological constraints to enhance accuracy. In particular, on the STB dataset, the model achieves a Mean Endpoint Pixel Error of 8.71 mm. Despite focussing solely on the area of the hand, both approaches enhance accuracy and address distinct challenges in 3D hand pose estimation, highlighting the diverse strategies employed in computer vision. However, they do not explicitly consider the spatial positioning of the hand relative to the camera surroundings, a notable limitation for virtual reality applications.

3.2 Other Models

He et al.'s study, proposing a technique to generate depth hand images based on ground-truth 3D hand poses using generative adversarial networks (GAN) and image-style transfer techniques, achieves mean joint errors of 8.41 mm and 6.45 mm on the MSRA and ICVL datasets, respectively (He et al., 2019). However, Wu et al.'s Capsule-HandNet, an end-to-end capsule-based network for estimating hand poses from 3D data, demonstrates slightly higher mean joint errors of 8.85 and 7.49 mm on the same datasets (Wu et al., 2020). Despite these nuanced differences, both approaches exhibit significant potential for enhancing hand pose estimation methodologies, each highlighting distinct strengths in modelling and inference techniques.

3.3 3D Hand Pose Databases

The study by Gomez-Donoso et al. (Gomez-Donoso et al., 2019) proposes the creation of a new multiview dataset for hand pose due to the shortcomings present

in existing datasets. These datasets are characterised by limited samples, inaccurate data or higher-level annotations, and a focus mainly on depth-based approaches, which lacked RGB data. The set comprises colour images of the hand and annotations for each sample, including the bounding box and the 2D and 3D location of the joints. In addition, a deep learning architecture was introduced for real-time estimation of the 2D pose of the hand. Experiments demonstrated the dataset's effectiveness in producing accurate results for 2D hand pose estimation using a single colour camera. The study made a significant contribution by providing a large-scale dataset of more than 26,500 annotations. This dataset includes colour frames from four different viewpoints, detailed annotations of 3D and 2D joint positions, and bounding boxes.

The study by (Chatzis et al., 2020) provides a review of various databases used in hand posture research and hand posture tracking. The datasets mentioned include ICVL, NYU Hand Pose Dataset, Big-Hand2.2M, MSRA15, Handnet, HANDS 2017, Syn-Hand5M, FreiHand, RHD, STB, EgoDexter, Dexter+Object, Dexter1, and SynthHands. Each dataset varies in size, resolution, number of subjects, and specific characteristics of the annotations provided. Furthermore, the text highlights the strengths and weaknesses of each dataset, including image quality, annotation accuracy, and specific challenges for hand tracking research.

4 MATERIALS AND METHODS

Two approaches are proposed as common techniques to detect depth in three-dimensional environments by using two cameras:

1. Depth map approach: In this method, the model directly analyses the depth map to detect the hand and estimate its 3D pose landmarks.
2. Stereo image approach: In this approach, the model analyses the 2D image to detect the hand and estimate its 2D pose landmarks in the two stereo images. Subsequently, the 2D landmark coordinates are merged into 3D coordinates using the depth information provided by both cameras.

The dataset created by Zhang et al. (Zhang et al., 2016) is used. This data set is selected for its compromise between size and robustness, as well as its inclusion of stereo images and depth maps. To process the data, both stereo images and depth maps are resized, transformed to greyscale, and normalised. The landmark coordinates present in the two matrix files (one

for the stereo images and another for the depth maps) are transformed into a data frame and reorganised according to their intended purpose, with each three columns representing a point. Each row represents the landmarks of the detected hand in a stereo picture or a depth map. See Figure 3 as a simplified visual representation of the methodology of this paper. All other specifications such as the cameras' properties, distances, units are written in Zhang's project.

Both methods are evaluated using the Endpoint Pixel Error (EPE) metric to provide a standardised measure for performance evaluation. Convolutional Neural Networks (CNNs) are selected as the main architecture for both methodologies because of their effectiveness in handling image processing tasks. For the depth image approach, the architecture shown in Figure 2 is used. In addition, the Mediapipe solution for the estimation of bidimensional hand pose is used as the basis for the second approach (google, 2024).

5 RESULTS AND DISCUSSION

The two approaches were successfully implemented, yielding intriguing results. In both methods, the predicted hand is correctly located according to the test data, with some minor details from the predicted finger position almost aligned with the test data, as expected. In Figures 4 and 5, the plot represents the position of the three-dimensional space of the hand, with the X, Y, and Z axes serving as a reference in millimetres (mm); the view is adjusted to focus solely on the position of the hand. As illustrated in Figures 4 and 5, the predicted data exhibit a high degree of alignment with the test data, with only minor discrepancies. The EPE scores indicate that the stereo image approach demonstrated a slight advantage over the depth map approach, with scores of 22.89 and 29.72 mm, respectively. The mean processing times for each had minimal differences, yet the stereo image approach exhibited a marginal advantage over the depth map approach, with respective scores of 2.3507 and 2.7012 ms.

Meanwhile, the depth map approach was trained for this study using the previous architecture depicted in Figure 2, the stereo image approach was built using the Mediapipe solution, indicating that the stereo image model is better trained than the depth map model. Nevertheless, during certain predictions, the Mediapipe solution encountered difficulties in detecting the hand because the environment being too dark. Consequently, it yielded deformed hands to no hands detected in some instances. The hand deformation can also happen when the cameras are not well synchro-

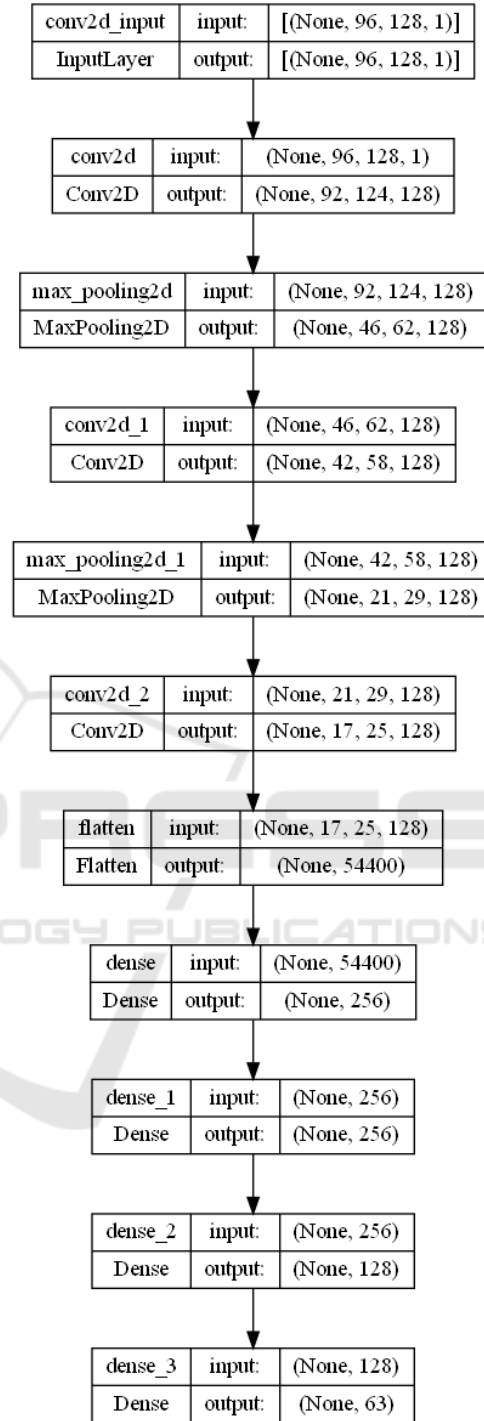


Figure 2: Architecture used for model CNN in the depth map approach.

nised when the image capture is made, noise in the model and ambient, actual model precision on the hand detection, among others. However, both techniques simply processes these inputs and, as a consequence, produces outputs that reflect these limita-

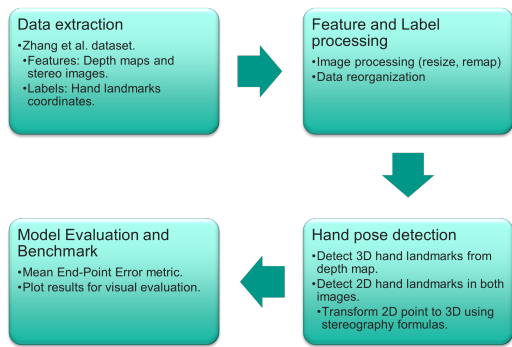


Figure 3: Methodology process.

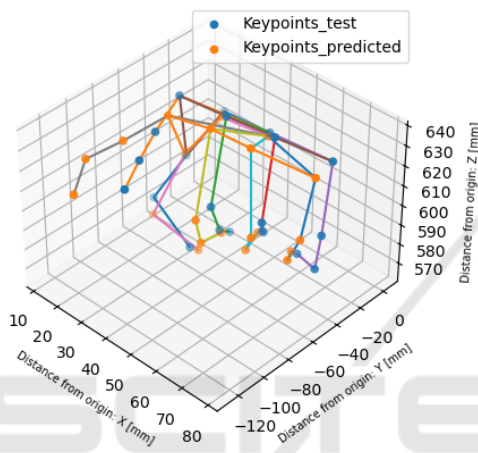


Figure 4: Predicted and test data comparison for the depth map approach.

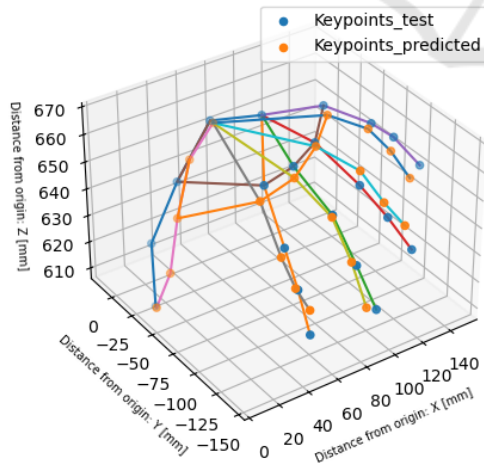


Figure 5: Predicted and test data comparison for the stereo image approach.

tions.

It is evident that even if both models were unable to achieve a higher precision compared to the results obtained in the models mentioned in Section 3, the hand itself would still maintain its structural

integrity and the position of the hand would remain accurate. It is also noteworthy that the approach utilising stereo images requires the model to be called twice, whereas the depth-map approach necessitates the processing of stereo images in order to generate the depth map, which consequently results in a slower processing time. Additionally, such processing may not yield optimal results, as environmental and camera variables may present challenges in processing the data, thereby increasing the time required for both calibration and accuracy. In contrast, the stereo images approach requires only the utilisation of the images themselves, with the detected coordinates undergoing a stereoscopic formula to calculate the position of the point in the three-dimensional plane. The stereo image model can be used as a rapid estimator for hand landmark annotation, including the incorporation of corrections to ensure that the resulting estimates are of an acceptable standard.

6 LIMITATIONS

The lack of processing power of the computer systems available at the time impeded the efficient execution of complex algorithms required for this task. Furthermore, the limited time available for training detection models constituted an additional challenge, as the accuracy of these models was highly dependent on the quantity and quality of the training data. The reduced resolution of the input images resulted in a lower level of hand analysis, which in turn affected the accuracy of the detection.

7 FUTURE WORK

Continued progress in hand pose detection for virtual environments necessitates a multifaceted approach that addresses several key challenges. In this context, research is proposed that focusses on three main areas with the objective of improving the effectiveness of existing systems. Firstly, the architecture of convolutional neural networks can be optimised through the introduction of innovative techniques, such as the addition of new and useful layers or the exploration of deeper structures. This will improve accuracy and computational efficiency. Secondly, a model of pose and hand gesture correction is proposed in order to enhance the precision of the representation of finger structure. It is proposed that the angles and proportions of the virtual skeleton of the hand need to be fixed via algorithmic or with artificial intelligence methods, as these are the primary issues encountered

in this study. Fixing these parameters can enhance the accuracy of the model. Third, in addition to the ongoing development and refinement of a hand detection model, consideration should be given to the integration of this model into virtual training, as previously outlined in the main objective of this study. In other words, the effectiveness of the model must be evaluated to determine whether the trainees are able to adapt to it. One potential methodology for evaluating the efficacy of the aforementioned approach is to devise a series of scenarios in which the trainee is required to position themselves and perform gestures in a manner that allows for the assessment of their adaptation to the hand pose detection model.

8 CONCLUSIONS

A comparative analysis of two distinct methodologies for hand pose estimation, one based on depth maps and the other on stereo images, has yielded significant insights into the relative strengths and limitations of each approach. Although neither approach yielded a near-optimal solution, both demonstrated effectiveness in accurately capturing the spatial position of the hand and constructing viable hand representations. These results suggest the potential for substantial improvements in accuracy, robustness, and adaptability through further refinement and optimisation of existing techniques. Consequently, continued research and development in this area could lead to more advanced solutions for applications such as virtual reality training and gesture-based control systems.

ACKNOWLEDGEMENTS

The authors thank Tecnológico de Monterrey for financial support to produce this work.

REFERENCES

- Alinezhad Noghre, G., Danesh Pazho, A., Katariya, V., and Tabkhi, H. (2023). Understanding the Challenges and Opportunities of Pose-based Anomaly Detection. In *Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, pages 1–9, Lübeck Germany. ACM.
- American Society for Surgery of the Hand (2024). Joints. Publication Title: Body Anatomy: Upper Extremity Joints | The Hand Society.
- Buran Basha, M., Ravi Teja, S., Pavan Kumar, K., and Anudeep, M. (2020). Hand poses detection using convolutional neural network. *International Journal of Scientific & Technology Research*, 9(1):1887–1891.
- Chatzis, T., Stergioulas, A., Konstantinidis, D., Dimitropoulos, K., and Daras, P. (2020). A Comprehensive Study on Deep Learning-Based 3D Hand Pose Estimation Methods. *Applied Sciences*, 10(19):6850.
- Gomez-Donoso, F., Orts-Escolano, S., and Cazorla, M. (2019). Large-scale multiview 3D hand pose dataset. *Image and Vision Computing*, 81:25–33.
- google (2024). GitHub - google/mediapipe: Cross-platform, customizable ML solutions for live and streaming media. Publication Title: GitHub.
- Gu, F., Fan, J., Cai, C., Wang, Z., Liu, X., Yang, J., and Zhu, Q. (2022). Automatic detection of abnormal hand gestures in patients with radial, ulnar, or median nerve injury using hand pose estimation. *Frontiers in Neurology*, 13:1052505.
- Guan, X., Shen, H., Nyatega, C. O., and Li, Q. (2023). Repeated Cross-Scale Structure-Induced Feature Fusion Network for 2D Hand Pose Estimation. *Entropy*, 25(5):724.
- Haji Mohd, M. N., Mohd Asaari, M. S., Lay Ping, O., and Rosdi, B. A. (2023). Vision-Based Hand Detection and Tracking Using Fusion of Kernelized Correlation Filter and Single-Shot Detection. *Applied Sciences*, 13(13):7433.
- He, W., Xie, Z., Li, Y., Wang, X., and Cai, W. (2019). Synthesizing Depth Hand Images with GANs and Style Transfer for Hand Pose Estimation. *Sensors*, 19(13):2919.
- Liu, S., Yuan, X., Feng, W., Ren, A., Hu, Z., Ming, Z., Zahid, A., Abbasi, Q., and Wang, S. (2023). A Novel Heteromorphic Ensemble Algorithm for Hand Pose Recognition. *Symmetry*, 15(3):769.
- Malik, J., Elhayek, A., and Stricker, D. (2019). WHSP-Net: A Weakly-Supervised Approach for 3D Hand Shape and Pose Recovery from a Single Depth Image. *Sensors*, 19(17):3784.
- Sharma, S. and Huang, S. (2021). An end-to-end framework for unconstrained monocular 3D hand pose estimation. *Pattern Recognition*, 115:107892.
- Wu, Y., Ma, S., Zhang, D., and Sun, J. (2020). 3D Capsule Hand Pose Estimation Network Based on Structural Relationship Information. *Symmetry*, 12(10):1636.
- Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., and Yang, Q. (2016). 3D Hand Pose Tracking and Estimation Using Stereo Matching. eprint: 1610.07214.
- Zhang, Q., Lin, Y., Lin, Y., and Rusinkiewicz, S. (2023). Hand Pose Estimation with Mems-Ultrasonic Sensors. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, Sydney NSW Australia. ACM.