



Combined Convolutional Neural Network (CNN) and Keypoint-based Method for Recognizing Finger Interaction Intent States

Yana Marisova  ¹ *

¹ National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (Ukraine).
Bachelor of Computer Science, Department of Artificial Intelligence.

* Corresponding Author, e-mail: marisova.ai@ukr.net

ARTICLE INFO

Research Article

Received:

10 October 2025

Revised:

25 November 2025

Accepted:

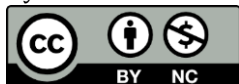
15 December 2025

Published online:

25 December 2025

Copyright © 2025

by author



This is an open access journal and all published articles are licensed under a Creative Commons Attribution—NonCommercial 4.0 International (CC BY-NC 4.0)

DOI: [10.5281/zenodo.19546851](https://doi.org/10.5281/zenodo.19546851)

ABSTRACT

This work addresses the challenge of developing intuitive and accessible control systems for upper-limb prostheses with particular emphasis on pediatric applications. Conventional approaches - such as myoelectric and mechanical control often suffer from limitations related to signal stability, usability or reliability. The proposed hybrid method introduces computer vision as an alternative command channel, enabling vision-based prosthetic control driven by the interpretation of user intention, without relying on wearable bioelectrical or tactile sensors. Intention states are inferred from observed hand interactions using a contact-based labeling logic and a combined CNN and keypoint-based framework. A hardware-software prototype was implemented using a microcontroller platform and a 3D-printed robotic finger, in which a vision-based AI model performs real-time intention-state classification and generates corresponding actuation commands. A preliminary cost and feasibility assessment indicates the potential suitability of the proposed approach for cost-sensitive assistive devices. Considering the increasing number of individuals injured as a result of russian military aggression, this solution provides a highly relevant and socially significant assistive mechanism for both civilians and military personnel. Additionally, this system holds significant potential for users with congenital limb malformations, facilitating a more active lifestyle via responsive robotic prosthetic fingers. Consequently, the proposed system represents a practical and scalable technology that should be made accessible to every individual in need.

KEYWORDS

prosthetics, assistive technology, vision-based control, robotics, hand keypoint detection, convolutional neural networks, human-machine interaction.

Introduction

The evolution of assistive technologies has led to increasingly sophisticated robotic hands. To fully utilize these mechanical capabilities in real-world scenarios, research is shifting towards advanced intent recognition methods that enable proactive rather than reactive control strategies (Villani et al., 2018; Mouchoux et al., 2022). Previously, prosthetic solutions were constrained by limited functional bandwidth, mechanical inaccuracies, latency in real-time control and prohibitive costs (Ghazaei et al., 2017; Marković et al., 2014). Advances in artificial intelligence, sensorimotor control modules and computer vision-driven interfaces have now facilitated the development of ergonomically optimized, intuitive prosthetic systems (Mouchoux et al., 2022). These innovations enable precise intention recognition, real-time actuation and adaptive feedback, restoring dexterity and eliminating biases based on external appearance, thereby approaching the performance of biological limbs.

It is also important to highlight that the demand for prosthetic devices has surged following the Russian invasion of Ukraine, as an increasing number of civilians and military personnel require new limbs to maintain multifaceted, independent lives. This situation presents novel challenges for engineers, who must ensure that next-generation bionic systems are safe, ergonomic, intuitive, easy to use, accessible to all segments of the population, capable of delivering a broad range of functionalities, effectively serving as true replacements for lost body parts.

The CV interface for robotic interaction represents a state-of-the-art convergence of information technology, rehabilitation engineering, robotics and AI. This integrated hardware-software system demonstrates a marked performance advantage over previous iterations, epitomizing the next generation of robotic integration into human life as a fully functional, context-aware assistive tool. In certain scenarios, such systems seamlessly support complex motor tasks. As the vanguard of engineering innovation, this platform continues to evolve rapidly, propelled by continuous advancements in sensorimotor processing, real-time control algorithms and AI-powered decision-making frameworks (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015).

Literature Review

The scholarly basis of computer-vision-driven intention recognition combines studies on deep convolutional architectures and works focused on assistive control interfaces. Foundational CNN approaches were proposed by LeCun et al. (1998), Krizhevsky et al. (2012), and Simonyan and Zisserman (2015), while further advances in residual and lightweight backbone design were presented by He et al. (2016) and Sandler et al. (2018). The problem of real-time hand tracking and landmark extraction, which is essential for stable gesture interpretation, was substantially advanced by MediaPipe Hands (Zhang et al., 2020). At the conceptual level, vision-based human intention recognition was systematized by Kurmankhojayev et al. (2021), and the use of deep learning for grasp-related prosthetic tasks was demonstrated by Ghazaei et al. (2017). Research on continuous semi-autonomous prosthesis control with a depth sensor (Mouchoux et al., 2022) and on stereovision for closed-loop grasping in hand prostheses (Marković et al., 2014) confirms the practical importance of visual channels for assistive robotics. Taken together, these studies show that reliable real-time prosthetic interaction requires a combination of expressive visual features, geometrically grounded hand representation, and stable decision logic, which substantiates the relevance of the proposed hybrid CNN and keypoint-based approach.

Problem Statement

The aim of the article is to develop and experimentally validate an intuitive and accessible upper limb prosthetic control system, focused in particular on a children's audience, based on a hybrid computer vision approach that provides recognition of user intentions without the use of wearable bioelectric or tactile sensors, as well as to evaluate its functional effectiveness, cost, and potential for implementation in socially significant assistive technologies.

Methods and Materials

This study employs a combination of open-source computational frameworks, analytical models and hardware developments, culminating in a fully functional apparatus for fabrication and operational deployment. The development covers the complete pipeline from data collection and model training to system integration and functional validation. Prior to experimental implementation, a rigorous theoretical analysis was conducted, including comparative evaluation and selection of state-of-the-art computational tools and neural architectures. This involved a rigorously structured analysis of foundational computational frameworks, systematic diagnosis of operational bottlenecks, real-time profiling of detection-model performance, formal justification for the incorporation of specific neural architectures, and a comprehensive quantitative evaluation of the system's functional efficiency.

This study evaluated two CNN backbone architectures: ResNet50V2 for accuracy-oriented performance and MobileNetV2 for low-latency operation (He et al., 2016; Sandler et al., 2018). The hardware platform comprises an Arduino UNO R4 WiFi microcontroller, an MG996R servo motor and a 3D-printed finger prototype, with data acquisition facilitated by a Meta Ray-Ban camera. A convolutional neural network (CNN) was trained to analyze input frames and classify them into four distinct intention states: "Idle", "Hold", "Intent Grasp" and "Intent Release". The system is implemented in Python, leveraging TensorFlow and Keras frameworks, the OpenCV library and Google Colab for model training. PyCharm IDE is employed for development, debugging, project structuring and seamless integration with version control systems (Git).

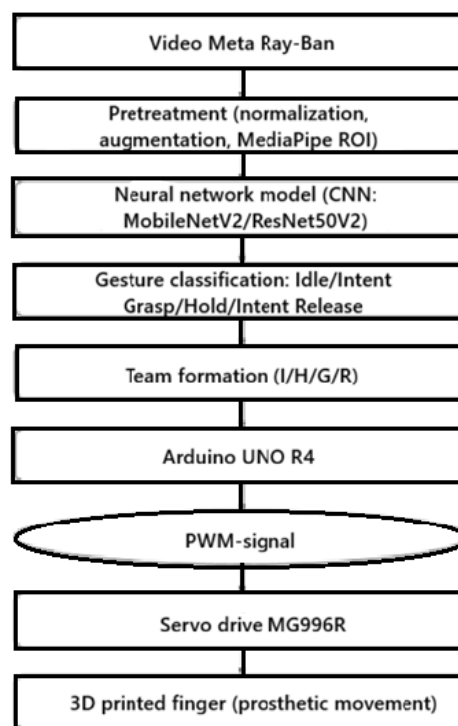


Figure 1. System architecture

Figure 1. Shows the relationship among all components employed to ensure the proper functioning of the model and general architecture

The proposed hybrid architecture establishes an effective vision-based control interface for a robotic finger prototype, enabling real-time sensori motor coupling, adaptive response mechanisms and highly accurate intention decoding. This development demonstrates the feasibility of hybrid vision-based intent recognition for assistive robotic control and next-generation prosthetic control technologies.

Results and Discussion

Synergy of CNN features and keypoint-based geometry – enhanced recognition accuracy of finger-movement intentions

Comparative analysis indicates that isolated CNN-based or keypoint-based paradigms exhibit inherent methodological limitations, failing to provide sufficiently robust intent recognition within real-world human-machine interaction contexts. CNN-based classifiers exhibit strong capability for extracting high-level visual descriptors - such as texture signatures, spatial contours, geometric morphology and global contextual cues (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). Their performance degrades significantly under domain-shift conditions, including variations in illumination, background clutter, camera perspective, hand orientation and occlusion patterns. CNN inference imposes substantial computational overhead, which may induce latency, throughput bottlenecks and instability in real-time control applications. Crucially, CNN alone does not provide sufficient granularity for interpreting micro-kinematic changes in finger-phalange articulation.

Keypoint-based skeletal models, such as MediaPipe Hands or similar landmark-estimation architectures, offer high resilience to environmental perturbations and provide anatomically grounded representations of the hand pose, capturing joint coordinates, inter-phalange angles and kinematic topology (Bazarevsky et al., 2020; Zhang et al., 2020). Despite this, keypoint-only pipelines suffer from limited intent-level discriminability: gestures with similar 2D skeletal projections may encode distinct interaction intents, making them difficult to differentiate (Molchanov et al., 2015). Moreover, high-velocity finger trajectories often exceed the temporal resolution of landmark detectors, causing frame-to-frame jitter, label ambiguity and error propagation across the classification pipeline.

These constraints justify the introduction of a hybrid, feature-fusion paradigm that integrates complementary strengths of CNN-derived deep visual embeddings and keypoint-based geometric descriptors (Li & Wu, 2019; Srivastava & Salakhutdinov, 2012). The synergistic architecture yields a multimodal representation that is functionally expressive for intention recognition and anatomically precise. In this configuration, keypoints effectively reduce spatial noise within the ROI, functioning as a dynamic localization and alignment module. The CNN processes the normalized hand image, enabling refined extraction of visual micro-features associated with finger-contact events and intention-specific motion patterns.

This multimodal strategy results in significantly enhanced classification accuracy and operational stability, achieving up to 72.9% accuracy with ResNet50V2 on the proposed four-class intent taxonomy ("Idle", "Hold", "Intent Grasp", "Intent Release"), which is also reflected in the metric diagrams in Figures 2 and 3.

Importantly, the hybrid pipeline demonstrates robustness unattainable by either modality in isolation, making it a compelling solution for next-generation prosthetic control systems, real-time robotic manipulation and advanced human-machine frameworks.

An advanced hybrid recognition framework combining convolutional neural representations with keypoint-based geometric descriptors

At the preliminary stage of the study, the MediaPipe Hands framework was utilized as a high-precision hand-pose estimation module (Zhang et al., 2020). It performs real-time detection of 21 anatomically grounded keypoints and estimates 21 anatomically grounded hand landmarks with approximate 3D coordinates. It also generates a robust, geometry-normalized ROI that serves as the foundation for subsequent feature extraction and intention-recognition processing. After the region of interest is identified, the extracted data are fed into the CNN for intent or gesture classification.

This establishes a complementary interaction between the two subsystems: the CNN is responsible for interpreting high-level visual semantics, while the keypoint-based module ensures ROI stabilization, and provides precise geometric, kinematic information. Together they form a synergistic pipeline that combines features relevant to intention recognition understanding with anatomically grounded structural accuracy.

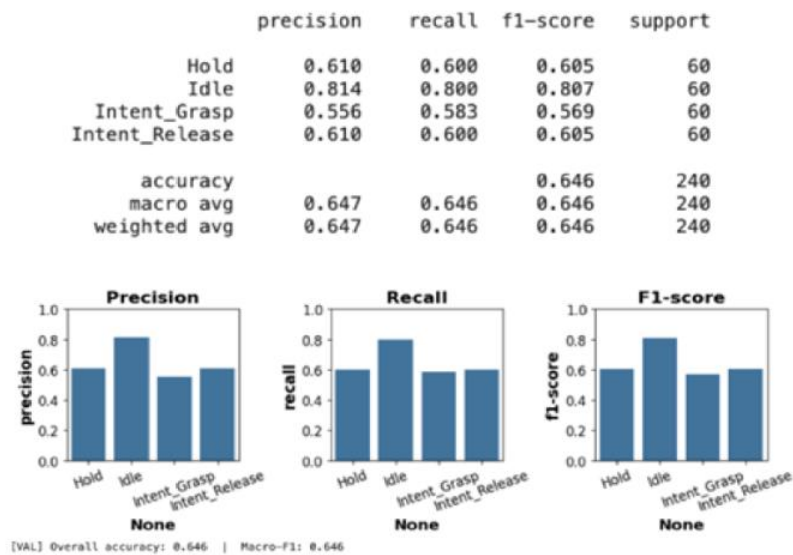


Figure 2. Model evaluation on MobileNetV2

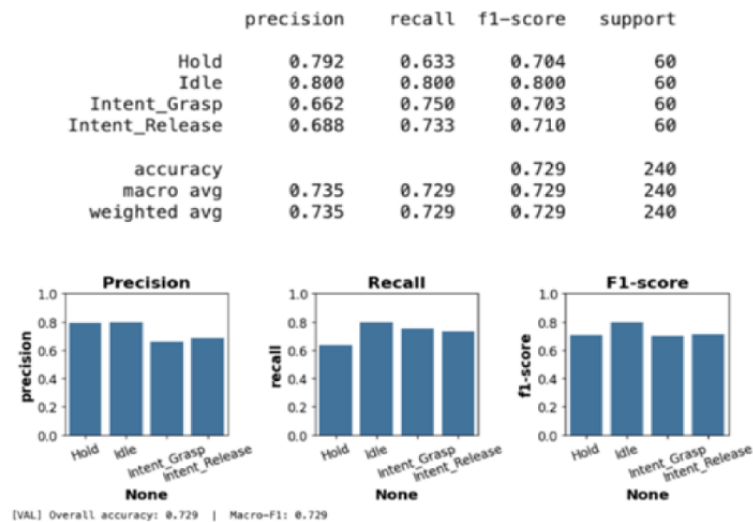


Figure 3. Model evaluation on ResNet50V2

Figure 2 and Figure 3. Architecture assessment metric diagrams for four classes

This functional separation allows the system to fuse the geometric precision of landmark-based detection with the broad generalization capacity of deep learning models, yielding a more robust and functionally consistent recognition pipeline.

A unique four-class intention framework: an overview of the underlying mechanism

A custom dataset comprising 4 intention states was created to train the model on control commands. The CNN models were trained on the following classes: “Idle”, “Hold”, “Intent Grasp” and “Intent Release”. A contact-based labeling logic was employed, relying on the interaction between the thumb and the index finger. Binary indicators were used, where 0 denotes the absence of contact and 1 denotes the presence of contact. Based on this principle, each class is encoded as follows: “Idle”: 0 - 0 - the hand does not interact with any object; “Hold”: 1 - 1 - stable contact between the thumb and the index finger; “Intent Grasp”: 0 - 1 - the index finger is already touching the target object while the thumb is not yet in contact, indicating the initiation of a grasping action; “Intent Release”: 1 - 0 - the thumb maintains contact while the index finger has already lost it, signaling the intention to release the object.

Data collection for the dataset was performed using photo-video smart glasses. This first-person perspective is essential for ensuring the correct operation of the prototype, as it captures realistic visual conditions - hand position, orientation, viewpoint. It replicates how a user naturally perceives and interprets their own hand movements during tool manipulation.

During experimental testing, several bottlenecks were identified-primarily misclassifications between “Idle” - “Intent Grasp” and “Hold” - “Intent Release”, as these state transitions determine the initiation or termination of prosthetic movement. To mitigate such errors, enhance robustness against random information spikes, and improve overall system stability and predictability, a decision-filtering mechanism was introduced. It combines probability thresholding with transition inertia, ensuring that class changes occur only when the model consistently maintains a high confidence level.

This approach mitigates false activations, thereby enhancing the reliability of transitions between intention states.

Model architecture, feature-fusion mechanism, and decision-making logic

The trained convolutional neural networks MobileNetV2 and ResNet50V2 were employed as backbone architectures (He et al., 2016; Sandler et al., 2018). These distinct backbone architectures were evaluated to address different operational constraints: MobileNetV2 was selected for its low latency in embedded applications, while ResNet50V2 was utilized for its higher representational capacity. A transfer-learning strategy was adopted, allowing the parameters of the base network to be initially frozen, followed by selective unfreezing of the upper layers to fine-tune high-level features (Yosinski et al., 2014; Weiss et al., 2016). A key element of the implementation is the feature-fusion mechanism positioned after the backbone layers: the output feature tensor is processed through parallel pathways - global average pooling and global max pooling. The resulting descriptors are then concatenated, providing both smoothed (average-pooled) and high-salience (max-pooled) representations of the extracted visual features (Lin et al., 2014). In the final classification stage, a dense layer with L2 regularization, batch normalization, ReLU activation and Dropout was applied to mitigate overfitting and enhance generalization. A concluding Softmax layer produced the probabilistic class distribution across the four intention states.

Experimental results demonstrate that the selected architectures maintain high stability and predictive accuracy with minimal signs of overfitting, thereby supporting the effectiveness of the proposed backbone configuration and feature-fusion strategy for real-time intent recognition.

To construct the decision-making mechanism, a multi-level stabilization pipeline was developed. The first component is the implementation of a debounce algorithm, which ensures robustness against random spikes in the input stream and improves the overall stability of the system. The debounce strategy also establishes the temporal consistency required for correct and safe operation of the robotic actuator (Klotzbuecher & Bruyninckx, 2012). In addition, a high confidence threshold is applied to reduce the likelihood of false-positive activations. This mechanism helps ensure that a state transition is triggered only when the model consistently maintains a sufficiently high posterior probability over a defined temporal window. A predefined permissible transition trajectory is enforced to regulate intent progression: “Idle” - “Intent Grasp” - “Hold” - “Intent Release” - “Idle”. At the level of a finite-state machine (FSM), all logically inconsistent or unsafe transitions are explicitly blocked.

This prevents erroneous command propagation, supports deterministic system behavior, and maintains semantic coherence of the control cycle.

Final improvements in stability and accuracy of the proposed system compared to using CNN-only vs CNN + MediaPipe-guided ROI

The outcome of this work is a hardware-software implementation serving as a proof-of-concept, which successfully integrates visual-intent analysis with physical actuation of a servo mechanism. Experimental evaluation confirmed the substantial advantages and practical usability of the proposed hybrid methodology. It is grounded in the synergistic interaction between CNN-based feature extraction and keypoint-driven geometric modeling. The obtained computational and experimental results demonstrate the successful achievement of the primary objective: the development of an operational prototype of a CV system capable of interpreting gesture intentions and translating them into physical actions of a robotic finger prototype.

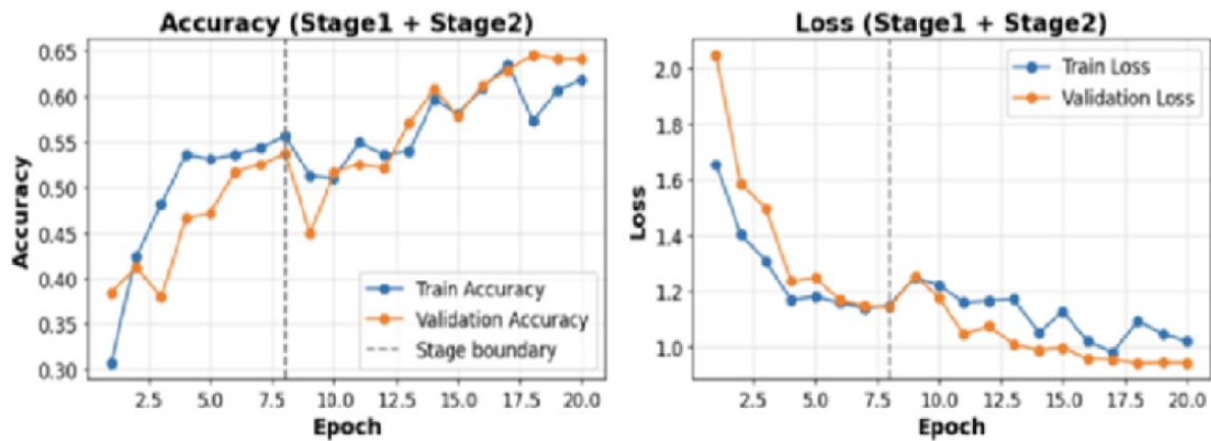


Figure 4. Training graphs on the MobileNetV2 network

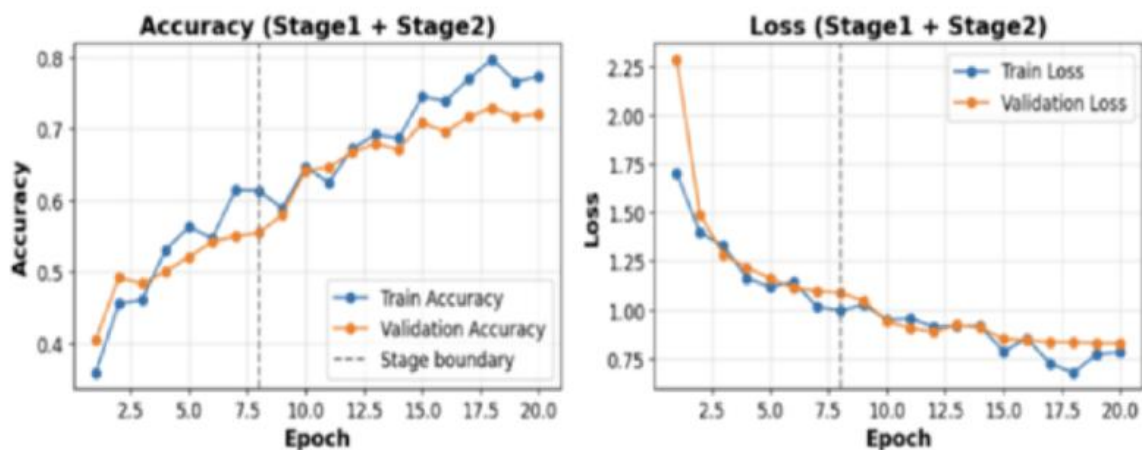


Figure 5. Training graphs on the ResNet50V2 network

Figure 4 and Figure 5 show the training graphs of selected networks with accuracy and loss metrics

An extended analysis of the confusion matrix demonstrated a robust discriminative capability of the proposed system, particularly in distinguishing the most critical transitional intention states “Intent Grasp” and “Intent Release”. These states exhibit subtle micro-kinematic differences that are often indistinguishable for CNN-only pipelines operating without keypoint-guided ROI stabilization, which lacks the semantic richness of visual contextual cues. Observed performance improvements suggest that combining CNN-based high-level visual features with keypoint-guided geometric information provides a more reliable decision space. Moreover, the incorporation of a stabilized ROI significantly reduced background-related false activations, a common source of error in standalone CNN-based detectors that process the full image without spatial constraints. By isolating the hand region and filtering out irrelevant visual noise, the system achieved higher robustness, improved precision and predictable behavior under varying environmental conditions.

This combination provides high classification accuracy, stable and reliable visual capture, continuous real-time performance with minimal or acceptable latency, ensuring the relevance and ergonomics of the system during operation.

Conclusion

This work demonstrates that integrating geometric keypoint-based methods with convolutional neural networks provides a feasible vision-based control approach for a robotic finger prototype intended for assistive applications. The proposed hybrid pipeline reduces background-induced

noise and allows the neural network to focus primarily on fine-grained hand dynamics relevant to intention recognition. The implemented hardware-software prototype confirms the feasibility of applying computer-vision-based control in an embedded system of this type. A key factor contributing to system performance is the use of a custom first-person dataset combined with a contact-based intention logic derived from thumb-index interaction. This methodological design enables reliable differentiation of interaction intents even when gestures appear geometrically similar in 2D projections but correspond to distinct functional actions.

The resulting prototype operates in real time with acceptable latency for interactive use in a prototype setting. Stabilization mechanisms reduce sporadic classification errors and support smooth and predictable actuator behavior. The conducted study indicates that vision-based control systems can serve as a promising alternative control modality to conventional myoelectric approaches at the prototype level, particularly for pediatric-oriented assistive device concepts. A preliminary economic assessment supports the practicality and cost feasibility of the proposed solution, highlighting its potential for further development and validation.

Overall, the presented system contributes an engineering foundation for future research in vision-driven robotic and assistive technologies.

References

- Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2020). BlazePose: On-device real-time body pose tracking. *CVPR Workshops*. <https://doi.org/10.48550/arXiv.2006.10204>
- 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>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.48550/arXiv.1512.03385>
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.48550/arXiv.1709.01507>
- Klotzbuecher, M., & Bruyninckx, H. (2012). Coordinating robotic tasks and systems with rFSM statecharts. *Journal of Software Engineering for Robotics*, 3(1), 28–56. <https://aisberg.unibg.it/retrieve/e40f7b86-2bd9-afca-e053-6605fe0aeaf2/52-254-1-PB.pdf>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kurmankhojayev, G., et al. (2021). Vision-based human intention recognition: A survey. *IEEE Sensors Journal*, (79), 30509–30555 <https://doi.org/10.1007/s11042-020-09004-3>
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*. <https://doi.org/10.1109/5.726791>
- Li, H., & Wu, X.-J. (2019). DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5), 2614–2623. https://blog.csdn.net/Pineapple_Daisy/article/details/136349283
- Lin, M., Chen, Q., & Yan, S. (2014). Network in Network. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1312.4400>
- Marković, M., Dosen, S., Cipriani, C., Popović, D., & Farina, D. (2014). Stereovision and augmented reality for closed-loop control of grasping in hand prostheses. *Journal of Neural Engineering*, 11(4), 046001. <https://doi.org/10.1088/1741-2560/11/4/046001>
- Molchanov, S., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. <https://www.scribd.com/document/876243377/IEEE-Conference-Template-1>
- Mouchoux, J., Dosen, S., et al. (2022). Continuous semi-autonomous prosthesis control using a depth sensor on the hand. *Frontiers in Neurorobotics*, (15). <https://doi.org/10.3389/fnbot.2022.814973>

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.48550/arXiv.1801.04381>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1409.1556>
- Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2012/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html>
- Tan, M., & Le, Q. (2019). EfficientNet. In *International Conference on Machine Learning*. <https://doi.org/10.48550/arXiv.1905.11946>
- Villani, V., Pini, F., Leali, F., & Secchi, C. (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, (55), 248–266. <https://doi.org/10.1016/j.mechatronics.2018.02.009>
- Weiss, K., Khoshgoftaar, T., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, (3), 9. <https://doi.org/10.1186/s40537-016-0043-6>
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*. https://papers.nips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html
- Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). MediaPipe Hands: Real-time hand tracking. In *CVPR Workshops*. <https://ijrpr.com/uploads/V6ISSUE11/IJRPR56027.pdf>