Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7, 150-164
2026
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v10i7.13302

A cross-subject multimodal static gesture dataset for a vision-based gestural calculator: Image–landmark Fusion

Halima Abdelmoumene[1*], Lila Meddeber[2], Nasr Eddine Berrached[3], Boutkhil Sidaoui[4]
[1,2,3]Department of Electronics, University of Science and Technology of Oran, Algeria; abdelmoumene@cuniv-naama.dz (H.A.).
[4]Department of Computer Science, University of Naama, Algeria.

Abstract: This paper proposes a multimodal approach dedicated to static hand gestures, integrated into a gesture recognition system based on computer vision. The system combines normalized RGB images, segmented landmarks, and 3D information, extracted using the MediaPipe library, to simultaneously capture both the visual appearance and the joint geometry of the hand. The dataset comprises 20 subjects, 13 classes of symbolic gestures, and 62,400 paired samples, with each instance consisting of a normalized RGB image of the hand and a vector of 63 normalized coordinates. Two unimodal branches are evaluated in a strict 5-fold cross-validation protocol: EfficientNet-B0 for the image and a lightweight MLP for the landmarks. The results show that the landmark branch achieves 98.99% accuracy, compared to 96.36% for the image branch. Late probability fusion offers the best overall performance, with 99.02% accuracy, 99.02% Macro-F1, and 99.82% Top-2 accuracy. These results validate, on the one hand, the strong discriminative power of hand geometry for static gestures, and establish, on the other hand, that visual information provides a significant complementary contribution, thereby contributing to overall robustness. Beyond the numerical results, the article discusses the role of cross-subject validation and the place of multimodal fusion in real-time natural interfaces in HCI and HRI.

Keywords: *Cross-subject evaluation, Gesture-based calculator, Hand gesture recognition, Human-computer interaction. Multimodal fusion.*

1. Introduction

Natural human-computer interfaces aim to bridge the gap between human intent and computer control by enabling spontaneous, direct, and contactless interactions. Hand gestures are particularly interesting because they are based on common movements in daily life and can be detected at low cost using a simple RGB camera [1, 2]. This approach applies to both human-computer interaction and human-robot interaction, especially when physical contact with a conventional interface is impractical, unhygienic, or cognitively unnatural [2, 3].

Even with recent advances in deep learning, vision-based gesture recognition still faces a major challenge: generalizing to users who have not been seen before. Hand shape, range of motion, gesture type, and acquisition conditions vary considerably from one person to another. In this context, randomly splitting the samples can give an overly optimistic view of performance, as subject-specific characteristics may carry over from training to testing. Cross-subject protocols are therefore more demanding, but they better reflect real-world usage conditions [1, 3, 4].

A second challenge concerns the representation of the gesture itself. RGB images capture rich cues regarding appearance, contour, shadow, and texture, but remain sensitive to lighting, apparent scale, and visual noise. Conversely, 3D landmarks produced by pipelines such as MediaPipe Hands explicitly describe the hand's joint structure, which facilitates compact representations that are more robust to background noise and often better suited for recognizing static poses [5]. However, landmarks can

History: Received: 17 April 2026; Revised: 18 June 2026; Accepted: 23 June 2026; Published: 14 July 2026
* Correspondence: abdelmoumene@cuniv-naama.dz

degrade in cases of self-occlusion, difficult viewpoints, or tracking failure; relying solely on them is therefore not always sufficient in more varied scenarios [5-7].

This complementarity explains the value of multimodal approaches. By fusing visual and geometric cues, we can leverage the strengths of each modality while mitigating their respective weaknesses. Recent work on gesture recognition, and more broadly on multimodal learning, shows that well-chosen fusion strategies often improve robustness when multiple complementary views of the same phenomenon are available [3, 6, 8].

This article presents a consolidated scientific version of a study on a vision-based gesture calculator. The system is based on 13 static gestures associated with digits and operators, captured for both hands from 20 subjects. Each example associates an RGB image of the segmented hand with a vector of 63 coordinates corresponding to the 21 MediaPipe 3D landmarks. A separate evaluation is conducted on an image branch, a landmarks branch, and then on a late fusion of probabilities, within the framework of a strict 5-fold cross-validation protocol across subjects.

The main contributions are as follows:

(i) The development of a multimodal image-landmark solution designed for rigorous cross-subject evaluation.

(ii) A comparative analysis of two specialized branches, one based on appearance and the other on geometry.

(iii) The integration of a simple, reproducible, and effective late-fusion strategy.

(iv) The interpretation of the results obtained with a view to developing natural, robust, and deployable gesture interfaces.

This manuscript is organized into six sections. The first section presents the state of the art on gesture-recognition systems, multimodal fusion, cross-subject data, and gesture interfaces such as gesture calculators; the second section details the proposed protocol for the gesture-recognition system used to implement a gesture-based calculator; the third section discusses the implemented methods and the experimental results, followed by the discussion and limitations section; and finally, the conclusion and future works.

2. Statement of the Art

2.1. Vision-Based Gesture Recognition

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. Please do not revise any of the current designations.

Vision-based gesture recognition has made significant progress, evolving from manual descriptors to deep neural networks capable of learning directly from images [1, 2]. Convolutional neural networks have improved the recognition of static and dynamic gestures, especially when trained on diverse datasets and evaluated using rigorous protocols [9-11]. A recent review on gesture recognition on edge devices highlights that evaluating accuracy should not be limited to a single performance metric. It must also account for the computational constraints of these platforms, robustness against contextual changes, and the ability to adapt to new environments. This is essential for interactive systems used in real-world conditions [6]. Recent reviews also highlight the importance of selecting the right modalities, datasets, and evaluation criteria suited to real-world conditions. Representations based on hand keypoints are becoming increasingly important. MediaPipe Hands enables the rapid extraction of 21 3D markers for interactive applications using a single RGB camera [5]. This advancement has enabled the development of geometric pipelines, particularly for static gesture recognition. In this context, finger position is a key factor. However, recent literature shows that relying solely on keypoints does not solve all problems [3, 5].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

2.2. Multimodal Fusion and Cross-Subject Evaluation

The topic of multimodal fusion follows this same logic of complementarity. Within the general framework of multimodal learning, Baltrušaitis et al. have shown that heterogeneous information sources can be combined at different levels to improve the system's robustness [8]. For hand gestures, the image + landmarks combination is particularly coherent, as the two modalities are synchronous, derived from the same sensor, and complementary: the image preserves appearance cues, while the landmarks explicitly represent the joint structure. This complementarity is also consistent with the state-of-the-art research on multimodal fusion, whether it concerns general reviews on the fusion of heterogeneous modalities or more robust strategies in the face of partial failure of a modality [12, 13].

Building on this discussion of complementarity, recent research also highlights the importance of dataset diversity [4]. Similarly, the SHREC benchmarks and continuous datasets such as IPN Hand have shown that reported performance depends heavily on the evaluation protocol, the type of gesture, and the distinction between visible and invisible subjects [14, 15].

2.3. Gesture-Based Calculator Interfaces and the Scientific Gap

The robustness of gesture-recognition systems for natural HMI interfaces depends primarily on inter-subject diversity, the evaluation protocol, the quality of hand detection, and the model's ability to utilize complementary modalities. They also note that real-time and human-robot interaction (HRI) scenarios require a trade-off between accuracy, computational cost, landmark stability, and generalization ability [16-21].

In the field of applications, gesture-based calculators represent a relevant use case, as they require reliable recognition of discrete symbols in an interactive context. Their simplicity facilitates a detailed analysis of the system's behavior. Several studies have demonstrated the feasibility of these interfaces. However, few combine a large-scale paired multimodal corpus, rigorous cross-subject validation, and an in-depth analysis of the system's robustness [22-25].

A recent review in robotics shows that, for natural human-robot interaction (NHRI), the main challenges are cross-domain generalization, the lack of suitable datasets, the scarcity of multi-user scenarios, and the diversity of real-time evaluations [3]. This finding underscores the importance of analysing the corpus, employing a multimodal approach, and conducting cross-subject evaluation. Furthermore, recent reviews on HRI and deep recognition pipelines indicate that a system's robustness depends as much on the experimental protocol as on the choice of architecture and the realism of the scenarios studied [26-28].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

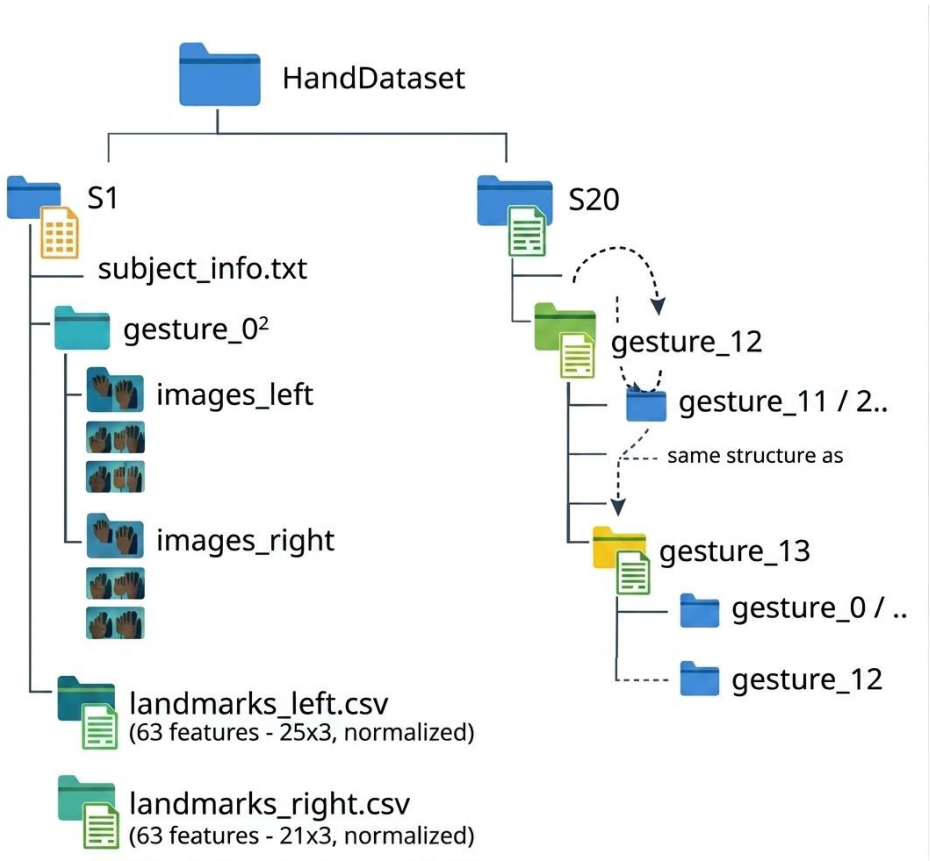

Figure 1.
Hierarchical organization of the multimodal dataset.

3. Multimodal Dataset and Static Gesture Recognition System

For this study, a database was designed. The dataset includes 20 participants, designated S1 through S20. For each participant, the data is organized hierarchically.

A main folder contains 13 subfolders, each corresponding to a gesture. Each subfolder contains samples from both hands as well as the associated joint landmark files. This organization allows for rigorous cross-validation among participants, as shown in Fig. 1.

The vocabulary comprises 13 static gestures, each associated with a calculator symbol: 0, 1, 2, 3, 4, 5, =, e, d, *, +, /, and -. The folders from gesture_0 to gesture_12 ensure consistency between acquisition, indexing, and learning, as shown in Fig. 2.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

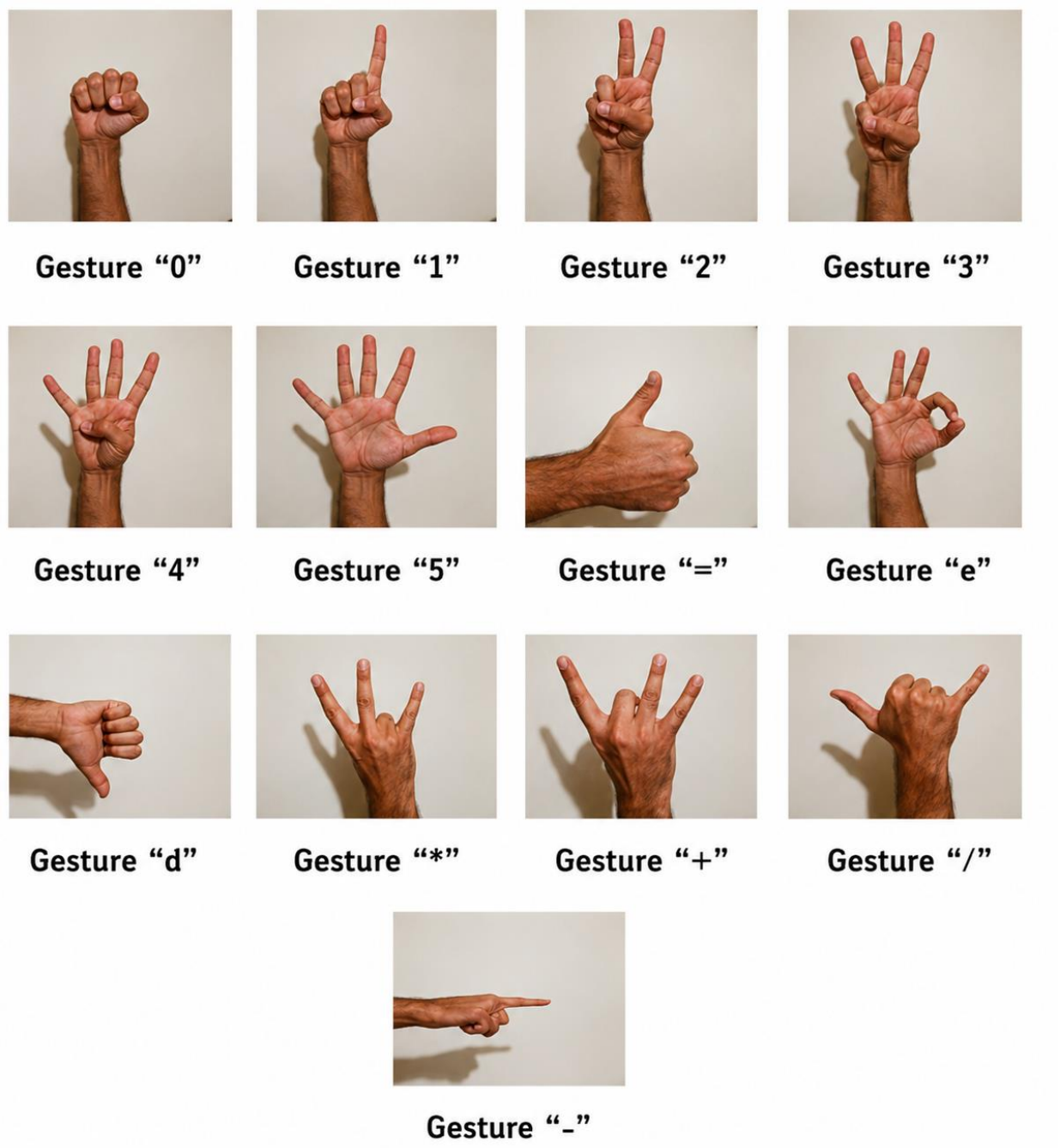

Figure 2.
The proposed gesture vocabulary for the gesture calculator is ordered from left to right, from Gesture_0 to Gesture_12.

Each sample is multimodal. The image modality consists of a segmented RGB cropped image of the hand, resized to 256 × 256 pixels during acquisition. The geometric modality is a 63-dimensional vector representing the (x, y, z) coordinates of the 21 landmarks on the hand, as shown in Fig. 3.

The landmarks are stored in a pre-normalized form to reduce sensitivity to global translation and facilitate comparison between participants.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

Table 1.
Statistical details of the multimodal database.

Element	Value
Number of subjects	20 (S1-S20)
Number of classes	13 static gestures
Samples per gesture and per subject	240 (120 left hand + 120 right hand)
Samples per subject	3,120
Total paired samples	62,400
Imaging modality	Segmented RGB crops, acquired at 256 x 256
Landmark method	21 3D points, or 63 normalized coordinates

Each participant performs 120 acquisitions per gesture and per hand, for a total of 240 per hand. This results in a total of 62,400 paired multimodal instances, as presented in Table 1. During indexing, image-landmark pairing is ensured by taking, for each gesture and each hand, the smaller of the number of available images and the number of rows in the corresponding CSV file. This verification prevents any desynchronization between modalities. This approach has three advantages: it allows for reliable comparison of unimodal and multimodal data, it is suitable for a 5-fold cross-validation, and it provides a consistent framework for a concrete symbolic application, which facilitates the analysis of class confusion and ensures robustness in interaction.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

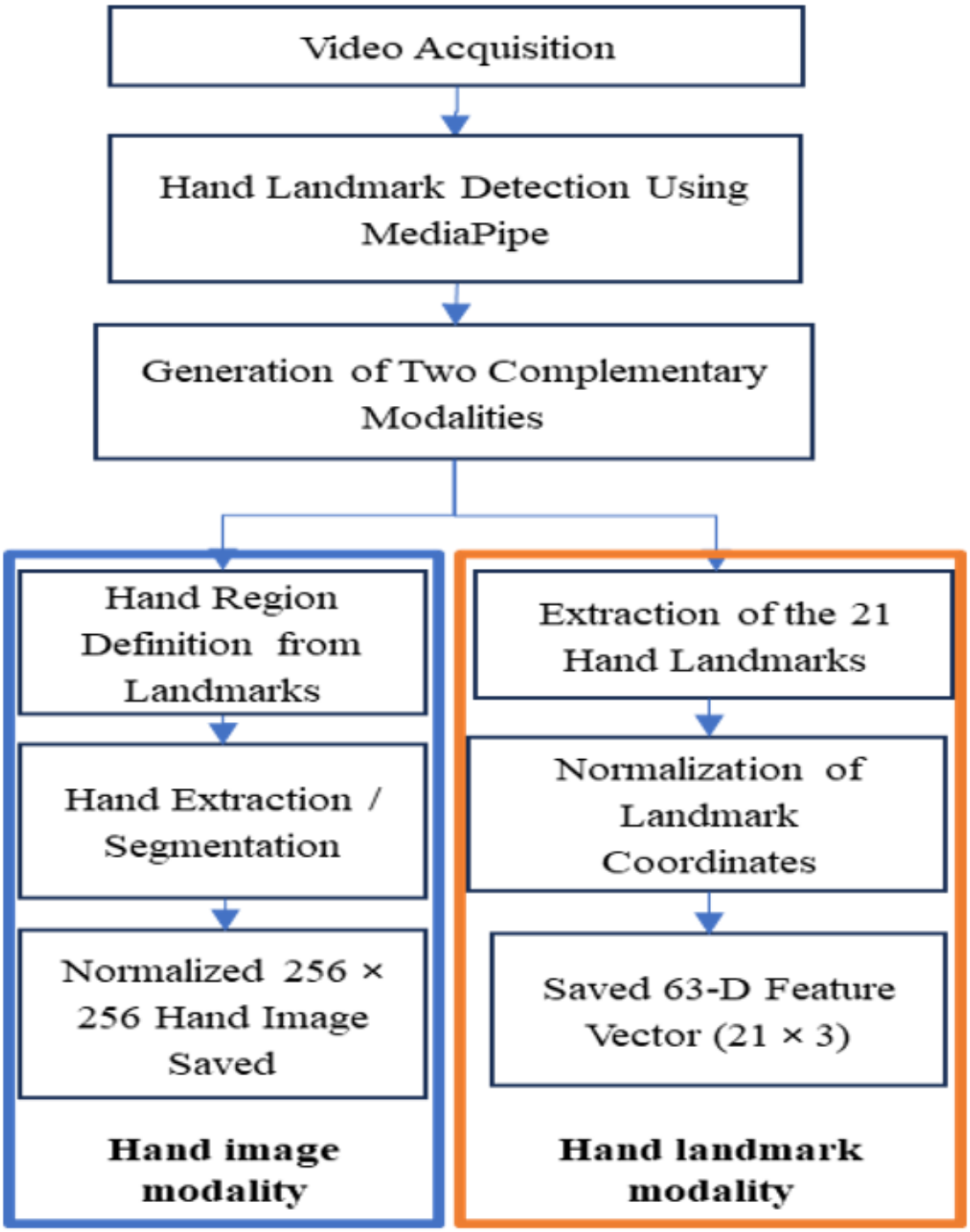

Figure 3.
Overview of the Multimodal Hand Data Acquisition Pipeline.

4. Materials and Methods

The image branch uses EfficientNet-B0 pre-trained on ImageNet [9].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

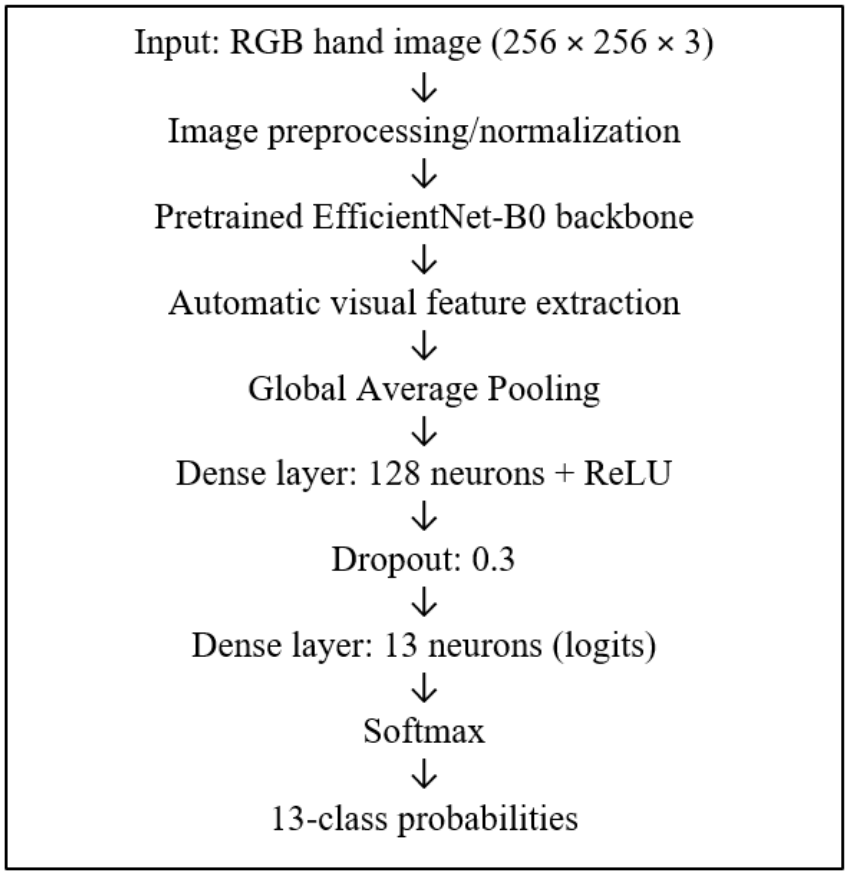

Figure 4.
EfficientNet-B0-Based Architecture for Static Hand Gesture Recognition.

Images are first cropped and stored at 256 × 256 pixels, then resized to 224 × 224 during training to match the backbone configuration. The training data undergoes slight augmentation, including small rotations, translations, scale changes, and photometric perturbations. This choice improves generalization without altering the meaning of the gesture.

The landmarks branch uses a multi-layer perceptron that takes a vector of 63 values as input. The architecture comprises fully connected layers of sizes 256, 128, and 64, with Batch Normalization, ReLU activation, and Dropout between each layer, as shown in Fig. 5.

This model is intentionally kept simple to effectively leverage an already highly informative geometric description of static poses, rather than maximizing complexity.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

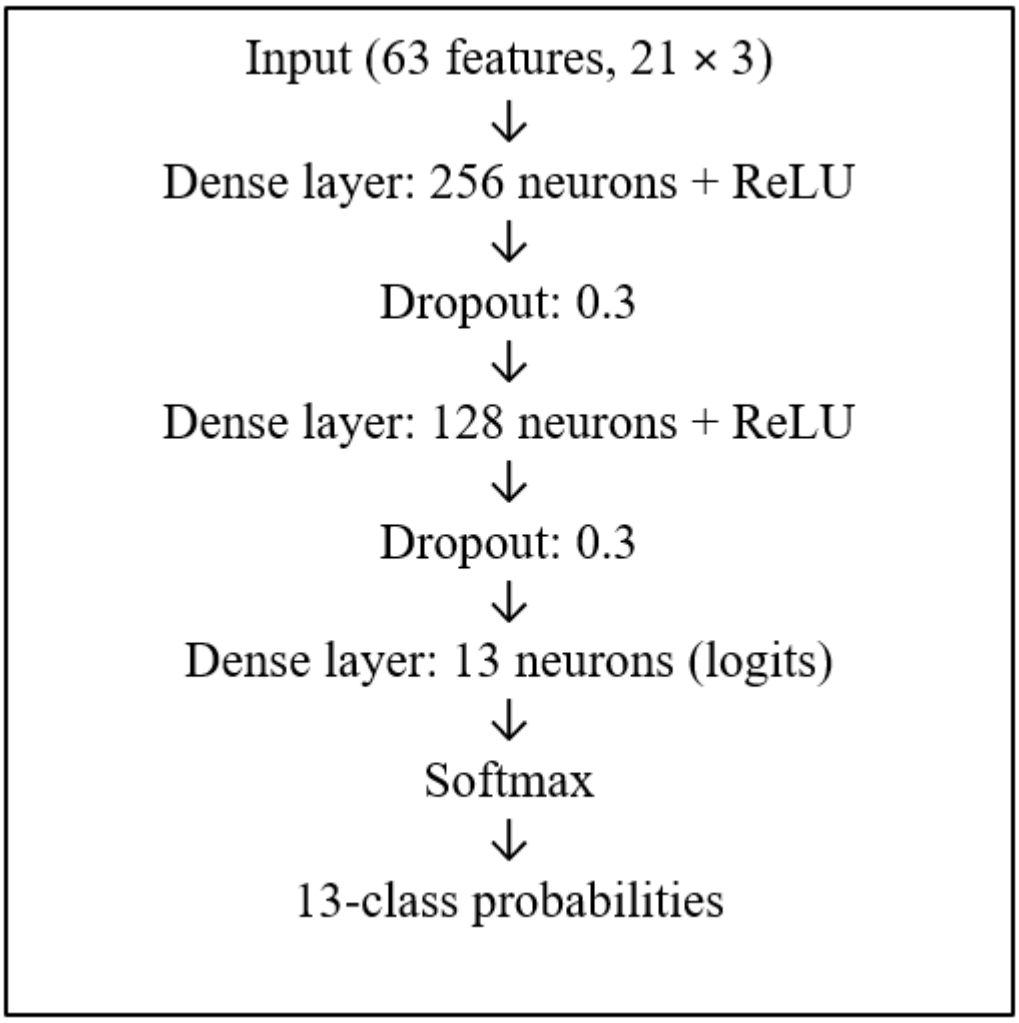

Figure 5.
MLP-Based Architecture for 63-D Hand Landmark Gesture Classification.

Both branches are trained separately using cross-entropy loss and the Adam optimizer. The main hyperparameters validated experimentally are: a batch size of 16, a maximum of 30 epochs for the image branch and 50 for the landmark branch, a patience of 8 epochs, a learning rate of 1e-4 for EfficientNet-B0 and 1e-3 for the landmark MLP, as well as a weight decay of 1e-4.

Multimodal fusion is performed at the decision level by taking the arithmetic mean of the output probabilities, using Eq. (1) as follows:

$P_{fusion} = 0.5\ P_{im} + 0.5\ P_{Lm}$ .....□□□

This simple choice avoids overcomplicating the architecture, keeps each modality independent, and makes the results easier to interpret. This simplicity also facilitates reproducibility and allows for a clear measurement of each branch's contribution. In deep learning, this design remains consistent with standard practices in modern pipelines.

Efficient convolutional backbones leverage advancements from residual networks, while the training of unimodal branches relies on well-established building blocks such as Batch Normalization, Adam optimization, Dropout, ReLU activations, and a more general theoretical framework for deep learning [29-31].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

5. Experimental Configuration

The evaluation is based on a 5-fold cross-validation across subjects. In each fold, four subjects are reserved for testing, and the remaining sixteen subjects are used for training, as shown in Fig. 6.

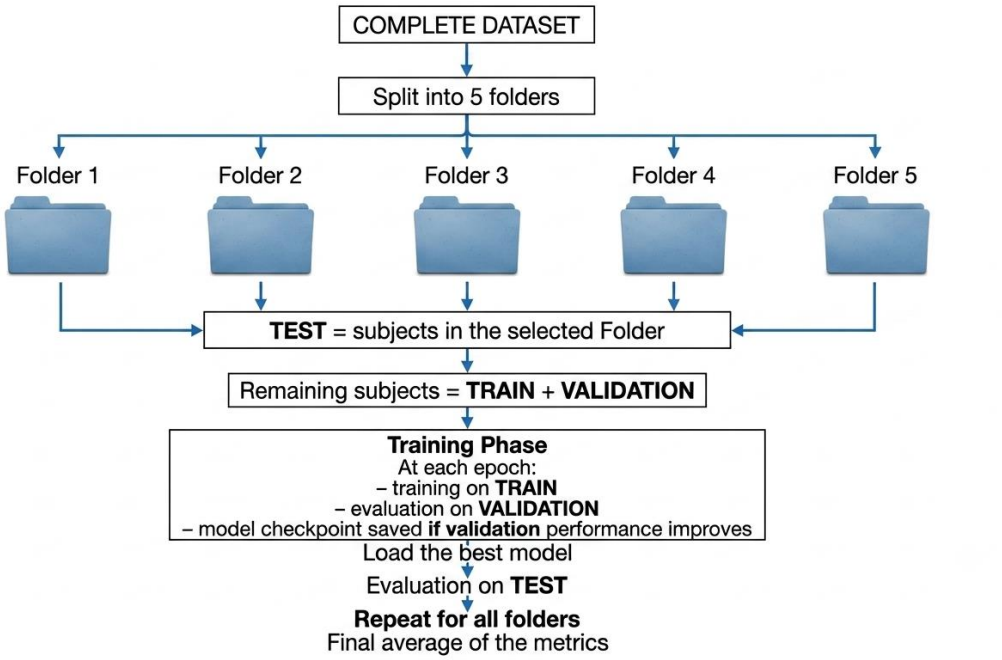

Figure 6.
Flowchart of the 5-Fold Cross-Subject: Training, Validation, and Testing Procedure.

Definition of folds (tested subjects):

- Fold 1: {S1, S2, S3, S4}
- Fold 2: {S5, S6, S7, S8}
- Fold 3: {S9, S10, S11, S12}
- Fold 4: {S13, S14, S15, S16}
- Fold 5: {S17, S18, S19, S20}

Within the training set, approximately 15% of the samples are reserved for validation to guide early stopping and the selection of the best models. This protocol eliminates any information leakage between samples and provides a realistic estimate of generalization ability. Three main metrics are reported: top-1 accuracy, Macro-F1, and top-2.

Accuracy measures the overall proportion of correct predictions. Macro-F1 evaluates the balance of performance across classes, which remains important even with a balanced corpus.

The Top-2 accuracy indicates how often the correct class is found among the two most likely classifications; this metric is useful for analyzing residual errors in a symbolic vocabulary where certain postures may be similar.

The choice of this protocol directly follows the recommendations of recent literature, which emphasizes the need to separate users between training and testing to avoid an artificial overestimation of performance [3, 4, 6].

6. Results and Discussion

The average results across the five folds first confirm that the recognition problem under study is well-defined: all three configurations achieve a high level of performance. However, the value of the

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

study lies not only in the absolute values of the scores, but in the differences between modalities and in the stability observed from one-fold to the next.

The image-only model achieves 96.36 ± 2.13% accuracy, 96.26 ± 2.29% Macro-F1 score, and 98.17 ± 1.74% Top-2 accuracy. These results show that a visual classifier based on EfficientNet-B0 already effectively captures the overall structure of postures, despite inter-subject morphological variability, as summarized in Table 2.

Table 2.
Average performance across 5 folds in cross-subject validation (Values are reported as mean ± standard deviation across folds; upper error bars were clipped at 100%.

Method	Accuracy (%)	Macro-F1 (%)	Top-2 (%)
Landmarks alone (MLP)	98.99 ± 1.30	98.98 ± 1.32	99.57 ± 0.86
Images only (EfficientNet-B0)	96.36 ± 2.13	96.26 ± 2.29	98.17 ± 1.74
Late fusion (0.5 / 0.5)	99.02 ± 0.90	99.00 ± 0.92	99.82 ± 0.32

The landmark-only model significantly outperforms the image-only model, with 98.99 ± 1.30% accuracy, 98.98 ± 1.32% Macro-F1, and 99.57 ± 0.86% Top-2 accuracy.

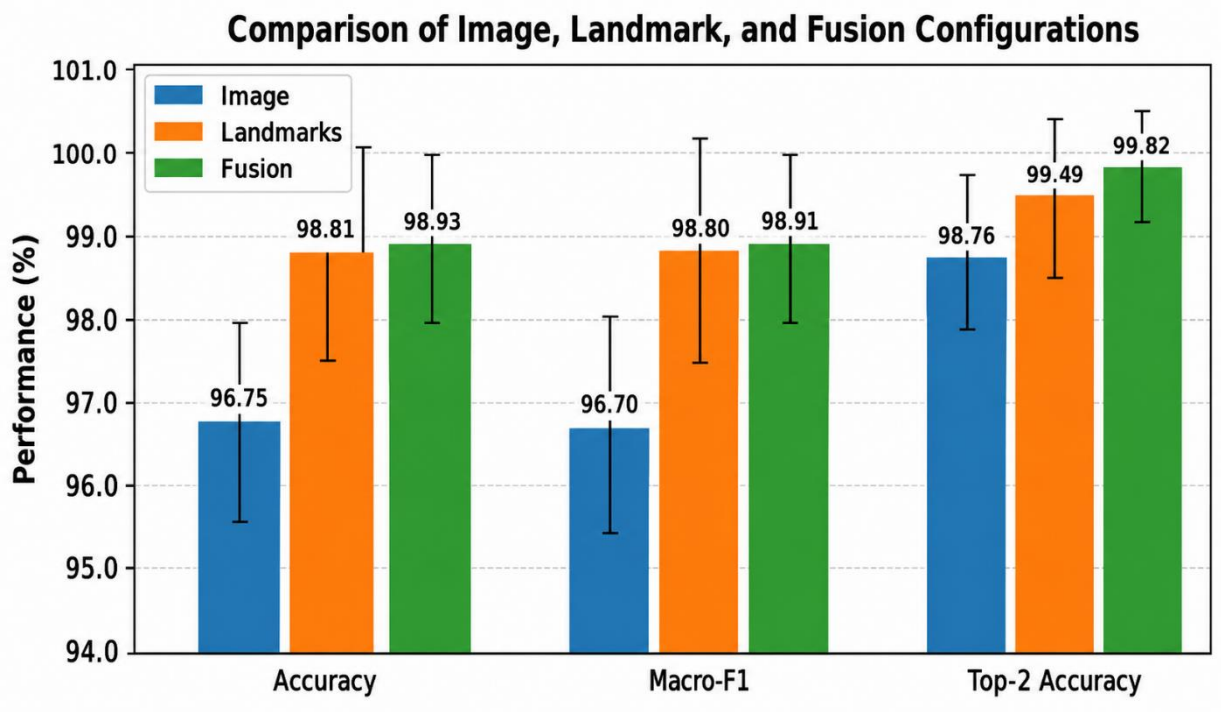

Figure 7.
Comparison of the performance of the three configurations (image, landmarks, fusion) in terms of accuracy, Macro-F1, and Top-2 accuracy.

This superiority indicates that, for static gestures, the normalized hand geometry constitutes the most directly discriminative source of information.

Late fusion provides the best overall performance with 99.02 ± 0.90% accuracy, 99.00 ± 0.92% Macro-F1, and 99.82 ± 0.32% Top-2 accuracy. The absolute gain relative to the landmark-only model remains moderate, which is expected given the very high initial performance of this approach. However, this gain is consistent and is accompanied by a reduction in standard deviation, suggesting improved stability across folds, as shown in Figure 7.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

Fold 5 clearly illustrates this trend. On this representative fold, the image branch achieves 98.67% accuracy, the landmarks branch 99.72%, while the fusion reaches 99.92%. Here again, the role of fusion is not to correct a weak system, but to consolidate an already high-performing system by absorbing some of the errors specific to each modality, as summarized in Table 3.

Table 3.
Representative observation of fold 5.

Configuration	Accuracy (%)	Top-2 (%)	Comment
Images only	98.67	99.60	Strong visual branch but inferior to landmarks
Landmarks only	99.72	99.96	Dominant geometric branch for static gestures
Late fusion	99.92	100.00	Best compromise between precision and robustness.

These results can be explained by several factors. First, a normalized representation of landmarks is well suited for recognizing static gestures. The joint structure provides more useful information than raw appearance, which remains more sensitive to acquisition conditions. The results also show that the image provides added value, even though landmarks remain the primary source of information.

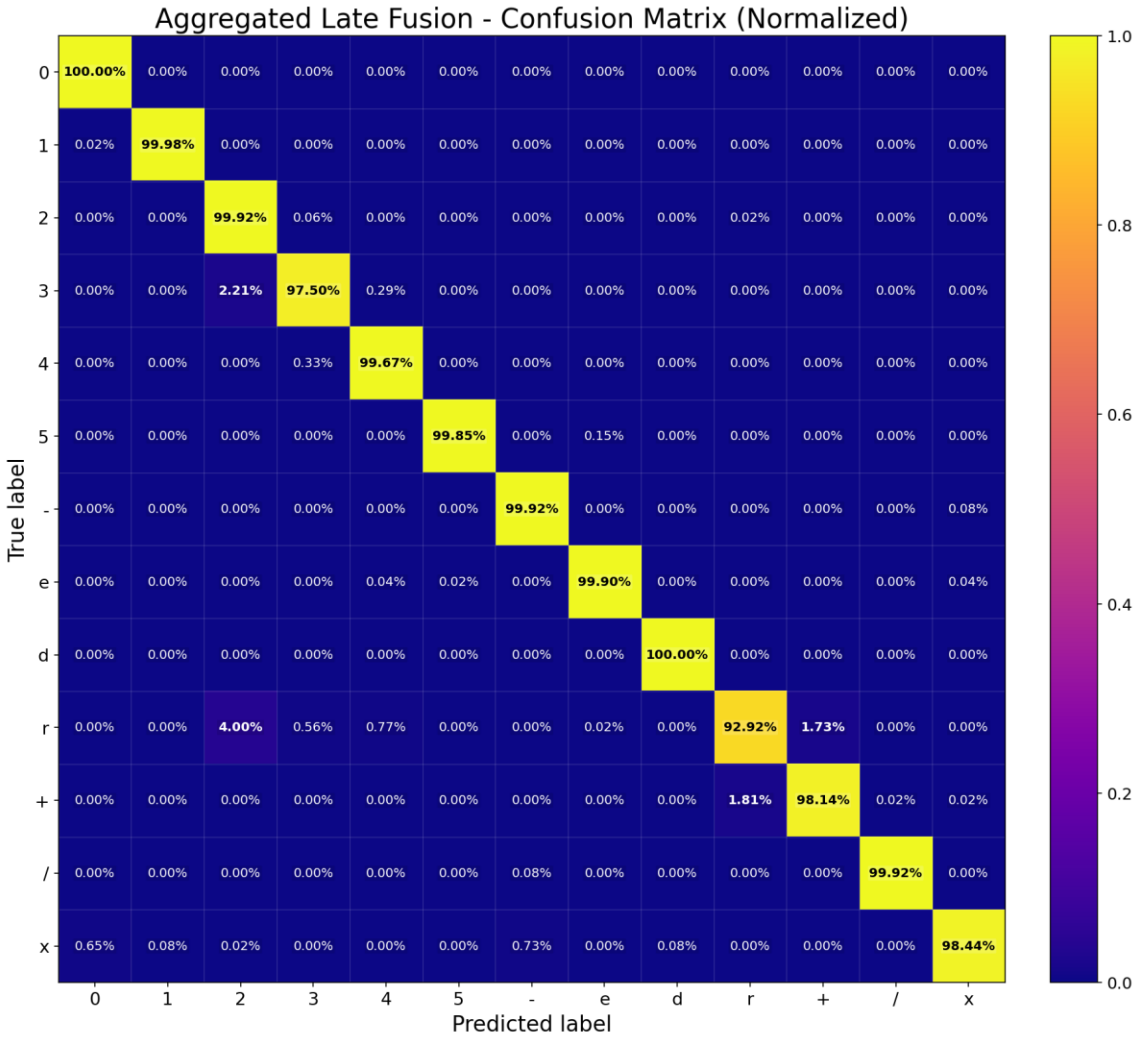

Figure 8.
Aggregated confusion matrix of the late fusion model.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

The improvement observed during fusion, though modest, suggests that certain geometrically ambiguous cases are better resolved using visual cues such as the outline of the hand, the thickness of the fingers, or configurations that are difficult to represent using coordinates alone. Furthermore, reducing variability between folds is important in practice. To ensure a natural interface, the model must maintain high average accuracy and remain stable with new users. Late fusion contributes to this stability and improves overall accuracy, as shown in Fig. 8.

Finally, these findings are consistent with recent literature. Reviews show that robustness depends on architecture, evaluation protocols, subject diversity, and the ability to utilize different sources of information [3, 6, 11]. The results presented here confirm this trend.

Although the results are encouraging, several limitations remain. The first concerns the static gesture vocabulary studied. The thirteen gestures analyzed fall under posture recognition, not continuous gesture recognition. Temporal transitions, intermediate movements, and dynamic gestures are therefore not taken into account, with the aim of reducing the gesture vocabulary for a truly natural interface. Regarding acquisition conditions, the corpus includes genuine variability across subjects and natural postural variations, but it remains more controlled than fully unstructured contexts. More complex backgrounds, extreme angles, significant occlusions, or scenes with multiple users would pose more challenging tests.

Another limitation concerns real-time performance. It is important to evaluate and adapt latency, which is essential for interactive applications.

7. Conclusion

This paper presents a multimodal method for a static gesture recognition system for a vision-based gesture calculator. To this end, a database was constructed containing 62,400 paired samples of 13 static gestures from 20 subjects (normalized RGB images and a vector of 63 coordinates for the 21 MediaPipe landmarks).

The main conclusion drawn from this is as follows: for a static gesture vocabulary, normalized 3D landmarks are the most informative modality. Nevertheless, fusion with the image provides a significant additional benefit and substantially improves inter-subject stability, as well as the average performance metrics of 99.02% accuracy and 99.00% Macro-F1 for late fusion, obtained using a strict five-fold cross-validation protocol, demonstrating that the proposed approach is both accurate and methodologically sound.

Beyond the scores, this work lays a solid foundation for robust, reproducible, and scientifically sound gesture-based interfaces. It paves the way for adaptive fusion, consolidated real-time evaluation, and expansion into advanced human-robot interaction applications.

The future directions for this work are clear. The first step will be to integrate a complete real-time evaluation system, including MediaPipe detection, cropping, inference of both branches, and fusion on the target platform. Next, it will be useful to explore adaptive fusion to adjust the use of the image and landmarks based on the quality of the capture. Another avenue is to extend the system to dynamic gesture recognition and human-robot interaction scenarios, where posture recognition could be combined with trajectory cues for more natural and less tiring commands. Finally, incorporating usability and cognitive-load metrics would allow for a better correlation between the algorithm's performance and the naturalness of the interaction.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302

References
[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artificial Intelligence Review*, vol. 43, pp. 1-54, 2015. https://doi.org/10.1007/s10462-012-9356-9

[2] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications," *Communications of the ACM*, vol. 54, no. 2, pp. 60-71, 2011. https://doi.org/10.1145/1897816.1897838

[3] R. Jalayer, M. Jalayer, C. Orsenigo, and M. Tomizuka, "A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction," *Robotics and Computer-Integrated Manufacturing*, vol. 97, p. 103110, 2026. https://doi.org/10.1016/j.rcim.2025.103110

[4] A. Kapitanov, K. Kvanchiani, A. Nagaev, R. Kraynov, and A. Makhliarchuk, "HaGRID--HAnd gesture recognition image dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[5] F. Zhang *et al.*, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020. https://doi.org/10.48550/arXiv.2006.10214

[6] E. Fertl, E. Castillo, G. Stettinger, M. P. Cuéllar, and D. P. Morales, "Hand gesture recognition on edge devices: Sensor technologies, algorithms, and processing hardware," *Sensors*, vol. 25, no. 6, p. 1687, 2025. https://doi.org/10.3390/s25061687

[7] K. Gao *et al.*, "Challenges and solutions for vision-based hand gesture interpretation: A review," *Computer Vision and Image Understanding*, vol. 248, p. 104095, 2024. https://doi.org/10.1016/j.cviu.2024.104095

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2018. https://doi.org/10.1109/TPAMI.2018.2798607

[9] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," presented at the International Conference on Machine Learning, 2019.

[10] R. Tripathi and B. Verma, "Survey on vision-based dynamic hand gesture recognition," *The Visual Computer*, vol. 40, pp. 6171-6199, 2024. https://doi.org/10.1007/s00371-023-03160-x

[11] C. Cui, M. S. Sunar, and G. E. Su, "Deep vision-based real-time hand gesture recognition: A review," *PeerJ Computer Science*, vol. 11, p. e2921, 2025. https://doi.org/10.7717/peerj-cs.2921

[12] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, pp. 345-379, 2010. https://doi.org/10.1007/s00530-010-0182-0

[13] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692-1706, 2015. https://doi.org/10.1109/TPAMI.2015.2461544

[14] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, "IPN Hand: A video dataset and benchmark for real-time continuous hand gesture recognition," presented at the 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4340–4347). IEEE, 2021.

[15] A. Caputo *et al.*, "SHREC 2021: Skeleton-based hand gesture recognition in the wild," *Computers & Graphics*, vol. 99, pp. 201-211, 2021. https://doi.org/10.1016/j.cag.2021.07.007

[16] M. Emporio, A. Ghasemaghaei, J. J. Laviola, and A. Giachetti, "Continuous hand gesture recognition: Benchmarks and methods," *Computer Vision and Image Understanding*, vol. 259, p. 104435, 2025. https://doi.org/10.1016/j.cviu.2025.104435

[17] P. Marques, P. Váz, J. Silva, P. Martins, and M. Abbasi, "Real-time gesture-based hand landmark detection for optimized mobile photo capture and synchronization," *Electronics*, vol. 14, no. 4, p. 704, 2025. https://doi.org/10.3390/electronics14040704

[18] J. Xie, Z. Xu, J. Zeng, Y. Gao, and K. Hashimoto, "Human–robot interaction using dynamic hand gesture for teleoperation of quadruped robots with a robotic arm," *Electronics*, vol. 14, no. 5, p. 860, 2025. https://doi.org/10.3390/electronics14050860

[19] T.-H. Nguyen, B.-V. Ngo, and T.-N. Nguyen, "Vision-based hand gesture recognition using a YOLOv8n model for the navigation of a smart wheelchair," *Electronics*, vol. 14, no. 4, p. 734, 2025. https://doi.org/10.3390/electronics14040734

[20] L. Fiorini *et al.*, "Daily gesture recognition during human-robot interaction combining vision and wearable systems," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23568-23577, 2021. https://doi.org/10.1109/JSEN.2021.3108011

[21] S. Zhang, H. Zhou, R. Tchantchane, and G. Alici, "Hand gesture recognition across various limb positions using a multimodal sensing system based on self-adaptive data-fusion and convolutional neural networks (CNNs)," *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18633-18645, 2024. https://doi.org/10.1109/JSEN.2024.3389963

[22] H. Abdelmoumene, L. Meddeber, O. Ghali, and Y. Amellal, "Natural human-machine interaction using static hand gestures for a gestural calculator system with DNN," *International Journal of Advanced Computer Science & Applications*, vol. 17, no. 1, p. 666, 2026. https://doi.org/10.14569/IJACSA.2026.0170164

[23] F. Al Farid *et al.*, "A structured and methodological review on vision-based hand gesture recognition system," *Journal of Imaging*, vol. 8, no. 6, p. 153, 2022. https://doi.org/10.3390/jimaging8060153

[24] K. Foteinos *et al.*, "Visual hand gesture recognition with deep learning: A comprehensive review of methods, datasets, challenges and future research directions," *arXiv preprint arXiv:2507.04465*, 2025. https://doi.org/10.48550/arXiv.2507.04465

[25] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, 2020. https://doi.org/10.3390/jimaging6080073

[26] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: A review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581-1606, 2024. https://doi.org/10.1007/s40747-023-01173-6

[27] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[28] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311-324, 2007. https://doi.org/10.1109/TSMCC.2007.893280

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. https://doi.org/10.1038/nature14539

[30] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," presented at the International Conference on Learning Representations (ICLR 2021), 2021.

[31] Y. Wu and T. S. Huang, *Vision-based gesture recognition: A review. In A. Braffort, R. Gherbi, S. Gibet, J. Richardson, & D. Teil (Eds.), Gesture-Based Communication in Human-Computer Interaction (Lecture Notes in Computer Science)*. Berlin, Germany: Springer, 1999.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 150-164, 2026
DOI: 10.55214/2576-8484.v10i7.13302