Data collection and annotation framework for vision-based finger intent recognition

Yana Marisova[1]

Received: 2025-06-13
Accepted: 2025-07-14
DOI: https://doi.org/10.5281/zenodo.19297947

Abstract. This study presents a functionally grounded framework for finger-intention recognition in robotic prototype, focusing on the rigorous collection, processing and annotation of egocentric visual data. Recordings were captured using smart glasses in realistic home environments, simulating object manipulation scenarios relevant to pediatric assistive applications. To mitigate the subjectivity inherent in user intention, a contact-based annotation logic was developed, ensuring objective class boundaries between transitional hand states. The methodology prioritizes data-centric principles, employing region of interest extraction, class balancing and multi-stage augmentation to enhance model robustness against occlusion and lighting variations. Training of MobileNetV2 and ResNet50V2 architectures on the curated dataset demonstrated stable learning dynamics and accurate detection of pre-shaping movements. The proposed framework closes the gap between raw visual input and real-time control requirements, supporting reliable vision-based actuation grounded in observed hand-object interactions.

Key words: dataset creation, annotation, intent recognition, computer vision, robotic control, keypoints.

[1] Bachelor of Computer Science,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
https://orcid.org/0009-0006-9631-331X

Introduction

The integration of computer-vision systems into robotic mechanisms transforms a prosthesis into an active vision-based interface that perceives the environment from a first-person perspective to infer operational intent based on hand kinematics [1]. While neural network architecture plays a role, system performance is primarily driven by the quality and representativeness of the training data. In vision-based control, errors such as false positives or delayed responses may trigger unintended actuation and compromise system stability, significantly impacting user experience [2].

Developing a robust framework for dataset reliability and annotation consistency is essential, particularly when designing for scenarios involving small-object manipulation. The framework presented in this study is grounded in the principles of Data-Centric AI, emphasizing systematic improvements in data quality rather than relying solely on hyperparameter tuning [3]. It directly addresses key challenges associated with egocentric vision, including limb occlusion, the dynamic nature of transitional hand states and the operational constraints imposed by real-time processing requirements [4].

This study contributes a structured data acquisition methodology specifically designed for egocentric vision in assistive control prototypes. The proposed approach seeks to align laboratory-based performance metrics with the real-world functional demands of real-time pediatric robotic prostheses.

Materials and methods. In this study, the selection of sensing and recording tools was driven by the data requirements of a vision-based intention recognition model, with particular emphasis on capturing realistic first-person visual conditions relevant to assistive use cases. Meta Ray-Ban smart glasses were specifically selected for their ability to record visual data from a true first-person perspective. By positioning the camera at eye level, the system provides a gaze-aligned optical vector that closely reflects the user`s natural attentional focus. This configuration is critical, as hand trajectories during object-reaching motions typically evolve from the periphery toward the center of the visual field, producing characteristic scale and perspective transformations that the recognition system must learn to interpret as precursors of intentional action [5].

First-person acquisition also introduces inherent visual complexities. The user`s hand frequently occludes the manipulated object and the fingers may partially occlude one another, causing the thumb to disappear from the camera`s field of view. The use of Meta Ray-Ban glasses facilitated the acquisition of ecologically valid data reflecting frequent occlusions, challenging the model to interpret incomplete or degraded visual information. The rigid head-mounted camera produces subtle, continuous changes in viewpoint due to natural head movements. Such variations enrich the dataset with realistic spatial diversity, providing critical training conditions for models expected to operate reliably in real-world scenarios.

Consequently, the recorded footage forms a continuous, unstructured video sequence. To transform this material into a curated dataset suitable for model training, a dedicated Python-based processing pipeline was developed. This pipeline automated frame extraction and performed preliminary consistency checks. The next section outlines the methods employed to correlate, filter and remove non-informative samples.

Results

Data acquisition methodology for a finger-intention recognition system:

The development of a computer vision model requires careful selection of both the programming environment and the technological tools used for neural network training, as well as seamless integration with the robotic system hardware. A combination of software and hardware resources was employed, chosen based on functionality, widespread adoption and suitability for the tasks at hand.

Prior to data acquisition and hardware testing, photo and video materials were captured under controlled conditions to optimize model training. This setup enables recording of so-called “pre-shaping” hand movements, which serve as strong predictors of the “Intent Grasp” prior to actual object contact [6]. The first-person perspective ensures accurate perception and interpretation of limb movements, while variations in angles and lighting enhance model adaptability and robustness across diverse operational conditions.

Preprocessing involved extraction of the hand region of interest (ROI) using the MediaPipe framework, followed by frame normalization and scaling [7]. MediaPipe was integrated for hand keypoint detection and ROI extraction, allowing the model to focus computationally on the most informative areas of each frame, reducing noise and improving gesture classification accuracy. A custom Python script was then used to automatically segment videos into individual frames, producing a raw dataset for subsequent annotation and construction of the final gesture dataset. The choice of Python as the main programming language was motivated by its comprehensive set of libraries for image analysis and deep learning, along with its simplicity and adaptability, which facilitate efficient implementation and integration of various software modules (see: Figure 1. System architecture).

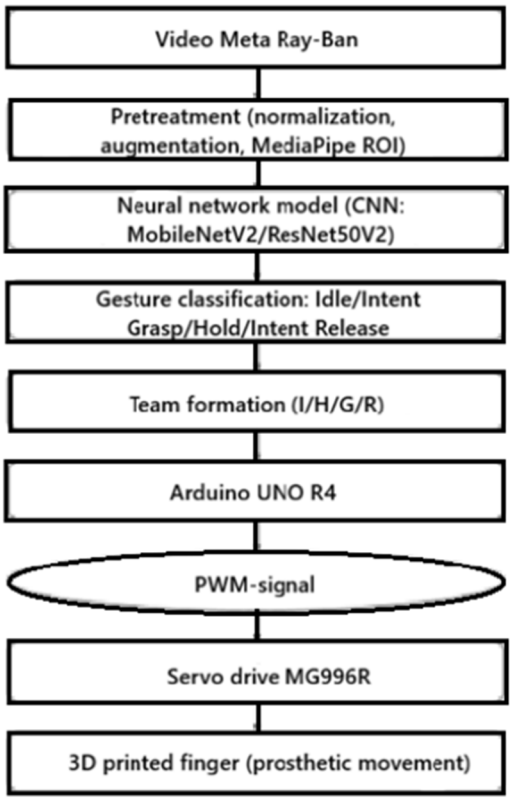

Figure 1. System architecture

Figure 1. System architecture. The hardware control components fall outside the scope of this study, which focuses on dataset construction and annotation

Frames that were blurred, underexposed or duplicated were removed to maintain dataset quality. The most challenging frames corresponded to boundary cases, where the distance between the finger and the object was less than a millimeter. In these instances, critical cues included subtle shadows cast by the finger or the deformation of soft fingertip tissue, indicating applied pressure.

Another important consideration was the choice of sampling frequency. Because human movements are inherently smooth, consecutive frames in a 30 fps or 60 fps video stream are

often nearly identical. Excessive correlation between frames can lead to model overfitting and artificially inflated accuracy metrics on the validation set.

Manual for research data annotation:

Annotator subjectivity can lead to overly detailed action labeling and, consequently, unpredictable robotic behavior. Variability in annotator labeling, such as one annotator marking “Intent Grasp” during hand movement and another only when the hand reaches the object, produces training data with ambiguous class boundaries. A neural network trained on such conflicting examples cannot establish a clear switching rule, which in real-world use may lead to unstable or inconsistent predictions: the prototype could unexpectedly close due to an incidental gesture or begin “trembling” (rapidly switching between states). Objective contact-based logic represents the most effective strategy to overcome subjectivity in intention labeling.

“Intention” is a cognitive process that may not always have a clear or precise visual manifestation. Gestures that appear different externally may be equivalent in 2D modeling and therefore fail to accurately reflect true human intent in practice. To address this challenge, a strict contact-based annotation logic was developed, relying on the interaction between the thumb and index finger with the object. Two binary features were defined: “Thumb Contact”, indicating physical contact of the thumb with the object and “Index Contact”, indicating physical contact of the index finger. Based on the combination of these features, four classes were created, forming the foundation of the proprietary Gestures Dataset.

Within this framework, the “Idle” state corresponds to moments, when the hand is in free space or approaching the object, with no contact initiated. The fingers may be relaxed or slightly flexed, but do not form a characteristic grip for any specific object. The “Intent Grasp” state occurs when the index finger or other fingers, excluding the thumb, have already touched the object, providing an initial support point, while the thumb has not yet closed the grip. This represents the “aiming” moment toward the target object, with pixels corresponding to the finger and object already in contact. The “Hold” state represents firm engagement of the object between the thumb and index finger or the palm, during which the hand may move the object in space and the prototype`s instant-release function is disabled. Finally, the “Intent Release” state is characterized by the index finger breaking contact with the object while the thumb may still maintain contact, reflecting the loss of stable contact with both fingers.

This annotation strategy ensures that the dataset captures subtle, but meaningful variations in hand-object interactions, providing a consistent and objective framework for training vision-based intention recognition models under conditions that closely approximate real-world use.

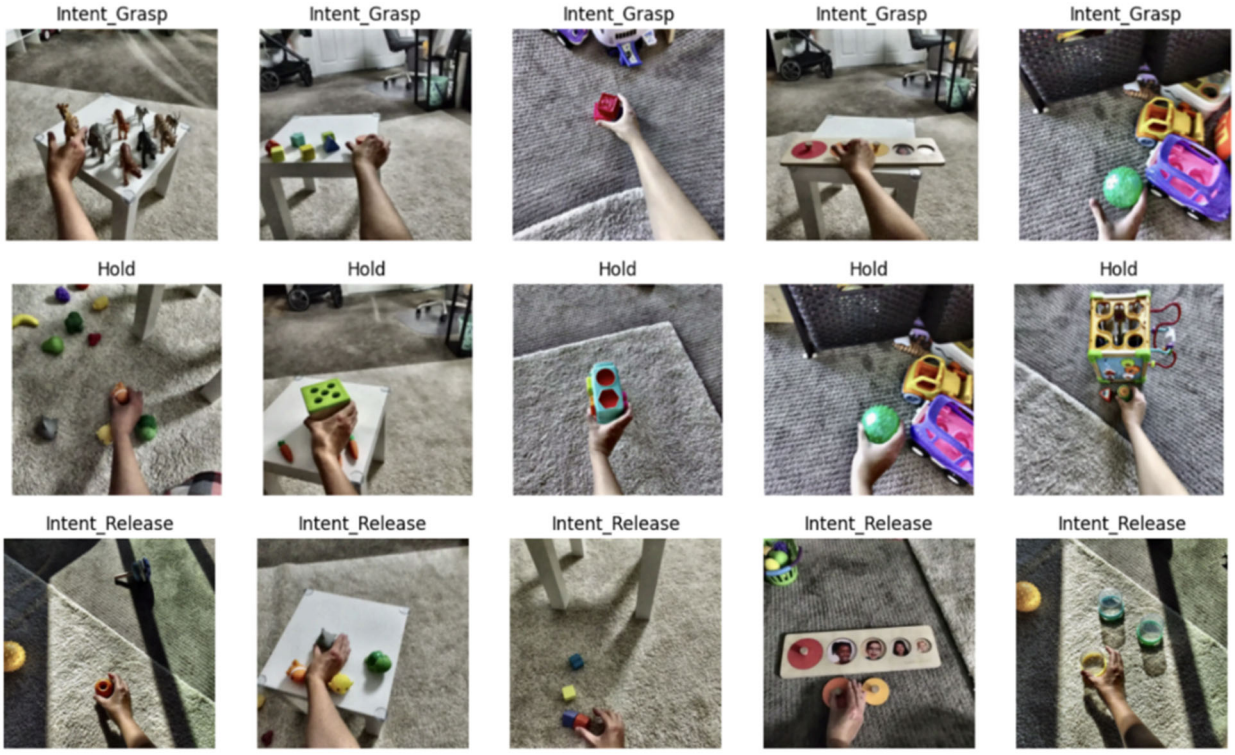

Figure 2. Representative frames from the first-person perspective dataset demonstrate the four finger intention classes:

"Idle", "Intent Grasp", "Hold" and "Intent Release"

When grasping an object, a person typically first stabilizes its position with the remaining fingers (or the index finger) and then closes the grip with the opposed thumb, until all fingers contact the object.

Among these classes, "Intent Grasp" and "Intent Release" are particularly critical, as they ensure correct prototype function and proper alignment with human intent. Detection of the "Intent Grasp" class enables the mechanism to act preemptively, anticipating the user`s action by fractions of a second and compensating for system latency - while the human already intends the action, the machine may not yet have fully responded. Similarly, recognizing the "Intent Release" class allows the system to detect the person`s intention to release the object before the hand fully withdraws, supporting smooth and synergistic human - robot interaction.

Dataset qualification for model training:

A critical step in data preparation was the reduction of input dimensionality through region of interest (ROI) extraction. Feeding the full frame from the smart-glasses camera directly into the neural network is both computationally inefficient and potentially detrimental to accuracy, as the network may inadvertently learn background features rather than focusing on the primary object of interest - the human hand [8].

Focusing the system on the hand-object interaction region was critical at this stage to ensure meaningful feature extraction. The resulting dataset included 1200 frames, evenly balanced across the four classes, with 300 images per class (see: Figure 3. Class balance histogram).

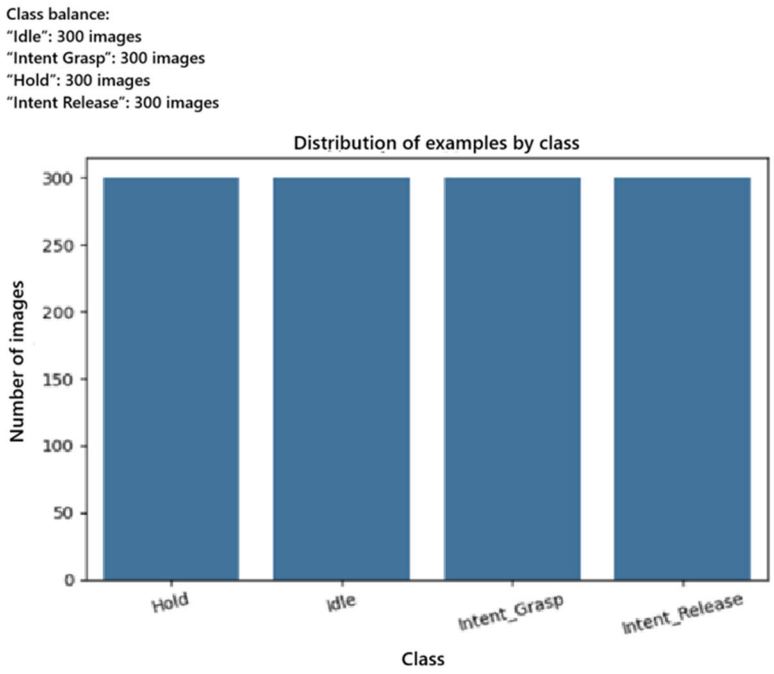

Figure 3. Class balance histogram

This uniform distribution eliminates class imbalance, which otherwise biases the network toward overrepresented classes and reduces recognition accuracy for less frequent states [9]. In video materials, the "Idle" state occupies the majority of frames, while transitional states such as "Intent Grasp" and "Intent Release" last only fractions of a second. Training on an improperly balanced distribution of media data would result with consistency in the model predicting "Idle", achieving high nominal accuracy, but being entirely nonfunctional in practice.

As described earlier, automatic ROI extraction was implemented using the MediaPipe Hands framework. Based on hand keypoints, a minimal area bounding rectangle was constructed and scaled by a factor of 1.8, isolating the informative region of the frame and reducing background noise.

Figure 4. Impact of region of interest (ROI) extraction. Left: original wide-angle frame. Right: extracted and normalized hand region, used for training

The data preparation stage addresses critical challenges in model training. ROI extraction focuses the network on visually relevant hand-object interaction patterns, reducing background-induced noise. Constructing the dataset in this manner allows the model to prioritize human intention over statistical or environmental artifacts, forming a solid basis for consistent and reliable recognition in vision-driven systems.

The role of multispectral conditions:

To ensure the model`s ability to operate correctly under conditions not represented in the training set, data variability is critically important. Given that the raw dataset comprised only 1200 frames, which is relatively small for deep learning, the framework extensively employs methods to artificially increase variability. From the outset, recording conditions were designed to introduce natural variation in the data.

The network was trained to ignore background features such as furniture, floor textures and room decor, while variations in lighting ensured robustness to brightness and contrast changes.

Augmentation techniques, including geometric transformations, simulated differences in camera position and object distance, enabling the model to recognize gestures independent of frame orientation.

Photometric transformations were employed to enhance robustness under varying lighting conditions, while Gaussian blur simulated motion blur that naturally occurs during rapid hand movements, improving performance in dynamic scenes and optimizing prototype operation.

Cutout augmentation was particularly important for this model, as it simulates partial occlusions of the hand or fingers, forcing the network to make decisions based on visible portions of the hand, such as the palm or remaining fingers, when the index finger is occluded by an object.

Finally, MixUp - a nontrivial technique for classification tasks was used to regularize the model by encouraging linear behavior between classes. In the context of prototype control, this promotes more stable probabilistic outputs during transitional hand movements, facilitating subsequent signal filtering [10, 11, 12].

The application of this comprehensive augmentation strategy transforms a limited dataset into an extensive and highly variable training base, ensuring that the model can adapt to unpredictable real-world conditions while maintaining the stability required for reliable vision-based control.

Dataset-driven effects on robotic system performance:

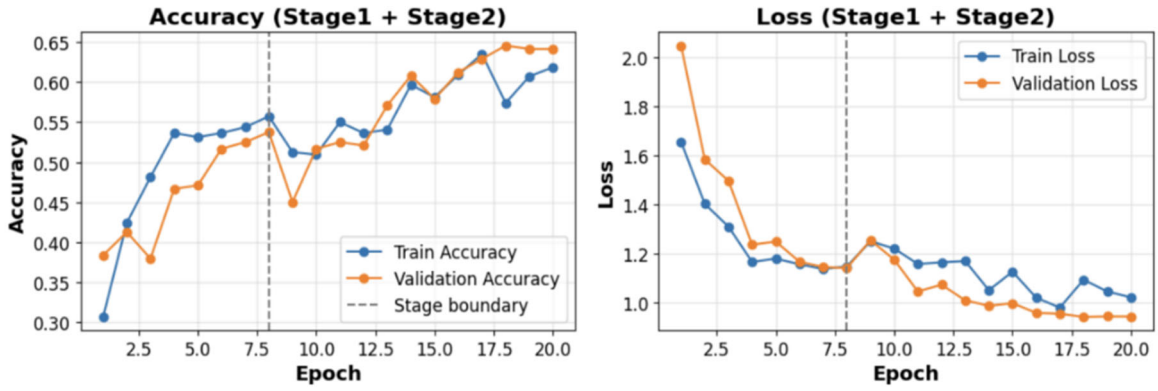

Figure 5. MobileNetV2 training results

Training the models demonstrated consistent performance across training and validation. Both MobileNetV2 and ResNet50V2 demonstrated smooth and consistent learning curves, with improvements in accuracy and loss and no notable gap between training and validation

metrics, reflecting stable generalization. MobileNetV2 reached an accuracy of 0.646, providing a practical balance between predictive performance and low computational demands.

MobileNetV2 provides adequate performance with low computational demands, whereas ResNet50V2 attains superior accuracy (0.729) and smoother learning curves, demonstrating improved generalization at the expense of increased hardware requirements [13, 14].

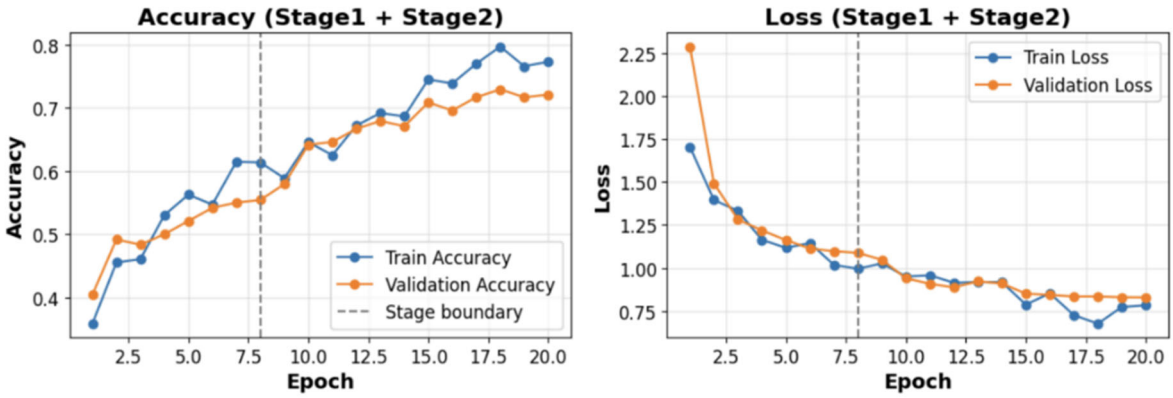

Figure 6. ResNet50V2 training results

Data quality directly influences the phenomenon of "jitter", characterized by rapid, chaotic switching between classes in consecutive video frames. In the presence of inconsistent labels, such as similar frames annotated as "Idle" and "Intent Grasp", the model becomes uncertain, resulting in high prediction entropy for borderline states. Recognizing that it is impossible to completely eliminate noise from the data, the framework incorporates post-processing techniques, including smoothing and debounce, to mitigate these effects [15].

Erroneous detection of "Intent Grasp" could lead to involuntary prototype closure. Through strict contact-based annotation logic, such cases were minimized by training the model to differentiate between "mere approach" and "active intent". Failure to recognize "Intent Release" could result in excessive object fixation [16]. Dataset balancing ensures sufficient weight for this class during training, enabling reliable hand opening.

A system-level safeguard, consisting of transition-permission logic and a defined action cycle, was applied to the neural network outputs to prevent abrupt or illogical state changes, ensuring smooth and stable operation.

Metric	MobileNetV2	ResNet50V2
Accuracy	0.646	0.729
Macro F1	0.646	0.729
Precision	0.556	0.662
Recall	0.583	0.750

Table 1. Summary table of metrics for both architectures

Conclusion

This study presents a comprehensive methodology for constructing a specialized visual dataset aimed at finger-intention recognition in robotic prosthetics. Several factors are critical for such a task, including data quality, multispectral variability, dataset stability, class balancing and the safety of mechanical control. The analytical framework developed in this

work enables the creation of a dataset that is well-structured, accurate, realistic and diverse. A key contribution is the implementation of a contact-oriented annotation logic, which removes the inherent subjectivity of intention interpretation. By relying solely on the interaction between the thumb and index finger, the system ensures reproducible class boundaries and prevents the emergence of ambiguous training signals.

Class balancing, region of interest extraction and a carefully designed augmentation strategy further enhance system robustness across diverse operating conditions, improve mechanical stability and facilitate seamless human-device interaction.

Both MobileNetV2 and ResNet50V2 demonstrated stable training behavior without signs of overfitting, highlighting the effectiveness of the curated dataset. While ResNet50V2 achieved superior accuracy, MobileNetV2 provides a favorable trade-off between computational demand and performance, making it particularly suitable for embedded robotic systems. Post-processing techniques serve as an additional stabilization layer, filtering noisy predictions and ensuring consistent state transitions.

Collectively, these results underscore that objective, contact-oriented annotation strategies combined with comprehensive data augmentation are essential for achieving robust and reliable intention recognition under realistic operating conditions.

References

1. Ghazaei, G., Alameer, A., Degenaar, P., Morgan, G., & Nazarpour, K. (2017). Deep learning-based artificial vision for grasp classification in myoelectric hands. Journal of neural engineering, 14(3), 036025. https://doi.org/10.1088/1741-2552/aa6802

2. Sensinger, J. W., & Dosen, S. (2020). A review of sensory feedback in upper-limb prostheses from the perspective of human motor control. Frontiers in Neuroscience, 14, 345. https://doi.org/10.3389/fnins.2020.00345

3. Mazumder, M., et al. (2022). Dataperf: Representative benchmarks for data-centric AI. DOI:10.48550/arXiv.2207.10062

4. Bambach, S., et al. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. Proceedings of the IEEE International Conference on Computer Vision (ICCV). DOI:10.1109/ICCV.2015.226

5. Land, M., et al. (1999). The roles of vision and eye movements in the control of activities of daily living. Neuroscience, 91(4), 1311-1325. DOI: 10.1068/p2935

6. Jeannerod, M. (1984). The timing of natural prehension movements. Journal of Motor Behavior, 16(3), 235-254. DOI: 10.1080/00222895.1984.10735319

7. Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. https://doi.org/10.48550/arXiv.1906.08172

8. Oudah, M., Al-Naji, A., & Chahl, J. (2020). Hand gesture recognition based on computer vision: a review of techniques. Journal of Imaging, 6(8), 73. https://doi.org/10.3390/jimaging6080073

9. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

10. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48. DOI:10.1186/s40537-019-0197-0

11. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1903.12261

12. Thulasidasan, S., et al. (2019). On mixup training: Improved calibration and predictive uncertainty in deep neural networks. Advances in Neural Information Processing Systems (NeurIPS), 32. https://doi.org/10.48550/arXiv.1905.11001

13. Sandler, M., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. CVPR. https://doi.org/10.48550/arXiv.1801.04381

14. He, K., et al. (2016). Identity mappings in deep residual networks. ECCV. https://doi.org/10.48550/arXiv.1603.05027

15. Kopuklu, O., et al. (2019). Real-time hand gesture detection and classification using convolutional neural networks. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). https://doi.org/10.48550/arXiv.1901.10323

16. Scheme, E., & Englehart, K. (2011). Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges. IEEE Transactions on Biomedical Engineering, 58(6), 1598-1609. DOI: 10.1682/jrrd.2010.09.0177