Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33625

Experimental Study on the Effects of Different Noise Types on Visual Gesture Recognition Performance in Complex Backgrounds

Yutong Liu

Leeds College, Southwest Jiaotong University, Chengdu, China
Sc232yl@leeds.ac.uk

Abstract. Gesture recognition has gradually entered the mainstream development direction of natural Human-Computer Interaction and is widely used in smart home devices, Virtual Reality technologies, Assistive systems, Intelligent mobile Devices, etc. Although the recognitions in real-world Visual Environment scenarios are frequently affected by complex background distractions, blurred movements, or missing Images. Based on the development of a test framework for evaluating how different factors affect model robustness using the HaGRID dataset. Three gesture classes, namely fist, like, and palm, are selected for experimentation. Since the original images contain considerable background content and the hand region occupies only a limited area, an automatic hand-cropping procedure is introduced before classification. A transfer-learning model built on MobileNetV2 is then trained, and three noisy test sets are constructed from the same clean cropped test set, corresponding to static clutter, dynamic interference, and missing information. The results show that the proposed preprocessing step improves recognition performance in complex scenes. On the clean test set, the model reaches an accuracy of 79.07% with a weighted F1-score of 0.84. Performance decreases under all noisy conditions. Among them, static clutter produces the largest decline, missing information has a moderate effect, and dynamic interference leads to the smallest reduction. In the class-level analysis, fist shows the strongest robustness, while palm is the most easily affected by noise.

Keywords: Gesture recognition, HaGRID, noise robustness, hand cropping

1. Introduction

With the rapid development of artificial intelligence, computer vision, and intelligent terminal technologies, human-computer interaction is gradually shifting from traditional keyboard-and-mouse input toward more natural and intuitive interaction modes. Because gesture recognition is contactless, direct, and convenient, it has shown broad application prospects in smart homes, virtual reality, in-vehicle interaction, rehabilitation training, and assistive healthcare [1-3]. From a technical perspective, gesture recognition mainly includes vision-based methods, wearable-sensor-based methods, and wireless-sensing-based methods. In comparison to other methods, vision-based gesture

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33625

recognition has relatively lower hardware requirements, abundant data sources, flexibility in application scenarios; It has thus received widespread attention in recent years [4-7].

Nevertheless, there are still many problems with its application in reality at present. Many of the existing researches have shown good recognitions under highly controlled environment or single background situations; However, their generalisation ability to more complicated Backgrounds is poor. Mainly due to the fact that disturbances in real images are no longer just random noises but rather result from various factors including background texture, foreground motion, local occlusion, imaging blurring and sampling errors simultaneously. These reasons impact the display effect and edges of gesture targets, reducing the effectiveness of feature learning by the model.

Disturbances in the visual gesture recognition problem are mainly divided into three types under complicated scenes. The first type of static background clutter arises primarily from stationary background objects, including walls, furniture, Doors and Windows, clothing texture etc. Interference of this kind increases the complexity of space in an image, reduces the difference between gesture edges and backgrounds; Therefore, it is more prone to learning features unrelated to gesture classes from the environment. The second is the dynamic random disturbance: moving foreground objects; non-target waving; semi-transparent occlusion; Local motion blur, etc. Such a disturbance changes the recognisable degree of the surrounding areas in this region to reduce the stability of target gesture recognition. The third is Information loss and sampling distortion: part occlusion; Reduced resolution, random crop, or omission of critical areas. Disturbances directly change the Shape of gesture and local Structure detail; They tend to affect those more sensitive to the position of fingers (sensory neurons) [8-10].

Although some researchers have proposed several methods to improve the problem of gestural recognition so far; However, there are still two unsolved problems now. The majority of existing research has only evaluated the recognition rate under clear conditions and lacked systematic tests in cluttered environments with multiple forms of interference. Second, different noise types may affect models through different mechanisms, yet relevant work often examines only overall accuracy changes and rarely analyzes category-specific differences under noise [5-7].

In this context, this paper will study visual gesture recognition under complex backgrounds and construct an experiment based on the HaGRID dataset with three types of noises: static clutter, dynamic interference and missing information. To deal with the problem that most of the original images have a high proportion of people's bodies and backgrounds but are difficult to find hands; Therefore, adding an automatic hand-cropping pre-processing method based on MediaPipe Hands. Meanwhile, MobileNetV2 serves as the classification back-end for balancing experimental speed and recognitions' accuracy. The main contributions of this paper are as follows:

(1) Complex-background gesture data are selected to construct a vision-based gesture recognition task closer to real-world scenarios;

(2) Automatic hand-cropping preprocessing is adopted to alleviate the problem that the background is too dominant and the hand region is too small in the original full images;

(3) Three noisy test sets are generated from the same clean test images, enabling fair comparison under different interference conditions;

(4) The model's noise robustness is analyzed from both overall performance and category-wise behavior, and the differences in the effects of different noise types on gesture recognition are summarized.

DOI: 10.54254/2755-2721/2026.GU33625

2. Related work

2.1. Overview of gesture recognition technology

Gesture recognition is an important branch of human-computer interaction, and existing studies have explored multiple modalities, including radar, Wi-Fi, electromyographic signals, and visual images. Ahmed et al. reviewed radar-based gesture recognition methods and pointed out both their potential advantages in contactless interaction and the modeling challenges they face in complex scenes [1]. Ahmed H. F. T. et al. summarized the development of device-free human gesture recognition using Wi-Fi CSI and emphasized that environmental interference and robustness are key challenges in wireless-sensing-based gesture recognition [2]. Related domestic research has also explored CSI-based gesture recognition and interference-robust feature modeling; for example, knowledge-distillation-based CSI gesture recognition methods have been proposed to improve performance in complex scenarios [3].

Compared with wireless sensing, visual gesture recognition depends more directly on image content. Reviews of visual gesture recognition indicate that the task usually involves detection and segmentation, analysis and modeling, and recognition and classification [4]. Research on dynamic visual gesture recognition further summarizes the pipeline into four key steps: gesture detection and segmentation, gesture tracking, feature extraction, and gesture classification, while noting that complex backgrounds, illumination variation, and noise can significantly affect the robustness of detection and segmentation [5,6].

2.2. Complex backgrounds and noise robustness

Complex backgrounds and noise have long been major challenges in visual gesture recognition. Several relevant studies have shown that under real-world conditions, background fluctuations such as tree shadows and variations in illumination levels due to different times and weather make hands more challenging to detect precisely than before during image acquisition alone fails to provide reliable data support [7].

In terms of image classification, noise and occlusion reduce the intra-class compactness and inter-class separation to degrade generalisation ability. In recent years, research on noise-resistant Image classification representation has also proven that Structure preservation and robust feature learning are important.

2.3. Lightweight models and hand detection methods

For the task of visual recognition in lightweight networks at this time have wide application ranges on edge devices and experiments to be fast enough for real-time processing needs. The research on MobileNetV2 shows that the lightweight performance of it mainly originates from inverted-residual bottlenecks, depthwise-separable convolution and adjustable-width multipliers to achieve a good tradeoff between efficiency and accuracy [11,12].

To extract the region of interest in the hand area, MediaPipe's hands module can be referenced to obtain high-precision hand keypoints and dynamic information. Some relevant research shows that this scheme is of high applicability for construction and deployment in RGB applications [13,14]. Therefore, based on the mediaPipe hands module in this paper will propose an automatic hand-cropping approach that emphasises the gesture region more strongly.

DOI: 10.54254/2755-2721/2026.GU33625

3. Dataset and experimental settings

3.1. Dataset selection and category screening

In this study, HaGRID will be used to obtain primary data. Compared with the dataset of simple background and high homogeneity, HaGRID has richer indoor backgrounds, human postures, illumination variations, as well as more complex environments; It is closer to real-world hand-drawing recognition scenarios.

Balance between interpretability and practicability, select three kinds of gestures in the data: fist; Like; Palm. These kinds of gestures' forms are typical cases. The fist is a closed-form gesture; has relatively obvious directional features, and the palm form of gesture depends on a more complete arrangement of fingers. Three classes have sufficient distinction; however, they are also somewhat related to each other, allowing for research on their categoriespecific recognition effects under different Noise Conditions.

3.2. Data split

Dividing the collected data into a training set, a validation Set and a clean test set respectively. The training set is used for model training; The validation Set monitors its accuracy while learning; The Clean Test Set evaluates it afterwards to ensure fairness. To ensure that the test set is independent of both the training and validation sets.

Based on the clean test set, three noisy test sets are further generated: a static clutter test set, a dynamic interference test set, and a missing-information test set. All three noisy test sets are generated one-to-one from the same clean test images, ensuring that the only varying factor in the comparison is the noise type.

3.3. Hand-cropping preprocessing

Initial experiments showed that directly classifying the original full images from HaGRID yielded poor performance. The reason is that the original images contain large areas of the human body and background, while the hand region that actually needs to be recognized is relatively small, causing the model to learn background features unrelated to the task.

To address this issue, MediaPipe Hands is used to automatically localize hands in the images. Based on the detected keypoints, a hand bounding box is computed, expanded, and cropped into a square region. For a small number of samples where detection fails, center cropping is used as a fallback strategy. The cropped images are uniformly rescaled to the same size before being used in training and testing. Experiments have shown that after hand-cropping, the proportion of the gesture area in the image has been effectively improved and background noise eliminated; this helps bring a more favourable condition for training to enter the system of gestures. Figure 1 shows that the cropped images have a larger area of hands and relatively fewer irrelevant backgrounds compared with the entire image at this stage.

DOI: 10.54254/2755-2721/2026.GU33625

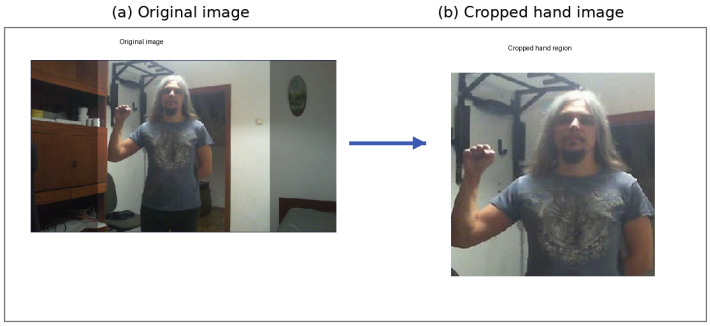

Figure 1. Illustration of hand-cropping preprocessing (comparison between the original image and the cropped result) (picture credit: original)

3.4. Noise modeling

To examine the robustness in response to various disturbances, three noised test sets were produced from the same clean-cropped test set: static clutter, dynamic interference, and missing data.

The static clutter is used to simulate strong-background-Texture enhancement, fixed occlusion, local Shadowing Increase background Complexity. Specifically, these include variations in brightness, Gaussian noise, stationery colour-block occlusion, and local shadow effects. The aforementioned type primarily disturbs changes in edges among the hand area and background pixels.

Dynamic Interference generates moving foreground objects, non-target behaviours, local motion blur, and partial occlusion of the background. Given that the experiments use static pictures, dynamic deviations are generally represented as stripe-style motion-blur, partial opacity of the background, and directionally scattered lines. Noise in this category is primarily local blurriness or obscured details in a small area.

The missing part of the information simulates local damage, decreased resolution, and incomplete image content. Random occlusion, downsample followed by upsample, and random crop followed by restoration back to the original size. This kind of noise directly destroys the local gesture structure and is more prone to impact classification that relies on finger-span information. Representative examples of the clean and noisy samples are shown in Figure 2.

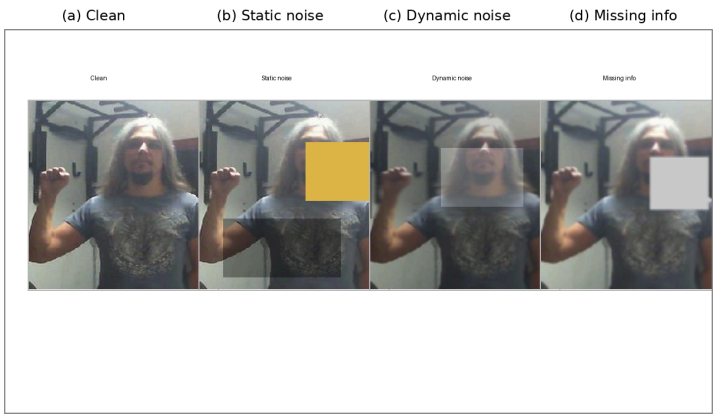

Figure 2. Illustration of the clean test set and the three noisy test sets (picture credit: original)

DOI: 10.54254/2755-2721/2026.GU33625

3.5. Model and parameter settings

The image classification module with transfer learning on the Edge Impulse platform is used, and the model structure is MobileNetV2 96×96 0.35. This model is a lightweight convolutional neural network with good training efficiency and deployment friendliness, making it suitable for rapid experimental validation.

The training parameters are set as follows: input size 96×96, learning rate 0.0005, 20 training epochs, data augmentation enabled, and 3 output classes. During the experiments, the training and validation sets are used in the training stage, while the final performance evaluation is uniformly conducted on the Model testing page to avoid mixing validation and test results.

4. Experimental results and analysis

4.1. Baseline on the clean test set

After hand-cropping preprocessing, the model achieves an accuracy of 79.07% and a weighted F1-score of 0.84 on the clean test set. These results indicate that, after focusing on the hand region, the model is able to distinguish the three gesture classes (fist, like, and palm) with relatively stable performance under complex-background conditions.

4.2. Overall results under different noise conditions

With the model kept unchanged, the three noisy test sets are evaluated separately. The results are shown in Table 1.

Table 1. Performance comparison under different test conditions

Test Set	Accuracy	Weighted F1
Clean	79.07%	0.84
Static Noise	64.94%	0.74
Dynamic Noise	74.51%	0.81
Missing Info	71.24%	0.79

The overall results indicate that all three noise types degrade model performance, but to different extents. Using the clean test set as the baseline, the accuracy drops are 14.13 percentage points for static noise, 4.56 percentage points for dynamic interference, and 7.83 percentage points for missing information. The corresponding relative declines are approximately 17.9%, 5.8%, and 9.9%, respectively. Therefore, under the current experimental setting, the severity of the three noise types can be ranked as follows: static clutter > missing information > dynamic interference. The overall comparison of Accuracy and Weighted F1 under the four testing conditions is shown in Figure 3.

DOI: 10.54254/2755-2721/2026.GU33625

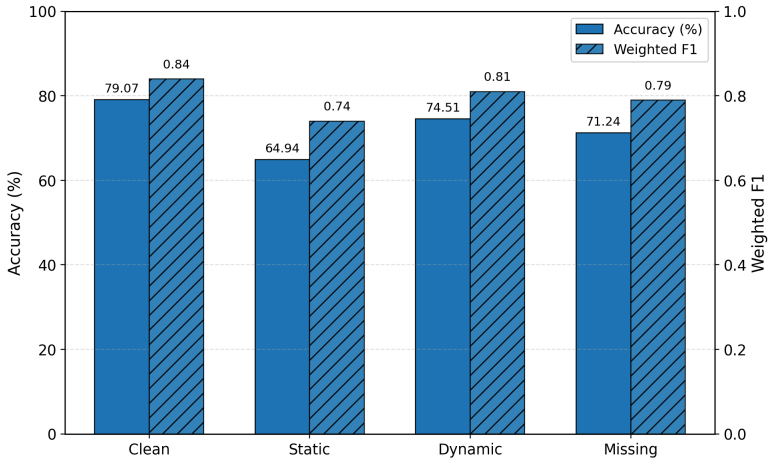

Figure 3. Comparison of Accuracy and Weighted F1 under different noise conditions (picture credit: original)

4.3. Category-wise analysis

4.3.1. Robustness of fist

In all test Conditions, FIST has relatively stable recognitional Accuracy to retain more noise tolerance. Because the first has a small whole shape and clear closed-form features. Despite obvious background clutter and partial structural damage, it can be determined that most basic forms of the target are recognizable quite accurately.

4.3.2. Stability of like

Maintains a relatively stable performance across the clean test set and all three noisy test sets. Given that this kind of gesture usually has clear directions, so as to facilitate recognitions through the model. Under noisy Conditions, it also shows some degree of misclassification or ambiguous prediction to varying degrees.

4.3.3. Sensitivity of palm

The palm class is the least robust among them. It is already starting to deteriorate slightly in good environment; The drop-off under static clutter and information deficiency will be more significant. This shows that open-handed movements are mainly related to the whole fingers, palm features and expansion effects.

4.3.4. Effect of static clutter

The experimental results show that static clutter has a greater impact on model performance than dynamic interference and missing information. This indicates that under the current task setting, fixed background textures, static occlusion blocks, and shadows are more likely to continuously damage gesture contours and edge information. By contrast, dynamic interference in this experiment is mainly implemented as local blur and local foreground disturbance, and its range of damage is relatively limited, resulting in a smaller overall effect.

DOI: 10.54254/2755-2721/2026.GU33625

Although missing information also damages local structures, its specific effect depends on the location and extent of the missing region, which does not always cover the most critical discriminative area. Therefore, its impact lies between that of static clutter and dynamic interference.

4.4. Result discussion

The experiments in this paper show that preprocessing has a significant influence on performance in complex-background gesture recognition tasks. Directly classifying the original full images results in poor performance of the model in learning gesture-specific features. After hand cropping, the result becomes much better, which shows that focusing on the hand region can help improve the model performance under complex backgrounds. The above conclusions are also in accordance with the proposal for target-region extraction before classification in complex-background gesture recognition research [7].

At the same time, different noise types do not affect the model in the same way. Stahty clutter has caused the greatest loss; therefore, background control, Occlusion Suppression and Key-Region Enhancement need to be particularly given attention during actual applications. The small effect size of dynamic interference suggests that the present model is somewhat robust to local blurring and minor Dynamic Disturbances.

Category-specific Results show that the geometry of gesture affects noise Resistance more prominently. closed-contour gesture is relatively stable; However, the spread-finger type or detailed Edge detection can be affected by noise. This finding is informative for class design and model optimization in future, more complex multi-class gesture recognition systems.

5. Conclusion

This paper investigates noise robustness in visual gesture recognition under complex backgrounds. Based on the HaGRID dataset, an experimental framework involving four conditions—clean testing, static clutter, dynamic interference, and missing information—is constructed, and a lightweight MobileNetV2 transfer learning model is used for validation. The experimental results show that hand-cropping preprocessing can significantly improve gesture recognition performance in complex backgrounds. The model achieves 79.07% accuracy and a weighted F1-score of 0.84 on the clean test set. All three noise types lead to performance degradation, among which static clutter has the largest effect, followed by missing information, while dynamic interference has the smallest effect. In addition, different gesture categories exhibit clear differences in robustness, with fist being the most stable and palm the most vulnerable.

In summary, robustness in complex-background gesture recognition depends not only on the network structure itself, but also on how the input region is represented and on the type of noise involved. Empirical analysis shows that the selection of pre-processing targets for regions in practice has helped raise the recognition accuracy of hand movements significantly higher. Nevertheless, this study only considers three gesture categories and noise simulation on static images. Future work can extend this framework to handle multiple gesture categories, disturbances from real videos, and strong models to enhance robustness comprehensively.

References

[1] Ahmed, S., Kallu, K.D., Ahmed, S. and Cho, S.H. (2021) Hand Gestures Recognition Using Radar Sensors for Human-Computer-Interaction: A Review. Remote Sensing, 13(3), 527.

DOI: 10.54254/2755-2721/2026.GU33625

[2] Ahmed, H.F.T., Ahmad, H. and Aravind, C.V. (2020) Device Free Human Gesture Recognition Using Wi-Fi CSI: A Survey. Engineering Applications of Artificial Intelligence, 87, 103281.
[3] Huang, Z., Zhu, H., Gong, H., Yang, M. and Wu, F. (2025) CSI Gesture Recognition Based on Knowledge Distillation. Piezoelectrics & Acoustooptics, 38(11), 1990-1999.
[4] Zhu, Y., Yang, Z., & Yuan, B. (2013, April). Vision based hand gesture recognition. In 2013 international conference on service sciences (ICSS) (pp. 260-265). IEEE.
[5] Hrishikesh, P., Akshay, V., Anugraha, K., TR, H. S., & Jyothisha, J. N. (2024). Vision based gesture recognition. Procedia Computer Science, 235, 303-315.
[6] Xiao, C. and Min, H. (2025) A Gesture Recognition Network Based on Global and Local EMG Feature Interaction. Control Theory & Applications, 42(3), 609-617.
[7] Jing, G., Cheng, J. and Ku, X. (2016) A Survey of Visual Gesture Recognition. Computer Science, 43(6A), 43-48.
[8] Xie, Y. and Wang, Q. (2021) A Survey of Vision-Based Dynamic Gesture Recognition. Computer Engineering and Applications, 57(22), 1-10.
[9] Wang, R., Wu, S., Zhang, M. and Wang, X. (2023) A Review of Vision-Based Neural-Network Methods for 3D Dynamic Gesture Recognition. Computer Science. https: //doi.org/10.11896/jsjkx.230200205
[10] Zhou, W. and Chen, K. (2022) A Lightweight Hand Gesture Recognition in Complex Backgrounds. Displays, 74, 102226.
[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4510-4520).
[12] Yang, J. (2025) Lightweight Network Optimization and Hardware Implementation Based on MobileNetV2. Master's Thesis, Heilongjiang University.
[13] Amprimo, G., Masi, G., Pettiti, G., et al. (2024) Hand Tracking for Clinical Applications: Validation of the Google MediaPipe Hand (GMH) and the Depth-Enhanced GMH-D Frameworks. Biomedical Signal Processing and Control, 96, 106508.
[14] Yerimbetova, A., Berzhanova, U., Sakenova, B., Milosz, M., Daiyrbayeva, E., Bayekeyeva, A., Mamyrbayeva, O. and Telman, D. (2026) Sign Language Recognition Based on Deep Learning via MediaPipe for People with Speech and Hearing Impairments. Procedia Computer Science, 275, 359-368.