2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Reliability of Mobile Camera-based Hand Sign Recognition in Outdoor Environments

Paula Stocco Raymond Kim Calvin Stahoviak Carol Young David Wood Tamzidul Mina

***Abstract*— Hand sign recognition systems have the potential to allow intuitive visual communication between humans and robots. Current recognition models often lack validation in outdoor settings, and thus fall short toward field deployment on mobile platforms. In this work, we assess the precision and recall of hand sign recognition models in the field considering robot movement, distance and viewing angle of the human from the mobile vision system. Supervised models were custom trained on skeletal hand data and their F1 scores were compared with various available pre-trained models at varying distances and viewing angles. Statistical analysis presented in this paper shows that the distance to subject from the vision system had a statistically significant impact on hand sign recognition in outdoor environments with mobile robots, while the impact of viewing angle remained insignificant with the models tested.**

I. Introduction

Mobile robots are capable of working alongside first responders, assist in search and rescue operations, and agricultural work. Their adoption in these areas is limited by the expertise required to control or direct them, and communication infrastructure requirements. Hand gesture-based control offers a simple, intuitive alternative high-level communication method for human-robot teams working together enabling reliable operation without conventional wireless radio or tethered communication systems. This method would also be preferred in densely populated areas with heavy radio traffic or interference.

Hand gesture recognition (HGR) methods have been widely proposed in previous works as summarized in Section II. However, its reliability in outdoor real-world conditions under varying lighting, distance, and viewing angles of the subject from the vision system remains undetermined, preventing wide-scale field deployment. Prior work on gesture recognition also predominantly focuses on using static cameras, limiting its range of applications. To our knowledge, no published studies are dedicated to developing and evaluating the reliability of HGR from freely moving cameras in real-world outdoor settings. Without proper validation in outdoor conditions, methods created for indoor environments cannot be assumed to work effectively in outdoor applications.

This work was supported by Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Fig. 1. Experiment setup to assess reliability of hand gesture recognition from a mobile robot, varying distance and viewing angle of a subject's hand from the mobile robot. A stabilized vision system mounted on the robot tracks and collects images of the subject as the robot moves over a prescribed path. The images are processed to infer hand sign as a function of distance and viewing angle of the hand from the robot.

As a step toward achieving consistent results in hand gesture-based communication in outdoor environments, in this paper we compare the performance of a number of pre-trained and custom trained HGR models using a carefully designed outdoor test protocol with a moving camera system varying distance and viewing angle of the subject. The custom trained models focused on fast, efficient model development for outdoor human-robot interaction, rather than requiring large datasets and extensive training pipelines. Thus, this work presents two key contributions.

First, a comparative evaluation of two model training methods for HGR in outdoor environments is presented. Our focus is on practical deployment, comparing how models are trained and which training strategy works best for outdoor HGR. The first method uses neural network architectures originally developed for classification tasks and trained on large indoor datasets, where subjects face the camera directly. The second method uses smaller classification models trained on skeletal landmarks as a lossy, compact representation of hand gestures captured in motion. By focusing on how these methods generalize to outdoor dynamic scenarios, we demonstrate that 1) general models trained on full images indoors may not necessarily perform well in outdoor environments, and 2) the skeletal landmark-based approach can be trained for outdoor mobile-robot applications. These findings provide a unique comparison to guide HGR training decisions for deployment.

Second, we compare these models specifically for outdoor mobile robotic applications. Performance is assessed against hand viewing angle and distance to the hand. Unlike previous studies, which have not systematically examined perspective variations in outdoor setting, our work provides

a methodological evaluation of their impact on HGR model performance.

Related vision-based hand sign recognition work is covered in Section II. The testing plan, including hand sign data collection and outdoor mobile robot validation, is detailed in Section III. Results are presented in Section IV with discussion and conclusions in Sections V and VI respectively. We believe these findings will encourage further exploration of visual human-robot communication and further steps towards broader and more widespread use.

II. Related Work

Despite the development of numerous HGR tools, significant usability challenges remain, such as fast response time, high recognition accuracy, ease of learning, and user satisfaction [1]. For effective HRI, these vision tools need to be robust to complex backgrounds and be capable of real-time operation [2]. In the case of wide open spaces, varying scales of gestures must be addressed, as discussed in [3].

Several studies on direct image classification have tested integration with hardware, such as [4] which demonstrated real-time physical HRI using convolutional neural networks (CNNs) for HGR with robotic arms. Research into improved robustness against hand gesture scale and complex backgrounds, as in [5] which presents multiscale feature learning networks, extends the state of the art. An attention-based single shot multibox detector network has also been proven to be effective for long-range HGR, extending the recognition distance to 7 meters from a flying UAV [6].

Gesture recognition methods that operate directly on full images use implicit feature extraction, prompting further research into understanding this process. Convolutional long short-term memory (LSTM) recurrent neural networks have learned gestures of varying duration and complexity by identifying key features [7]. Multimodal gesture recognition methods using 3-D convolutional and convolutional LSTM networks have demonstrated simultaneous learning of spatio-temporal features have improved gesture recognition [8]. Using hand recognition algorithms to propose an explicit region of interest is enough to improve accuracy with RGB-D based methods [9].

Other research approaches identify landmarks, specifically the articulating joints of the hand, before attempting to recognize gesture as precise explicit features to reduce the feature space. [10] introduced a neural network based on Symmetric Positive Definite (SPD) manifold learning with novel edges and features to improve accuracy and prevent overfitting. Spatio-temporal graph convolutional networks effectively learning spatio-temporal features from skeletal data [11]. however, the success of landmark processing methods depends on the accuracy of landmark detection. FastHand ia a fast monocular hand pose estimation method for embedded systems tested in real-time on hardware [12], although because of its efficient implementation and accuracy [13] is the API we chose.

This research adds to these works by addressing the effect of various factors on gesture recognition accuracy including the distance and speed of a moving camera, the need for image stabilization, lighting conditions, and changes in viewing angle from frontal to profile perspectives.

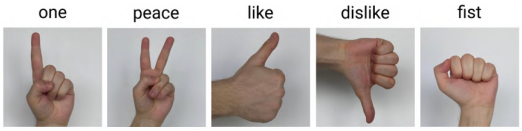

Fig. 2. Hand signs chosen for the reliability assessment of hand sign recognition. Images from HaGRID GitHub site, V2 [14].

III. Methodology

Five hand signs, shown in Fig. 2, were selected as target gestures for their potentially confounding characteristics. *One* and *thumbs up* test differentiation between the index finger and thumb, *dislike/thumbs down* tests differentiation of thumb direction relative to the wrist. *Two/peace* tests sensitivity to variations between the index and middle fingers, and *fist* assesses prediction accuracy from different viewpoints when landmarks are close together.

Training on a few select gestures test whether simpler, task-specific models improve real-time hardware control reliability compared to large general-purpose models trained on many gestures with full image inputs.

For the classification predictors, both pre-trained and custom models are used, described in Sections III-A and III-B, respectively. Publicly available models trained on large-scale image datasets were chosen because they leverage data that requires significant resources and time to collect and can be deployed immediately. Custom models were trained in order to assess whether teams can develop task-specific models with less data.

Finally, to evaluate the reliability of HGR methods from a moving camera, video was taken at varying distances and viewing angles outside, as detailed in Section III-C.

A. Existing Hand Sign Recognition Methods

HaGRID (Hand Gesture Recognition Image Dataset) is a repository of over half a million images of hand signs divided into 18 classes of gestures collected from $37,583$ unique subjects and pre-trained on this set [14]. HaGRID pre-trained models were chosen for using state-of-the-art architectures and being available online, and for covering a broad range of hand gestures rather than being customized for a specific use.

Available pre-trained gesture detection and classification models using this dataset include different versions of MobileNetV3 [15], ResNeXt [16] and ResNet [17]. MobileNetV3 is included in the study as it is optimized for embedded system applications making it a good fit for robotic applications; ResNeXt is included as it is reported to have the highest F1 score on the HaGRID dataset; and finally, ResNet is also included being the predecessor to ResNeXt for comparison.

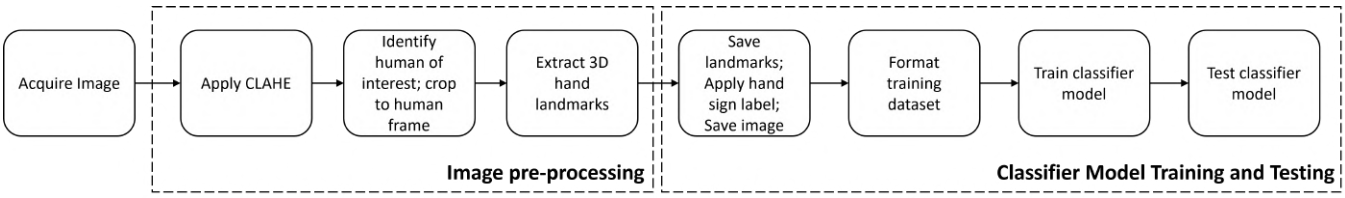

Fig. 3. Summary of image pre-processing, classifier model training and testing pipeline.

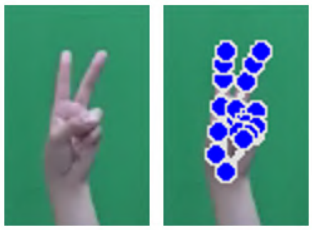

Fig. 4. Hand landmark detection of the *two/peace* sign in front of a green screen from collected training dataset.

B. Custom Hand Sign Recognition Model Training

A robust hand sign image dataset was collected from two volunteer participants with variable hand sign finger spacing and hand pose orientation. The training data collection procedure is presented in Section III-B.1.

The collected image datasets was preprocessed for color correction and cropped to the identified subject following the process outlined in Fig. 3. Contrast Limited Adaptive Histogram Equalization (CLAHE) was used on the validation data images to redistribute the luminance values of the image and improve contrast. The subject of interest was detected using YOLOv3 and each frame was cropped to the identified person bounding box. The cropped subject images for the validation dataset were saved for hand gesture classifier model validation. Yolov3 was used in this setup due to its proven stability, compatibility with existing legacy systems.

Google AI Edge MediaPipe [13] was used to identify hand landmarks from the dataset images. MediaPipe was chosen because of its mature application programming interface (API) which has been optimized for use on mobile devices, making it easy to use and suitable for running in real time. Detected hand landmarks from the images were labeled and saved for custom classifier model training.

In terms of custom classifier selection, three different models were chosen to make hand gesture predictions based on the 42 landmarks extracted from each image. These include classical classifiers such as k-nearest neighbor (KNC) and random forest (RFC), as well as a Multilayer Perceptron (MLP) supervised learning model. The custom MLP model consists of three fully connected layers with ReLU activations. The MLP's five logit outputs are used to assess confidence for each possible hand sign, with the highest confidence hand sign chosen as the final prediction.

1) Training Data Collection Protocol: The custom classifier models' training dataset was collected indoors under controlled 6500K lighting and a fixed uniform background.

Video of subjects seated in front of a green screen was captured using a GoPro Hero 11 action camera with image stabilization enabled at 60fps, 1080p resolution from a distance of 2 meters. Two equally spaced and independently controlled overhead lights, mounted on the ceiling, one on each side of the subject illuminated the hand.

For each hand sign shown in Fig. 2, the following timed sequence of prescribed motions were video recorded.

1) Hold up hand sign (natural pose) for 5s
2) If relevant, vary finger spacing continuously to the full possible range felt comfortable for 10s
3) Pitch hand pose about wrist continuously to the full possible range comfortable forward and backward for 10s
4) Roll hand pose about wrist continuously to the full possible range comfortable side-to-side for 10s
5) Yaw hand pose about wrist continuously to the full possible range comfortable clockwise-to-counter clockwise for 10s
6) Randomly change hand pose orientation in pitch, roll, yaw about wrist and change finger spacing for 10s.

The above sequence was repeated under four different lighting conditions: 1) no light, 2) left overhead light, 3) right overhead light, 4) both overhead lights switched on. The no light condition provided a low uniform lighting condition without any directional effects. Figure 4 shows an example hand with superimposed identified landmarks using Mediapipe from the training dataset.

C. Outdoor Testing Data Collection Protocol

Testing data was collected outdoors with natural lighting to emulate field deployment conditions, using a mobile robot to capture videos in motion. The test setup was designed to assess the effect of the subject's distance and viewing angle from the camera on hand sign recognition. The outdoor hardware testing setup is shown in Fig 5. A custom payload consisting of a stabilized linear FOV (field-of-view) GoPro 11 camera on an active gimbal was set to capture video at 60fps, 1080p resolution mounted on a SPOT quadruped robot. An AprilTag was also mounted for robot/camera position tracking from a remote RealSense D455 camera with a neutral density filter [18], [19], [20].

Outdoor test data was collected from 5 volunteer test subjects. Each subject was seated at the designated coordinate frame origin O_p, as shown in Fig. 6, and were instructed to hold up each of the five hand signs. The SPOT quadruped

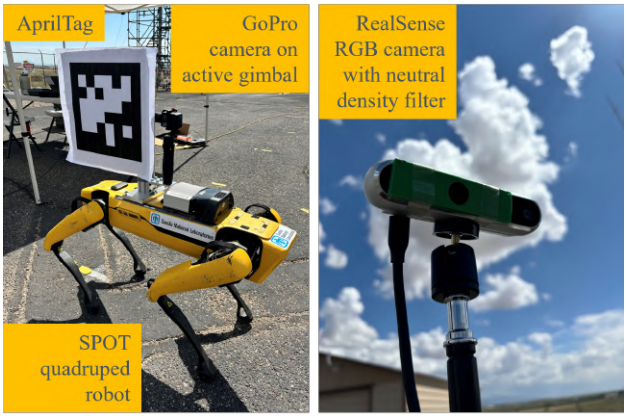

Fig. 5. A SPOT quadruped robot was mounted with a stabilized camera on an active gimbal for data collection. An AprilTag based tracking of robot position relative to the subject was set up using a remote RealSense camera.

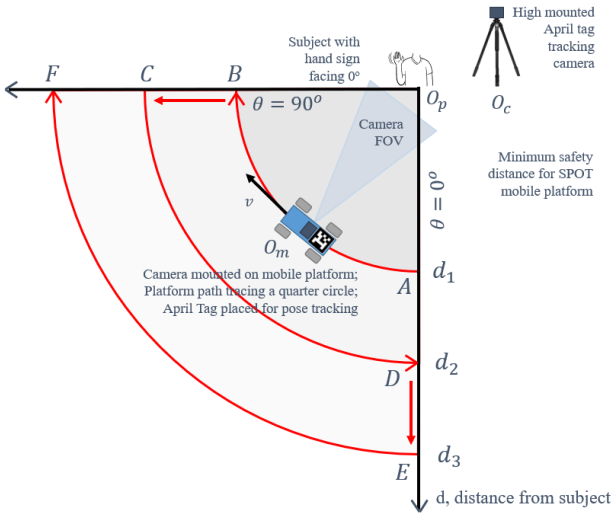

Fig. 6. Validation data collection setup and path tracing circuit ABCDEF.

robot was programmed to follow a prescribed path ABCDEF and collect video data of the hand sign at reference distances (radius) $d_{ref} = 1.83,\ 3.66,\ 5.49$ m from $\theta = 0^o$ (at subject front view) to 90^o (at subject side view) with real-time path tracking control using location feedback from the Apriltag tracking remote RealSense camera placed at a fixed location O_c relative to O_p. The minimum distance of 1.83 m was set based on the safe operating distance of the SPOT platform. The circuit was completed 3 times at 3 different speeds, $v =$, 0.3, 0.6, 0.9m/s for each hand sign. Sample views from the SPOT robot of the subject at different angles is shown in Fig. 7.

All experiments were conducted during the summer months in either sunny or partly cloudy weather conditions with clear visibility. Images collected outdoors for different subjects at different times of the day suffered from large variations in lighting due to the absence (partly cloudy conditions) or the relative position of the sun (morning, afternoon) to the subject resulting in inconsistent image

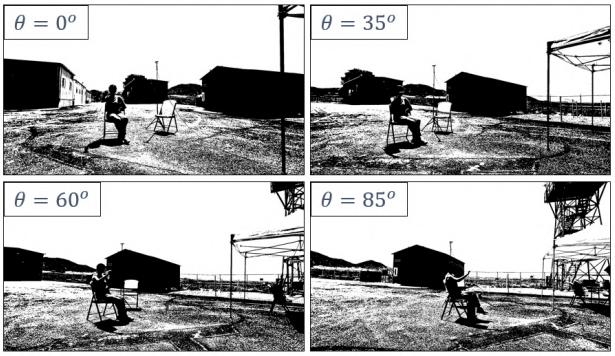

Fig. 7. SPOT robot view of participant holding up the *fist* hand sign over $0^o \leq \theta \leq 90^o$ circular motion at $d_2 = 3.66$ m.

Fig. 8. Hand landmark detection from the SPOT is achieved by first detecting the participant. The cropped human frame is used to detect the hand landmarks.

contrasts; the outdoor images were therefore preprocessed following the same procedure as the training dataset outlined in Fig. 3. CLAHE was used to enhance each image, then subjects were identified using YOLOv3 and the image was cropped to the human frame. The corrected images were classified by the pre-trained models. For the custom models, MediaPipe was first used to find landmarks for classification which were then used for prediction; Fig. 8 shows landmark detections on images taken from SPOT.

IV. Results

Various evaluation metrics are commonly used to compare classifiers. Accuracy measures the proportion of correctly classified instances but does not account for class imbalance or prediction bias; precision and recall assess the correctness and completeness of positive predictions, respectively. The F1 score is used in this analysis as a balanced measure of precision and recall.

To quantitatively analyze the effect of viewing distance and angle, the outdoor validation data was divided into bins. The quadrant representing the robot's prescribed path, defined by the range $0^o \leq \theta \leq 90^o$ and $0 \leq d_{ref} \leq 5.49$ m, was partitioned into 18 bins. Each bin corresponded to $\theta = 15^o$ in the range $0^o \leq \theta \leq 90^o$ in viewing angle at reference distances $d_{ref} = 1.83,\ 3.66,$ and 5.49 m. Images from all subjects were then assigned to these bins based on the SPOT's polar coordinates at the time of image capture.

For this study, video was collected at various robot speeds using a GoPro Hero 11 action camera equipped with internal image stabilization and mounted on an active gimbal. Image sharpness quantified by the variation of the image Laplacian [21] at different camera shutter speeds and mobile platform

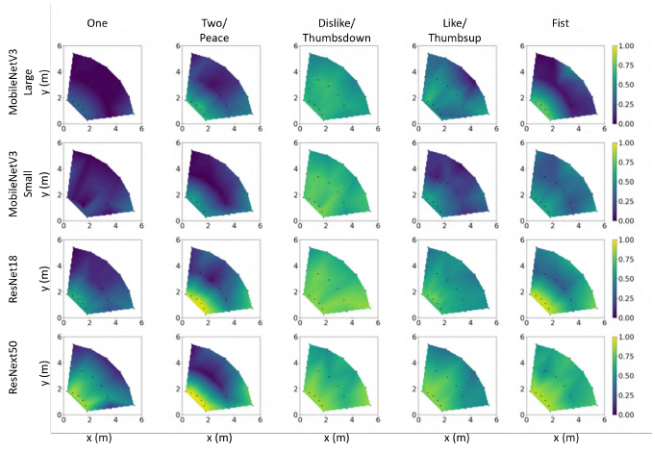

Fig. 9. Hand sign classification F1 score comparison heat map of various available pre-trained classification models using the Hagrid dataset relative to the origin O_p.

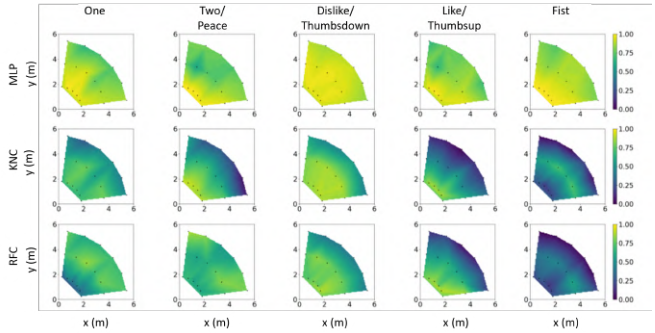

Fig. 10. Hand sign classification F1 precision score comparison heat map of custom trained classification models using the collected training dataset relative to the origin O_p.

speeds (0.6, 0.9, 1.2m/s) remained consistent at all three tested mobile platform speeds with higher shutter speeds. Hand signs were detected in all cases, suggesting that the range of mobile platform speed tested in this study did not affect hand sign recognition.

A. Hand Gesture Recognition Reliability

The F1 scores for each bin were calculated as a measure of hand sign recognition reliability. These scores were computed for every combination of hand sign and classifier model and visualized as heat maps for the available pre-built classifiers and the custom built classifiers, shown in Fig. 9 and 10 respectively. Distances and angles shown are relative to the human subject on the xy plane with the subject sitting at the origin O_p directly facing the y-axis.

The relative performance among the pre-built classifier models generally agree with the HaGRID dataset results. The ResNet and ResNeXt pre-built classifiers have higher F1 scores over the tested range compared to the MobileNetV3 models, with ResNeXt having higher scores than ResNet. F1 scores for ResNeXt were generally higher at greater distances from the subject for the 'one', 'thumbsup/like', and 'fist' hand signs, while scores for 'peace' and 'thumbsdown' remained similar across distances.

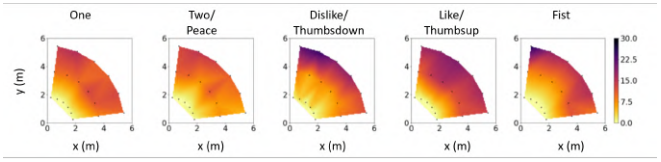

Fig. 11. Percentage of images unable to detect a hand landmarks using MediaPipe with distance and viewing angle relative to the origin O_p..

For the custom classifiers, the KNC and RFC achieved their highest F1 scores for the hand signs 'one', 'two/peace', and 'thumbs down/dislike' even as the distance from the subject increased, outperforming 'thumbs up/like' and 'fist'. In comparison, the MLP classifier consistently produced higher F1 scores even at further viewing distances for all hand signs, suggesting that of all the tested models, it maybe the most suitable for outdoor deployment.

The performance of custom skeletal data classifiers relies on the performance of the landmark identifier used. Fig. 11 illustrates MediaPipe's landmark detection performance for each hand sign during outdoor validation testing, with darker colors indicating a lower percentage of images with landmarks detected. MediaPipe Hands was trained on both real and generated hand images, utilizing a palm detector, feature extractor, and custom CNNs. As the viewing distance increased, a higher percentage of images lacked landmark detection, preventing classification, similar to patterns observed with image-based pre-trained classifiers.

B. Statistical Analysis

The F1 score heat maps presented in Section IV-A suggest that for most of the tested hand gesture classifier models and hand sign pairs, the HGR accuracy was impacted more by the distance to the human subject than viewing angle.

Ordinary Least Squares (OLS) regression models were fitted to the set of F1 scores (dependent variable) against the mobile camera distance (R) and viewing angle (θ) (independent variables) for all bins, considering each combination of classifier model and hand sign. The calculated p-values are used to assess the relative effect of these factors on different models and hand signs, where $\alpha \leq 0.05$ is set as the threshold for statistical significance.

The p-values obtained for the viewing angle (θ) in the range $[0^o, 90^o]$ consistently remained above the $\alpha = 0.05$ threshold for all model and hand sign pairs suggesting no statiscal significance of viewing angle on the F1 score, except for ResNet18 (thumbs down) with $p = 0.023$. p-values for viewing angles are therefore omitted for brevity.

Table I summarizes OLS regression results with F1 score by bin as the dependent variable and distance R, and each classifier-hand sign pair as independent variables. $p \leq 0.05$ values indicating a rejection of the null hypothesis are presented in bold. Results show that R significantly affects HGR reliability for almost all model-hand sign pairs except for the fist, consistent with the heat map observations made in Section IV. The F1 score for hand sign 'one' with

MobileNetV3 (MNV3) Small remained relatively low and uniform across the range of R as observed in Fig. 9; the p-value obtained was high for this hand sign-model pair. Further investigation is required to understand why the F1 score remained uniformly low with R. The 'peace' hand sign produced a low p-value for model RFC, but was not statistically significant.

TABLE I
OLS REGRESSION p-VALUES FOR F1 SCORE BY HAND SIGN AND MODEL. BOLD VALUES INDICATE STATISTICAL SIGNIFICANCE ($p \leq 0.05$).

Hand Sign / Model	One	Peace	Thumbs Up	Thumbs Down	Fist
MNV3 Large	**0.003**	**0.000**	**0.050**	**0.000**	**0.001**
MNV3 Small	0.949	**0.027**	**0.001**	**0.005**	0.085
ResNet18	**0.000**	**0.004**	**0.000**	**0.002**	**0.024**
ResNext50	**0.000**	**0.010**	**0.000**	**0.005**	**0.001**
MLP	**0.000**	**0.001**	**0.000**	**0.009**	**0.000**
KNC	**0.000**	**0.000**	**0.000**	**0.000**	0.305
RFC	**0.001**	0.121	**0.000**	**0.000**	0.097

C. Cross Validation

Each pre-trained and custom classifier was evaluated on the HaGRID 512px lightweight version of the full dataset as an assessment on datasets where the baseline models are expected to perform optimally. 10,000 images from each of the same five hand signs tested during validation (see Fig. 2) were selected as a representative subset. The HaGRID pre-trained models demonstrated consistent relative performance as in the HaGRID full dataset study, shown in Table II. Notably, the MLP model maintained a high F1 score on these unseen images, which were captured independently in untested indoor environments.

TABLE II
F1 SCORES OF HAGRID DATASET PREDICTIONS BY ALL MODELS.

Model	F1 Score
MobileNet Large	0.63
MobileNet Small	0.52
ResNet18	0.84
ResNext50	0.93
MLP	0.81
KNN	0.41
RFC	0.31

V. DISCUSSION

Several factors affect HGR and may account for the observed results of the validation studies. The accuracy of hand sign classification could be impacted by the effective similarity between different signs. For example, the 'thumbsup/like' sign may have easily differentiable distinct features, whereas the variability in distance between 'two/peace' sign's upright fingers may add confusing variability. Further topographical analysis could provide further insight.

The quality of images on mobile platforms may be lower than images used in training, whether with pre-trained or custom models. In this study. the mobile camera setup was limited to operate at a 1080p resolution, and captured images from 1.83 m to 5.49 m. Future work on determining the effects of higher resolution cameras on model performance must be explored.

Motion blur is an important factor that can degrade image and therefore HGR quality. The SPOT robot during motion also generated a fair amount of vibrations during each leg impact with the ground. The set up used in this study utilized the built-in image stabilization feature of the GoPro 11 camera along with an active gimbal to reduce motion and vibration effects on the image acquisition. Effects of higher platform speeds or traversal over rough terrain must be must of accounted for in the vision system setup on a robot platform with proper vibration isolation and image stabilization.

For HaGRID pre-trained models, which were trained on still images taken closer to the gesturing hand, further work could examine the dependence of F1 scores on distance with higher-resolution cameras, particularly for large-scale field deployments, to determine whether any distance-related issues can be mitigated. Furthermore, the HaGRID models were trained with thirteen hand signs in addition to the five hand signs chosen in this study. Both ResNet and ResNeXt models showed the most significant drop in F1 scores, potentially from confounding results with other hand signs. However, we note that, despite only being trained on two subjects, our skeletal based models performed better than the vision based models on five different subjects used in outdoor training and the 538 unique subjects from the indoor dataset. This cross-validation experiment measures the generalizability of our approach, even with little diversity in the training set for skeletal landmarks, there is sufficient similarity in the anatomical ratio across human hand-structures that the model generalizes to a wide variety of hands.

VI. CONCLUSION

Achieving secure and dependable communication between humans and robots is crucial for partnership in the field. The findings provide valuable insights into the limitations faced in current outdoor implementation on mobile robots, such as the need for image/video stabilization and challenges posed by outdoor lighting conditions, and strategies for training and updating HGR methods. During outdoor validation, viewing distance—rather than viewing angle—had a noticeable effect on model reliability across all tested configurations. Both training data and model architecture affect performance, with larger models trained on more extensive datasets not necessarily yielding improved scores, especially if they are trained for general use with many hand signs rather than focusing on specific cases.

While not an exhaustive benchmark of all published models, we address the key trade-offs between training efficiency, model size, and real-world adaptability. Several promising directions for future research could build on this work, such

as exploring the impact of model size and processing time during real-time use, which remains unaddressed here. We anticipate these findings will highlight key open challenges and inform future research and field deployment efforts in the future.

REFERENCES

[1] 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.

[2] Z. Xia, Q. Lei, Y. Yang, H. Zhang, Y. He, W. Wang, and M. Huang, "Vision-based hand gesture recognition for human-robot collaboration: a survey," in *2019 5th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2019, pp. 198–205.

[3] E. Bamani, E. Nissinman, I. Meir, L. Koenigsberg, and A. Sintov, "Ultra-range gesture recognition using a web-camera in human–robot interaction," *Engineering Applications of Artificial Intelligence*, vol. 132, p. 108443, 2024.

[4] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, and A. Cherubini, "Towards real-time physical human-robot interaction using skeleton information and hand gestures," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–6.

[5] H. Liang, L. Fei, S. Zhao, J. Wen, S. Teng, and Y. Xu, "Mask-guided multiscale feature aggregation network for hand gesture recognition," *Pattern Recognition*, vol. 145, p. 109901, 2024.

[6] L. Zhou, C. Du, Z. Sun, T. L. Lam, and Y. Xu, "Long-range hand gesture recognition via attention-based ssd network," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1832–1838.

[7] E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," *Neurocomputing*, vol. 268, pp. 76–86, 2017.

[8] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-d convolution and convolutional lstm," *Ieee Access*, vol. 5, pp. 4517–4524, 2017.

[9] X. Ma and J. Peng, "Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information," *J. Sensors*, vol. 2018, no. 1, p. 5809769, 2018.

[10] X. S. Nguyen, L. Brun, O. Lézoray, and S. Bougleux, "A neural network based on spd manifold learning for skeleton-based hand gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 036–12 045.

[11] Y. Li, Z. He, X. Ye, Z. He, and K. Han, "Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition," *EURASIP Journal on Image and Video Processing*, vol. 2019, pp. 1–7, 2019.

[12] S. An, X. Zhang, D. Wei, H. Zhu, J. Yang, and K. A. Tsintotas, "Fast-hand: Fast monocular hand pose estimation on embedded systems," *J. Systems Architecture*, vol. 122, p. 102361, 2022.

[13] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand trackings," Fourth Workshop on Computer Vision for AR/VR (CV4ARVR), Google, 2020. [Online]. Available: https://arxiv.org/abs/2006.10214

[14] 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, pp. 4572–4581.

[15] A. Howard, R. Pang, H. Adam, Q. Le, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, and Y. Zhu, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2019, pp. 1314–1324.

[16] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 07 2017, pp. 5987–5995.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] M. Kalaitzakis, B. Cain, S. Carroll, A. Ambrosi, C. Whitehead, and N. Vitzilaios, "Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers," *J. Intelligent & Robotic Systems*, vol. 101, pp. 1–26, 2021.

[19] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4193–4198.

[20] M. Krogius, A. Haggenmiller, and E. Olson, "Flexible layouts for fiducial tags," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1898–1903.

[21] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3. IEEE, 2000, pp. 314–317.