Hand Gesture Recognition Using MediaPipe Landmarks and Deep Learning Networks

Manuel Gil-Martín[1], Marco Raoul Marini[2], Iván Martín-Fernández[1], Sergio Esteban-Romero[1] and Luigi Cinque[2]

[1]*Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM), Av. Complutense 30, 28040, Madrid, Spain*

[2]*VisionLab, Department of Computer Science, Sapienza University of Rome, Via Salaria 113, Rome 00198, Italy*

Keywords: Hand Gesture Recognition, Human-Computer Interaction, MediaPipe Landmarks, Deep Learning.

Abstract: Advanced Human Computer Interaction techniques are commonly used in multiple application areas, from entertainment to rehabilitation. In this context, this paper proposes a framework to recognize hand gestures using a limited number of landmarks from the video images. This hand gesture recognition system comprises an image processing module that extracts and processes the coordinates of 21 hand points called landmarks, and a deep neural network module that models and classifies the hand gestures. These landmarks are extracted automatically through MediaPipe software. The experiments were carried out over the IPN Hand dataset in an independent-user scenario using a Subject-Wise Cross Validation. They cover the use of different landmark-based formats, normalizations, lengths of the gesture representations, and number of landmarks used as inputs. The system obtains significantly better accuracy when using the raw coordinates of the 21 landmarks through 125 timesteps and a light Recurrent Neural Network architecture (80.56 ± 1.19 %) or the hand anthropometric measures (82.20 ± 1.15 %) compared to using the speed of the hand landmarks through the gesture (72.93 ± 1.34 %). The proposed framework studied the effect of different landmark-based normalizations over the raw coordinates, obtaining an accuracy of 83.67 ± 1.12 % when using as reference the wrist landmark from each frame, and an accuracy of 84.66 ± 1.09 % when using as reference the wrist landmark from the first video frame of the current gesture. In addition, the proposed solution provided high recognition performance even when only using the coordinates from 6 (82.15 ± 1.16 %) or 4 (81.46 ± 1.17 %) specific hand landmarks using as reference the wrist landmark from the first video frame of the current gesture.

1 INTRODUCTION

Hand gesture recognition consists in detecting the movement that people perform using their hands. This technology could be useful to develop human computer interaction systems and could improve the user experience across a wide variety of domains. For example, it could be seen as the basis for sign language understanding and hand gesture control applications. For instance, a person could ask to take a picture using the front camera of a smartphone by opening and closing the hand palm. In these applications, it is crucial to accurately recognize the hand gesture to perform specific actions with smart devices, a computer, or an automatic transmission machine.

Multiple previous works have been focused on human activity recognition to optimize the physical activity classification using wearables or cameras (Gil-Martin, San-Segundo, Fernandez-Martinez, & Ferreiros-Lopez, 2020, 2021; Gil-Martín, San-Segundo, Fernández-Martínez, & de Córdoba, 2020; Zhang et al., 2017). However, there exists a lower number of works focused on detecting hand poses or gestures. Most of these works used images as inputs of their systems and follow a hand localization step as the first stage. Afterwards, they extracted handcrafted features or descriptors (Trindade, Lobo, & Barreto, 2012) from the hand and fed them to an inference algorithm that classifies the different hand poses or gestures. For example, a previous work (Mantecon, del-Blanco, Jaureguizar, & Garcia, 2019)

Gil-Martín, M., Marini, M. R., Martín-Fernández, I., Esteban-Romero, S. and Cinque, L.
Hand Gesture Recognition Using MediaPipe Landmarks and Deep Learning Networks.
DOI: 10.5220/0013053500003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 24-30
ISBN: 978-989-758-737-5; ISSN: 2184-433X

segmented the image into the hand in different regions and obtained the Histogram of Oriented Gradients (HOG) and a Local Binary Pattern (LBP) from each region. Afterwards, they combined k-means and Support Vector Machines (SVM) to classify the hand poses, obtaining an F1-score near 96% using data from 25 subjects and 16 different hand poses. Similarly a previous work (Bao, Maqueda, del-Blanco, & Garcia, 2017) fed a deep Convolutional Neural Network (CNN) to directly classify hand poses in images without any previous segmentation. They classified the hand pose with average accuracy of 97.1% in the images with simple backgrounds and 85.3% in the images with complex backgrounds. They used a dataset with data from 40 subjects and seven different hand poses. Another work (Gil-Martín, San-Segundo, & de Córdoba, 2023) used a normalization over the hand landmarks to detect the same poses, achieving robust performance even when the images had complex backgrounds. This approach has the advantage of sending less information to the recognition module compared to traditional computer vision approaches (landmark coordinates vs. a full image). Another previous work (Benitez-Garcia, Olivares-Mercado, Sanchez-Perez, Yanai, & Ieee Comp, 2021) used the raw images to classify hand gestures via training a ResNeXt-101 model, achieving a 86.32% of accuracy when evaluating 13 subjects.

This paper is focused on exploring the impact of different landmark-based input formats over the gesture recognition task, rather than optimizing the deep learning architecture for obtaining the maximum performance. This work uses a state-of-the-art deep learning architecture to understand which input formats—specifically raw hand landmark coordinates, speed of coordinate movement, and anthropometric measures—yield the most informative representations for gesture recognition. The primary contributions of this research are as follows:

- Analyze different landmark input formats: raw coordinates, speed of coordinates movement, and anthropometric measures.
- Investigate the effects of various normalization techniques on raw landmark coordinates.
- Minimize the number of landmarks used in the recognition process while keeping a high recognition performance.

2 MATERIAL AND METHODS

This section describes the dataset used, the landmarks information extraction, and the proposed model architecture.

2.1 Dataset Description

We have used the public dataset called IPN Hand to evaluate our system.

The IPN Hand dataset (Benitez-Garcia et al., 2021) includes hand gestures performed by 50 subjects for interaction with touchless screens. It contains 4,218 gesture instances and 800,000 frames. It includes 13 gestures performed with one hand: pointing with one or two fingers, clicking with one or two fingers, throwing up/down/left/right, opening twice, double-clicking with one or two fingers, zooming in, and zooming out. During data collection, each subject did different gestures with three random breaks in a single video. The subjects used their own PC or laptop to collect the RGB videos, which were recorded in the resolution of 640x480 with the frame rate of 30 fps.

This dataset is a great choice for hand gesture recognition due to its extensive and varied content regarding instances, gestures and subjects, ensuring a diverse representation of hand motions. This diversity enhances the dataset's applicability for real-world human-machine interaction applications, such as touchless screens and virtual reality interfaces, where accurate gesture recognition is crucial. Its realistic data collection, with subjects using their own devices to record gestures in varied environments, ensures that models trained on the dataset can generalize well to practical usage scenarios.

2.2 Landmarks-Based Representations

The original images from the dataset were processed by the MediaPipe library to extract the x and y coordinates from specific points of the hand called landmarks. MediaPipe (Lugaresi et al., 2019; Quinonez, Lizarraga, & Aguayo, 2022) is a powerful library with the capacity to track pose and hands from input frames or video streams. This framework can extract 21 landmarks from the hand (including wrist and four points along the five fingers). To standardize the input data and ensure consistent processing, we applied zero padding at the beginning of each gesture sequence, thereby aligning all examples to the same length, ranging from 25 to 250 timesteps. This padding ensures that a uniform

length, although it becomes dominant in shorter gestures.

In our gesture recognition pipeline, we leverage three distinct types of landmark information formats to encapsulate various aspects of hand movements and spatial relationships. Firstly, we utilize the raw coordinates of landmarks extracted from the hand, providing direct spatial information about the hand's configuration in each frame of the gesture sequence. Secondly, we incorporate the speed derived from the computed derivatives of these coordinates, capturing the temporal dynamics and velocity of hand movements throughout the gesture. Lastly, we integrate anthropometric measures (Pheasant & Haslegrave, 2006) obtained by computing the Euclidean distances between pairs of landmarks, enabling us to encode spatial relationships in the hand gestures and structural characteristics inherent in the hand physiognomy.

For the first approach of landmark information (raw landmark coordinates), we applied two normalization techniques at the landmark level to consider hand translation. The first normalization method involved using the wrist landmark from each frame of the gesture sequence as a reference point. This process entailed subtracting the coordinates of the wrist landmark from those of all other landmarks within the same frame, aligning them relative to the position of the wrist. The second normalization technique employed the wrist landmark from the first frame of the gesture sequence as a constant reference point throughout. Regardless of subsequent frames, the coordinates of all landmarks were adjusted relative to the wrist landmark from the initial frame, ensuring uniformity and consistency in spatial representations across the entire gesture sequence. The formulas related to these normalizations are described in Equation (1) and Equation (2), respectively, where 0 landmark corresponds to the wrist, i ranges from 1 to 20 (for the other landmarks) and t refers to a specific frame in the gesture sequence. These normalization approaches could contribute to reduce variability and enhance the interpretability of landmark coordinates for the modeling and recognition architecture. We also evaluated the possibility of performing a scaling normalization, but no improvement was obtained. A possible reason for this is that the distance between the subjects and their laptops was relatively consistent across the dataset.

$$x'_{i,t} = x_{i,t} - x_{0,t} \quad (1)$$

$$y'_{i,t} = y_{i,t} - y_{0,t}$$

$$x'_{i,t} = x_{i,t} - x_{0,0} \quad (2)$$

$$y'_{i,t} = y_{i,t} - y_{0,0}$$

2.3 Model Architecture

The deep learning architecture used in this work learns the evolution pattern from landmark-based inputs using a Long Short-Term Memory (LSTM) layer of 100 neurons and classifies the examples using a Dense layer of 13 neurons, corresponding to the number of classes. The input shape of the architecture was bidimensional, including the timesteps of the gesture (from 25 to 250) and the number of channels (for example, 42 when using x and y coordinates from the 21 landmarks). The architecture included a dropout layer (0.3) after the LSTM layer to avoid overfitting during training. The last layer used a softmax activation function to offer the predictions of each class for every analysis gesture. We used categorical cross-entropy as loss metric and the root-mean-square propagation method as optimizer (Weiss, 2017). We adjusted the epochs and batch size of the deep learning structure: 50 and 25, respectively. Since the objective of this work is not optimizing the deep learning architecture, its configuration is a sample of a state-of-the-art RNN useful for modeling pattern sequence (Goodfellow, 2016), which is the case of gesture recognition task.

3 RESULTS AND DISCUSSION

To evaluate the system using the whole dataset in a subject-independent scenario, we followed a Subject-Wise Cross-Validation (SW-CV) alternative as data distribution. In this 10-fold CV methodology, the given data are divided into 10 groups or folds to train and test a system with different data subsets in such a way that the data from one subject is only contained in one subset. This process is repeated by changing the training and testing folds and the results are the average of the partial results obtained for all repetitions. This methodology simulates a realistic scenario where the system is evaluated with recordings from subjects different to those used for training.

As evaluation metrics, we used accuracy, which is defined as the ratio between the number of correctly classified samples and the number of total samples. This way, for a classification problem with

N testing examples and C classes, accuracy is defined in Equation (3).

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{C} P_{ii} \quad (3)$$

To show statistical significance values, we used confidence intervals, which include plausible values for a specific metric. We will assure that there exists a significant difference between results of two experiments when their confidence intervals do not overlap. Equation (4) represents the computation of confidence intervals attached to a specific metric value and N samples for 95% confidence level.

$$CI(95\%) = \pm 1.96\sqrt{\frac{\text{metric} \cdot (100 - \text{metric})}{N}} \quad (4)$$

Regarding the experiments, we firstly analyzed the effect of different landmark-based input formats: raw landmark coordinates, landmarks speed and anthropometric measures. In these experiments we used the 21 available landmarks and different lengths for each gesture (25, 50, 75, 100, 125, 150, 175, 200, 225 and 250 timesteps). Figure 1 shows the evolution of accuracy considering the length of the gestures and the different landmark-based input formats.

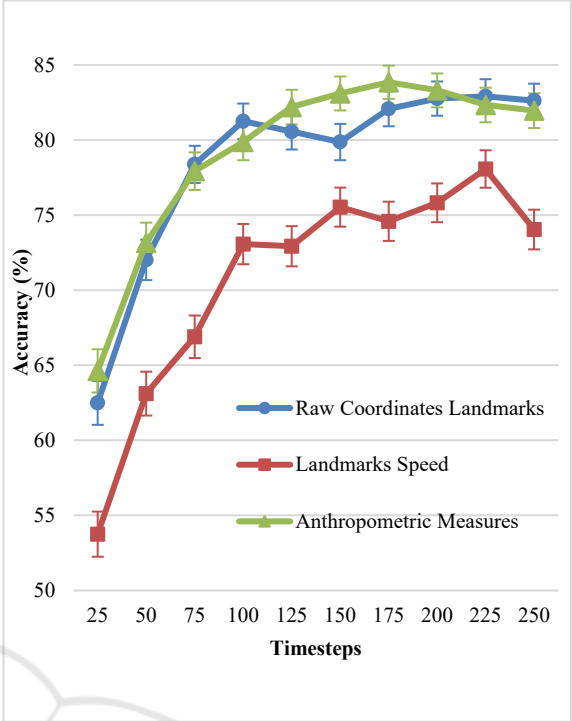

Figure 1: Accuracy depending on the number of timesteps used per gesture and the landmark format.

This figure shows how a saturation of performance is achieved when increasing the length of the gestures, reaching a competitive performance for each landmark representation at 125 timesteps, which is close to the mean of duration of the gestures in the dataset (140 timesteps). For example, at this gesture length, the system obtains an accuracy of 80.56 ± 1.19 % when using the raw coordinates, 72.93 ± 1.34 % when using their speed and 82.20 ± 1.15 % when using the hand anthropometric measures. As observed in the figure, when it comes to modeling gestures, finding the right balance in the amount of real data is crucial. Using too few points fails to capture the nuances of gestures adequately, potentially cutting off crucial information or having too much padding. Conversely, padding sequences unnecessarily lengthens them, posing challenges for recurrent layers to process efficiently.

Results also suggest that the raw landmark coordinates and the anthropometric measures offer significantly better performance compared to the landmark speed along the different timestep configurations.

Raw coordinates serve as fundamental building blocks for hand gesture recognition, offering direct spatial information about the positions of hand landmarks. Anthropometric measures derived from hand landmarks provide valuable insights into the structural characteristics and spatial relationships within gestures. This way, relative positions of hand key points could be useful to learn features based on the spatial arrangement of landmarks, facilitating the discrimination of complex gestures with subtle variations. However, landmark speed might not provide enough contextual information for understanding the hand gestures, because features such as direction, acceleration, and spatial relations between landmarks could be lost in this representation, which may struggle to accurately distinguish the hand gestures.

Second, we performed some experiments applying different landmark-based normalization over the raw coordinates. Figure 2 shows the evolution of accuracy considering the length of the gestures and the landmark normalization used.

This figure shows a tendency of performance improvement when applying a landmark normalization over the raw coordinates, which becomes significant when using 125 or 150 timesteps. For example, at 125 timesteps length, the accuracy of 80.56 ± 1.19 % obtained with the raw landmark coordinates is significantly improved until 83.67 ± 1.12 % when normalizing through the wrist landmark from the current frame, and until 84.66 ± 1.09 % when using as reference the wrist landmark

from the first video frame of each gesture. Normalizing the coordinates using the wrist landmark as reference removes the variability in the inputs among users. One of the reasons for this improvement is that thanks to these types of normalizations, the representation of examples of the same gesture become similar independently of the location of the hand through the images. For example, the representation of a hand gesture consisting in clicking with one finger could differ when the person performs the gesture at the right or left side of the image. However, thanks to normalizing using the wrist landmark, both representations become standardized since both use the coordinate origin as reference. Table 1 shows a summary of the results from the previous experiments using 125-timestep gesture length.

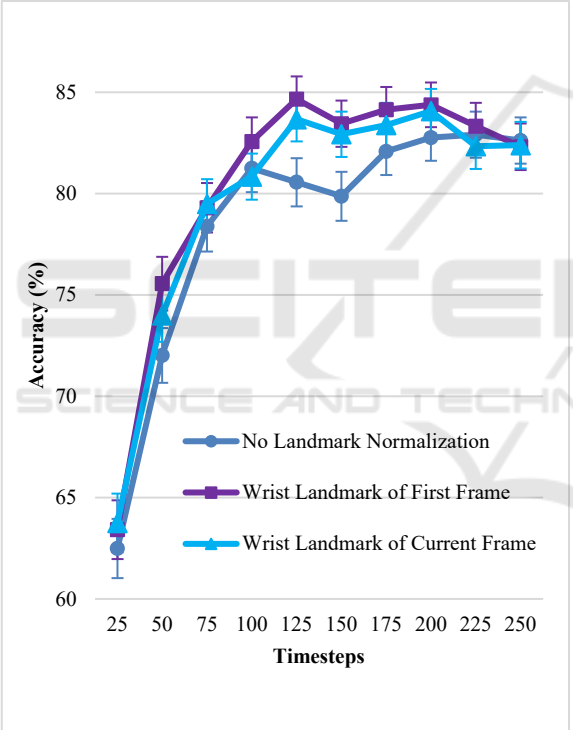

Figure 2: Accuracy depending on the number of timesteps used per gesture and the landmark normalization used over the raw landmark coordinates.

Comparing the state-of-the-art works, it is fair to highlight that the hand pose recognition usually gets higher performance because it deals with fixed images, which are more stable and easier to recognize. In contrast, the hand gesture recognition involves sequences of images of varying lengths, requiring normalization and handling of temporal dynamics, which complicates the recognition task.

Table 1: Accuracy for different landmark-based input formats using 125-timestep gesture length.

Landmark-based input	Accuracy (%)
Raw Coordinates + No Normalization	80.56 ± 1.19
Raw Coordinates + Wrist of Current Frame Norm.	83.67 ± 1.12
Raw Coordinates + Wrist of First Frame Norm.	84.66 ± 1.09
Landmarks Speed	72.93 ± 1.34
Anthropometric Measures	82.20 ± 1.15

For example, a previous work (Benitez-Garcia et al., 2021) used 37 subjects from this dataset to train a ResNeXt-101 model with 47.56 million parameters and evaluating the remaining 13 subjects, achieving an 86.32% accuracy. This potentially allows for overfitting, where the model performs well on the test data because it has been optimized for it. Using the same setup, the proposed approach of using the raw coordinates and the wrist of first frame normalization obtained an 86.38 ± 2.03 % accuracy using a significantly lighter model with only 58,614 parameters. In addition, this work offers a competitive 84.66 ± 1.09 % accuracy evaluating a larger and more diverse set of 50 subjects, making the task more challenging and the model's generalization performance more critical.

Finally, we analyzed if using a limited number of landmarks was informative enough to provide a high performance. Analyzing the nature of the dataset gestures, we observed that all the 13 classes compromised the movement of the thumb, index and middle fingers, and the ring and little fingers were slightly used in the gestures. This supports the idea that we observed in a previous study (Luna-Jimenez et al., 2023), indicating that for example the variance of the index finger was higher ($\sigma 2$ =0.031) than the one of the little finger ($\sigma 2$ =0.019). This way, we used the wrist and all fingertips landmarks (6 landmarks) and the wrist, and fingertips from thumb, index and middle fingers (4 landmarks) to perform the experiments. In this case, the anthropometric measures were computed using only these landmarks (Euclidean distances between those pairs of landmarks). Figure 3 shows the accuracy depending on the landmark-based input format and the number of landmarks per hand.

We observed that using the wrist and the fingertips landmarks could be enough to obtain a high recognition performance when using raw landmark coordinates as inputs. For example, when using the raw coordinates and the normalization using the first video frame, the system achieved 82.15 ± 1.16 % and 81.46 ± 1.17 % when using 6

and 4 landmarks, respectively. However, there exists a performance decrease when using the anthropometric measures and reducing the number of landmarks. This could be due to a loss of spatial relationships between the different landmarks (from 210 when using 21 landmarks to 15 and 6 when using 6 and 4 landmarks, respectively).

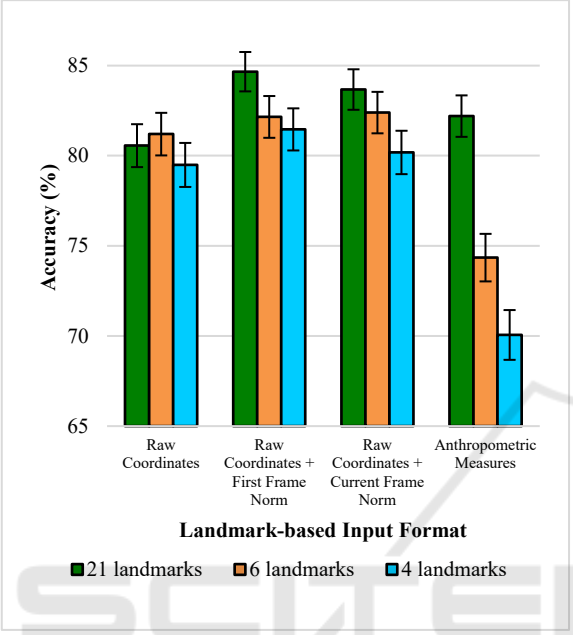

Figure 3: Accuracy depending on the landmark-based input format and the number of landmarks used per gesture.

4 CONCLUSIONS

This paper proposes a system to detect hand gestures using a limited number of landmarks from images. The proposed approach automatically extracts 21 MediaPipe landmarks (x and y coordinates of specific points) from the hand and feeds a deep neural architecture to model and recognize different hand gestures in a subject-independent scenario. This system analyzes the effect of using different landmark-based representations as inputs to a light RNN. It obtains significantly better accuracy when using the raw coordinates of the 21 landmarks or the hand anthropometric measures compared to using the speed of the hand landmarks. In addition, a performance tendency improvement is observed when using landmark-based normalizations over the raw coordinates. Moreover, using a limited number of hand landmarks (the ones with higher variance along the gestures) provides competitive gesture recognition performance.

As future work, it would be interesting to apply Multi-Modal Large Language Models to recognize gestures by feeding the models with the original images and landmarks representation. In addition, it would be interesting to apply this framework using other landmark detection models like OpenPose and for other datasets with a wider variety of gestures or microgestures (Chan et al., 2016) and/or related to sign language recognition. Finally, a real-time system that infers the prediction from a stream of a camera could be an application on a real use case of the proposed method.

ACKNOWLEDGEMENTS

This research was funded by Programa Propio 2024 from Universidad Politécnica de Madrid. Iván Martín-Fernández research was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). Sergio Esteban-Romero research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516).

This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), TRUSTBOOST (PID2023-150584OB-C21) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU/PRTR”.

REFERENCES

Bao, P., Maqueda, A. I., del-Blanco, C. R., & Garcia, N. (2017). Tiny Hand Gesture Recognition without Localization via a Deep Convolutional Network. *Ieee Transactions on Consumer Electronics, 63*(3), 251-257.

Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., Yanai, K., & Ieee Comp, S. O. C. (2021, Jan 10-15). *PIS Hand: A Video Dataset and Benchmark for Real Time Continuous Hand Gesture Recognition.* Paper presented at the 25th International Conference on Pattern Recognition (ICPR), Electr Network.

Chan, E., Seyed, T., Stuerzlinger, W., Yang, X.-D., Maurer, F., & Acm. (2016). User Elicitation on Single-hand Microgestures. [Proceedings Paper]. *34th Annual Chi Conference on Human Factors in Computing Systems, Chi 2016*, 3403-3414.

Gil-Martin, M., San-Segundo, R., Fernandez-Martinez, F., & Ferreiros-Lopez, J. (2020). Improving physical activity recognition using a new deep learning architecture and post-processing techniques. *Engineering Applications of Artificial Intelligence, 92*.

Gil-Martin, M., San-Segundo, R., Fernandez-Martinez, F., & Ferreiros-Lopez, J. (2021). Time Analysis in Human Activity Recognition. *Neural Processing Letters*.

Gil-Martín, M., San-Segundo, R., & de Córdoba, R. (2023). *Hand Pose Recognition through MediaPipe Landmarks*. Paper presented at the Modeling Decisions for Artificial Intelligence.

Gil-Martín, M., San-Segundo, R., Fernández-Martínez, F., & de Córdoba, R. (2020). Human activity recognition adapted to the type of movement. *Computers & Electrical Engineering, 88*, 106822.

Goodfellow, I. (2016). *Deep learning*. Cambridge, Massachusetts: Cambridge, Massachusetts The MIT Press.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., et al. (2019). MediaPipe: A Framework for Building Perception Pipelines. *ArXiv, abs/1906.08172*.

Luna-Jimenez, C., Gil-Martin, M., Kleinlein, R., San-Segundo, R., Fernandez-Martinez, F., & Acm. (2023). Interpreting Sign Language Recognition using Transformers and MediaPipe Landmarks. [Proceedings Paper]. *Proceedings of the 25th International Conference on Multimodal Interaction, Icmi 2023*, 373-377.

Mantecon, T., del-Blanco, C. R., Jaureguizar, F., & Garcia, N. (2019). A real-time gesture recognition system using near-infrared imagery. *Plos One, 14*(10).

Pheasant, S., & Haslegrave, C. M. (2006). *Bodyspace: Anthropometry, Ergonomics and the Design of Work*.

Quinonez, Y., Lizarraga, C., & Aguayo, R. (2022). Machine Learning solutions with MediaPipe. [Proceedings Paper]. *2022 11th International Conference on Software Process Improvement, Cimps*, 212-215.

Trindade, P., Lobo, J., & Barreto, J. P. (2012, 13-15 Sept. 2012). *Hand gesture recognition using color and depth images enhanced with hand angular pose data.* Paper presented at the 2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI).

Weiss, N. A. (2017). *Introductory statistics*: Pearson.

Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., & Li, Z. (2017). A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering, 2017*.