2024 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
July 15-19, 2024. Boston, MA, USA

2024 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM) | 979-8-3503-5536-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIM55361.2024.10637093

IHSR: A Framework Enables Robots to Learn Novel Hand Signals From a Few Samples

†Kanchon Kanti Podder, *Jian Zhang, *Yongshuai Wu
†Department of Electrical and Computer Engineering, Kennesaw State University, Marietta, GA 30060, USA
*Department of Information Technology, Kennesaw State University, Marietta, GA 30060, USA
Email: kpodder@students.kennesaw.edu, jianzhang@ieee.org, ywu26@students.kennesaw.edu

***Abstract*—This project introduces a framework to enable robots to recognize human hand signals, a reliable and feasible device-free means of communication in many noisy environments such as construction sites and airport ramps, to facilitate efficient human-robot collaboration. Various hand signal systems are accepted in many small groups for specific purposes, such as Marshalling on airport ramps and construction site crane operations. Robots must be robust to unpredictable conditions, including various backgrounds and human appearances, an extreme challenge imposed by open environments. To address these challenges, we propose Instant Hand Signal Recognition (IHSR), a learning-based framework with world knowledge of human gestures embedded, for robots to learn novel hand signals in a few samples. It also offers robust zero-shot generalization to recognize learned signals in novel scenarios. Extensive experiments show that our IHSR can learn a novel hand signal in only 50 samples, which is 30+ times more efficient than the state-of-the-art method. It also demonstrates a robust zero-shot generalization for deploying a learned model in unseen environments to recognize hand signals from unseen human users.**

***Index Terms*—Zero-shot Generalization, Gesture Recognition, Hand signals, Cross-modality Embedding**

I. Introduction

Significant advancements in recent robotics enable more and more robots to be deployed in open and human-shared environments in the short future [1], [2]. This trend demands effective communication among robots and non-technique people to ensure safety and foster efficient human-robot collaboration. Visual hand signals are essential in many circumstances of our daily lives and workspaces, especially in noisy environments such as construction sites and airport ramps. Their reliability, feasibility, and device-free make visual hand signals one of the best forms of interaction between humans and robots in loud and voice-prohibiting environments.

In recent years, we have seen more and more research [3]–[6] focused on the recognition of hand signals, such as the Air Marshalling signal. For example, the authors in [3] used Convolutional Pose Machines and learning classifiers to segment and classify gestures with 97% accuracy despite dataset diversity and ambient circumstances. Another study [4] introduced OpenMarshall, an 88% accurate gesture-tracking model for aircraft Marshalling signal recognition. This model struggled with Marshalling signal variability and environmental effects on signal visibility. Despite these advances, existing studies highlight several limitations, including hardware configuration dependence, the need for more diverse and extensive datasets to improve model robustness and accuracy, and poor generalization across novel conditions and environments (including diversities in backgrounds, human appearances, motion patterns, etc.). Furthermore, this is a challenging and open problem in robotics and human-robot collaboration systems. First, there are lots of different hand signal systems that are accepted in many small groups for specific purposes, such as Marshalling on airport ramps and construction site crane operations. Even the same type of hand signals have several variants, and many new customized hand signals have also been introduced by various groups for their requirements. This situation makes it an infeasible mission to design a method to recognize and react to all hand signals.

To address the above challenges, this work introduces Instant Hand Signal Recognition (IHSR), a learning-based framework with world knowledge of human gestures, to empower robots to communicate with humans with hand signals by learning from very few examples. The proposed IHSR comprises a Cross-modality Embedded Network for Gesture Recognition (CENGeR) to capture the temporal and spatial information of human gestures efficiently. We also proposed a self-supervised pre-training process to embed world knowledge of various human gestures by utilizing the massive task-unrelated gesture dataset from the internet. The main contribution of our work is summarized as follows:

- Our proposed method significantly improves the training efficiency by deploying a self-supervised pre-training process. Experiments show that it only consumes about 50 samples to learn a new hand signal.
- The embedded world knowledge of human gestures in our proposed method offers a supernal zero-shot generalization to handle domain shift, which is a problem arising in significant discrepancies between its application and training domain. Our learned model is robust to a variety of novel conditions, including new environmental backgrounds, different persons, motions, and more.
- We conduct rigorous experiments to enable a real mobile robot to recognize and react to tailored marshaling-liked hand signals.

The remainder of the paper is organized as follows. We present the problem statement in Section II and the proposed solution in Section III. We evaluate the proposed IHSR in Section IV and conclude this paper in Section V.

979-8-3503-5536-9/24/$31.00 ©2024 IEEE

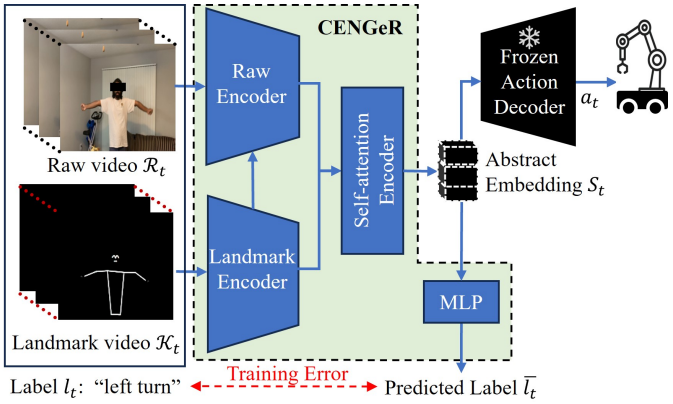


Fig. 1. High-level architecture of our IHSR. The green block highlights the proposed CENGeR. In this system, raw video $\mathcal{R}_t$ and landmark video $\mathcal{K}_t$ are the input, abstract embedding S_t is the intermediate output from CENGeR, while predicted label $\overline{l_t}$ and action a_t are the final outputs.

II. Problem Statement

In this project, we aim to design and develop a learning-based framework to enable various robotic platforms to communicate with their human co-workers by hand signals. To this end, we will train a model that can extract an abstract embedding S_t from visual observation $\mathcal{O}_t$ to represent a hand signal. Subsequently, S_t can condition a robotic action decoder to generate expected action a_t for current step t.

The visual observation, $\mathcal{O}_t = [\mathcal{R}_t, \mathcal{K}_t]$, comprises observed raw visual video $\mathcal{R}_t = [r_{t-T}, \ldots, r_t]$ and landmark video $\mathcal{K}_t = [k_{t-T}, \ldots, k_t]$. Both $\mathcal{R}_t$ and $\mathcal{K}_t$ are sequences with a length of $T \leq t$. A raw frame $r_t \in \mathbb{R}^{H \times W \times C}$ of video $\mathcal{R}_t$ will be acquired from a camera with C channels, H height, and W width. The skeleton-based landmark frame $k_t \in \mathbb{R}^{H \times W \times C}$ with the same dimensions as the raw frame. Thanks to the prevalent human skeleton detection methods, we can easily access k_t by converting an r_t or directly gain it from a 3D camera. Besides the observed $\mathcal{O}_t$, a label l_t is also requested and available at the training stage; it provides the expected action of the robot, for example, "move ahead" or "left," which offers ground-truth information for training.

III. Methods

We propose Instant Hand Signal Recognition(IHSR), a learning-based framework that empowers robots to communicate with humans using hand signals. The core of this framework is the Cross-modality Embedded Network for Gesture Recognition (CENGeR), highlighted as green in Fig. 1. CENGeR can efficiently and effectively extract a hand signal from raw video $\mathcal{R}_t$ and landmark video $\mathcal{K}_t$ and represent it as an abstract embedding S_t, which will subsequently guide the action decoder to generate expected robotic action a_t. To facilitate our proposed system to be trained by a non-technique user, we deploy a frozen and well-trained action decoder that can output expected a_t from S_t. In this way, to enable the robot to learn to react with a novel hand signal, we only need to train the CENGeR by minimizing the error between ground truth l_t and predicted $\overline{l_t}$, which is converted from S_t by the MLP module. This training error is highlighted by the red dot line in Fig. 1.

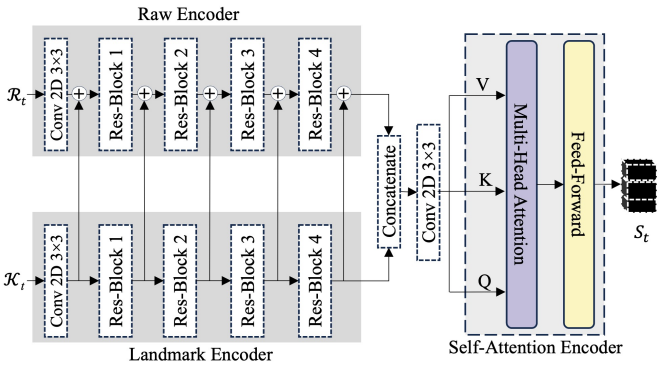


Fig. 2. Model architecture of CENGeR

A. CENGeR: Cross-modality Embedded Network for Gesture Recognition

A hand signal comprises several consecutive hand gestures, such as the "move ahead" in Marshaling signals: "Bend extended arms at elbows and move up and down from chest height to head." To effectively recognize hand signals, the proposed IHSR shall be able to detect and explore the gesture in each frame and explore the transition among consecutive gestures; in other words, it should efficiently capture the spatial and temporal patterns of hand gestures. To meet all these requirements, we design and develop CENGeR, whose architecture is illustrated in Fig. 2. It consists of two CNN-based encoders and a self-attention encoder to extract spatial and temporal features and represent them as an abstract feature S_t for the downstream action decoder.

1) CNN Encoders: The raw encoder and landmark encoder are deployed to extract spatial features in raw frame r_t and landmark frame k_t, respectively. In prior studies [7]–[9], Convolutional Neural Networks (CNNs) were utilized as encoders for feature extraction in gesture recognition tasks involving images and videos, motivating our exploration of novel CNN-based encoders for similar applications. These two encoders extract the gesture features from input visual observation frame by frame: $r_t \rightarrow e^{r_t}: \mathbb{R}^{1 \times d}$ and $k_t \rightarrow e^{k_t}: \mathbb{R}^{1 \times d}$, where d is the dimension of extracted features. Thus, we can represent the raw video $\mathcal{R}_t$ and landmark video $\mathcal{K}_t$ as extracted features $E_{\mathcal{R}_t} = [e^{r_{t-T}}, \ldots, e^{r_t}]$ and $E_{\mathcal{K}_t} = [e^{k_{t-T}}, \ldots, e^{k_t}]$, respectively. Then, we concatenate the $E_{\mathcal{R}_t}$ and $E_{\mathcal{K}_t}$ to form a comprehensive feature $\mathcal{E}_t$ to represent the visual observation $\mathcal{O}_t$ of a hand signal.

2) Cross-modality Embedding Fusion and Spatial Feature extraction: In our system, k_t is a sparse presentation with only positions of body joints. To augment these sparse gesture features with rich environmental and task context, we introduce a novel cross-modality embedding fusion technique by adding latent connections between the raw encoder and landmark encoder, as shown in Fig. 2. Both encoders comprise an equal number of blocks and produce outputs of matching dimension d_i at each block i: $e_i^{k_t} \in \mathbb{R}^{d_i}$ and $e_i^{r_t} \in \mathbb{R}^{d_i}$ are the output

embeddings from block i of raw and landmark encoders, respectively. This symmetry facilitates the cross-modality embedding fusion, where we augment the embeddings from the raw encoder using those from the landmark encoder. The latent connection between two encoders enables us to use the $e_i^{k_t} + e_i^{r_t}$ as the input of the raw encoder's block $i+1$ instead of using only the output from its block i. In this way, the raw encoder extracts the feature e^{r_t} from raw frame r_t and augments it by the corresponding e^{k_t}. This approach leverages the strengths of both encoders, combining detailed landmark and joint information with more general spatial features. This fusion enhances the overall feature representation, which is crucial for tasks requiring a nuanced understanding of human poses and gestures.

3) Self-Attention Encoder: With the raw encoder, landmark encoder, and the latent connection between them, we can efficiently extract the spatial hand gesture features from $\mathcal{O}_t$ and represent them as a comprehensive embedding $\mathcal{E}_t$. This subsection introduces a self-attention encoder to detect and extract the temporal features among consecutive gestures from $\mathcal{E}_t$ and represent them as an abstracted feature S_t. First, we deploy convolutional layers to reduce the channels of $\mathcal{E}_t$ to convert it as a more abstracted embedding $\overline{\mathcal{E}_t}$. The attention mechanism is introduced by the transformer [10] and has become a powerful tool in extracting temporal features [1], [11] and presented them as an abstract representation:

$$A_h(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

Where Q, K, and V denote queries, keys, and values in the attention mechanism, and d_k is the dimension of K. We deploy self-attention layers to capture all spatial gestures and temporal relationships to obtain S_t as:

$$S_t = A_h(\overline{\mathcal{E}_t}, \overline{\mathcal{E}_t}, \overline{\mathcal{E}_t}) \tag{2}$$

Thus, S_t is an abstract embedding that captures all spatial and temporal information of consecutive hand gestures for downstream components. Based on a S_t, the subsequent MLP module can predict hand signal label $\overline{l_t}$ for training or indication purposes. The deployed frozen action decoder also utilizes S_t to generate expectation action a_t for robots.

B. Train CENGeR

a) Pre-training process for world knowledge of human gestures: A properly designed pre-training process can significantly improve the learning efficiency of downstream tasks [12], [13]. An essential objective of our IHSR is to provide exceptional training efficiency to learn a new hand signal with minimal samples. To this end, we proposed a pre-training method based on Masked Autoencoder(MAE) to enable IHSR with world knowledge of human hand gestures, which significantly accelerates our downstream hand signal recognition. The MAE is a self-supervised learning algorithm with an encoder-decoder architecture to acquire visual representations from unlabeled data [14] by learning unmasked patches to reconstruct randomly masked-out patches, compelling the model to identify and represent essential image features, even partially or fully obscured.

The raw encoder and landmark encoder, without latent connection between them, will be pre-trained independently and separately. Two auxiliary decoders will be deployed to pair with our encoders to form two independent MAEs: Raw-MAE comprises the raw decoder and encoder, while Landmark-MAE comprises the Landmark decoder and encoder. We develop the auxiliary decoders with mirroring network architecture, exactly the same blocks but arranged in reverse order, with their corresponding encoder. These MAEs will be trained by different data sets with the same training architecture shown in Fig. 3: the encoder extracts discriminative features from the visible patches, distilling them into a compact representation; subsequently, the decoder reconstructs the image from the representation. Minimizing the error between the reconstructed and the original image enables the encoder to learn to capture the essence features.

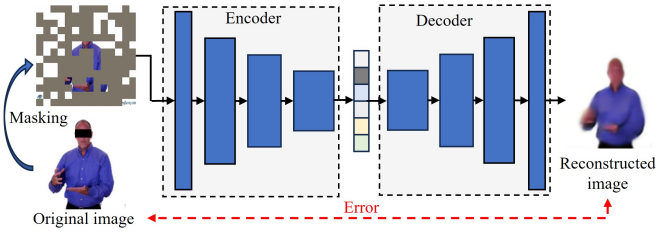


Fig. 3. The proposed MAE training architecture. The raw gesture images are deployed to train Raw-MAE, while landmark images are used to train Landmark-MAE.

By pre-training two separate MAEs, Raw-MAE and Landmark-MAE, we harness vast, task-unrelated datasets to imbue our model with world knowledge of human gestures and poses. This pre-trained knowledge, encapsulated within the raw encoder and landmark encoder, is directly applied to the CENGeR, enhancing spatial feature extraction across video frames. Massive, unrelated datasets allow our two encoders to generalize well, capturing a wide variety of visual patterns crucial for understanding complex gestures and poses from raw and landmark images, which facilitate subsequent task-specific applications.

b) End-to-end training for hand signals: As we mentioned before, once our CENGeR can recognize the hand signal and extract a related abstracted embedding S_t, the frozen robot action decoder will generate the expected robot action. This ensures a user without a robotic background to train our IHSR by providing a small training dataset: $\{(\mathcal{O}_1, l_1), \ldots, (\mathcal{O}_M, l_M)\}$, with M length. Our IHSR framework offers two end-to-end training methods, transfer-learning and fine-tuning, to train the CENGeR that enables a robot to recognize and react to hand signals from this dataset.

The CENGeR's Raw and landmark encoders in both training methods will be initialized with the well-trained models resulting from our pre-training process introduced above. In transfer learning, the raw encoder and landmark encoder are

frozen, with optimization focused on the self-attention encoder and MLP layers. Conversely, fine-tuning involves optimizing both the raw encoder and landmark encoder, along with the self-attention encoder and MLP layers. During this end-to-end training process of both methods, an observation $\mathcal{O}_t$, will be input to CENGeR to generate abstracted embedding S_t, then, the MLP will convert S_t to a predicted label $\overline{l_t}$. We deployed the Mean Square Error(MSE) as our loss function:

$$\mathcal{L} = \frac{1}{M}\sum_{t=1}^{M}(l_t - \overline{l_t})^2 \tag{3}$$

This strategy of using the pre-trained raw encoder and landmarks encoder imbues the CENGeR architecture with extensive world knowledge of human gestures and poses, significantly improving the learning efficiency of hand signal recognition to facilitate state-of-the-art results even with minimal datasets.

IV. Experiments

We implemented the proposed system using the PyTorch Lightning package and Python 3.8.0 and evaluated it on a computer with an NVIDIA GeForce RTX 3090 with 128GB of RAM. Below are further details of each experiment.

A. Experiments for Pretraining Models

Thanks to the openly accessible datasets: HaGrid dataset [15], Bangla Sign Language dataset [16], WASL-2000 dataset [17] and Arabic Sign Language dataset [18], we compose a pre-training raw dataset with 70K+ images. To form a landmark pre-training dataset, we used Mediapipe [19] to create landmark images from all raw images in the above dataset. A qualitative result is shown in Fig. 4: the pre-trained Raw-MAE and Landmark-MAE can both reconstruct images that keep essential visual information, which indicates that our raw encoder and landmark encoder are capable of efficiently capture essential features to represent the hand gestures.

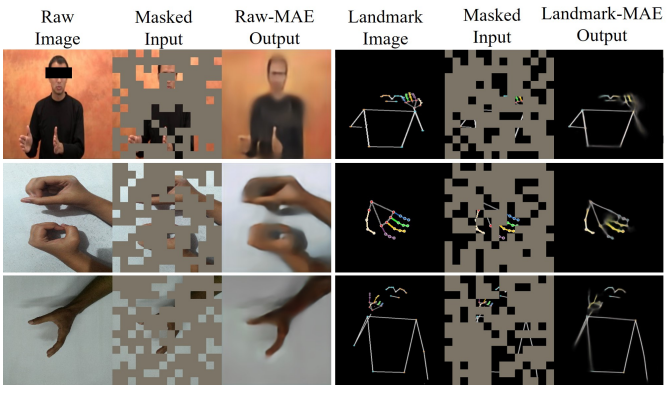


Fig. 4. Visual representation of image reconstruction of Raw-MAE and Landmark-MAE.

B. Experiments for Hand Signal Recognition

1) Training Setup and Dataset: To evaluate the performance of our IHSR, we create a customized AirMarshVideo dataset that includes three actions with labels: "Move ahead," "Right," and "Left." Its raw video $\mathcal{R}_t$ in an observation $\mathcal{O}_t$ is recorded by RGB cameras from several volunteers, while the landmark video $\mathcal{K}_t$ is generated using Mediapipe [19]. The sample of this AirMarshVideo dataset is given in Fig. 5. This dataset has 59, 60, and 61 data video samples in both raw and landmark for the "Move ahead", "Right," and "Left" classes, respectively. Each video sample has only 20 frames ranging from 3 to 6 seconds, each frame with a dimension of 224×224. We employed a cross-validation technique that divided this dataset into three cross-folds for the objectives of training, validation, and testing. In each fold, about 60% of video samples(around 36 samples) were for training, 15% of video samples(9 samples) for validation, and the remaining 25% of video samples(15 samples) for testing. In this way, we ensured the three folds had different training, validation, and testing samples. During 100-epoch training, the hyperparameters were updated using a validation loss-based early stop criteria with an early stop patience of 15 epochs. TABLE I provides more hyperparameters used in this study.

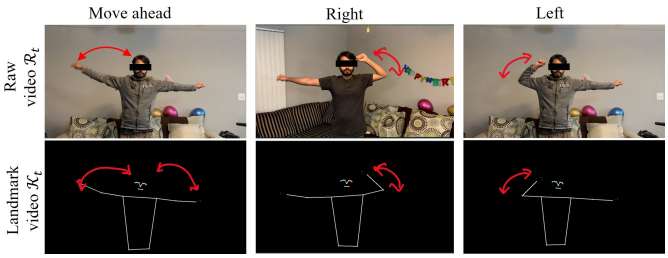


Fig. 5. A sample to illustrate three representative hand signals in our AirMarshVideo dataset.

TABLE I
Hyperparameters

Hyperparameter	Value
Batch Size	6
Initial Learning Rate	0.0001
Learning Rate Drop Factor	0.1
Learning Rate Drop Patience	15
Number of Epochs	100
Early Stopping Patience	35
Early Stopping Monitor	Validation Loss
Early Stopping Mode	min
Optimizer	Adam

2) Learning Efficiency: Based on the above three folds of datasets, we trained three CENGeR models using the fine-tuning method and the other three CENGeR models by transferring learning to assess the learning efficiency of our proposed IHSR framework. Then, we tested how accurately the learned models can predict the labels, $\overline{l_t}$, for all samples in the testing set of all three folds. Note that samples in the testing set were not included in the training and validation set; they may be subjected to totally different backgrounds, people, and gesture patterns. We also compared our predicted accuracy and average consumed training samples with two state-of-the-art Marshalling recognition models. TABLE II illustrates the experiment results.

TABLE II
LEARNING EFFICIENCY

Model		Accuracy	Average samples
CENGeR models	M_{t1}	91%	45
	M_{t2}	91%	
	M_{t3}	94%	
	M_{f1}	99%	
	M_{f2}	94%	
	M_{f3}	97%	
State-of-the-art models	CPM-classifier [3]	97%	1445
	OpenMarshall [4]	88%	–*

* Authors of [4] did not disclose the training samples. However, all our models are superior in accuracy.

In TABLE II, M_{t1}, M_{t2}, and M_{t3} denotes the transfer learning models that are trained on fold 1, fold 2, and fold 3. M_{f1}, M_{f2}, and M_{f3} are the fine-tuning models trained on fold 1, fold 2, and fold 3. The experimental results show that all our CENGeR models can learn to recognize a hand signal with only an average of 45 samples. Compared to the CPM-classifier, introduced by the latest work [3], which consumes an average of 1445 samples per signal to achieve the same level of prediction accuracy, our IHSR offers 30+ times improvements in learning efficiency. These results show that our proposed framework, IHSR, significantly improves learning efficiency, enabling it to learn novel hand signals after deployments.

3) Generalization to novel conditions: To assess the generalization robustness of our models to novel conditions, we invited several different volunteers to collect another 57 samples(17 samples for "move ahead", 20 for "right " and "left") of marshaling hand signals to form an unseen dataset. All samples in this set differ dramatically from our training set in the background, lighting, and user gesture patterns, which form a different sample distribution. Then, we deployed all the models we trained to predict all samples in our unseen dataset. TABLE III shows the predicted accuracy for unseen samples and compares them with the accuracy in seen scenarios that were introduced in the previous experiments. As mentioned before, our testing and training sets were randomly sampled from AirMarshVideo, so they shared similar scenarios and the same sample distribution; thus, the experimental results in the previous section IV-B2 represented well for seen scenarios.

TABLE III
GENERALIZATION

Scenarios	M_{t1}	M_{t2}	M_{t3}	M_{f1}	M_{f2}	M_{f3}
seen	91%	91%	94%	99%	94%	97%
unseen	93%	86%	88%	91%	95%	93%

From TABLE III, we can tell that all our models provide a very good zero-shot generalization for novel scenarios, especially the fine-tuning models that offer that same level of accuracy in seen and unseen scenarios. Fig. 6 illustrates the confusion matrices for our IHSR on unseen scenarios, evaluated via 3-fold cross-validation employing both transfer learning and fine-tuning methods. The fine-tuning models demonstrably excelled, achieving superior performance in true positives, true negatives, false positives, and false negatives. This enhancement highlights the IHSR's adeptness at generalizing to unseen subjects, backgrounds, and gesture patterns, underscoring the efficacy of the proposed training methodology and all network components in robust feature extraction and hand signal recognition.

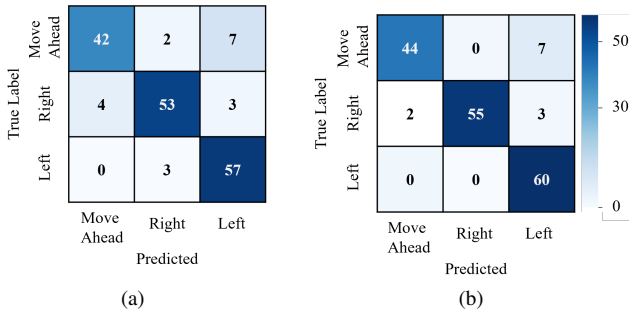


Fig. 6. Confusion matrices for IHSR in unseen scenarios: (a) Combined results from all transfer learning models M_{t1}, M_{t2}, and M_{t3}. (b) Combined results from all fine-tuning models M_{f1}, M_{f2}, and M_{f3}

Experimental results in TABLE III and Fig. 6 underscore the appropriateness of the employed training techniques, especially in scenarios demanding rapid learning from limited data. The superior performance of the fine-tuning models further emphasizes the potential of IHSR in enabling robots to comprehend complex hand gestures with minimal training data, achieving higher accuracy. This capability is crucial for robotics applications, where the ability to learn from a small set of examples can significantly expedite the training process and enhance the robot's adaptability to new tasks or environments.

4) Ablation Study: We performed ablation studies to assess the importance of components in our CENGeR by training models of 1) only raw encoder, 2) only landmark encoder, and 3) no cross-modality function that is no latent connection between the raw and landmark encoders. First, we trained M_t^r, M_t^l, and M_t^n as the ablation models 1), 2), and 3) using transfer learning, respectively. Second, the other three ablation models of 1), 2), and 3) were trained by fine-tuning technology are M_f^r, M_f^l, and M_f^n. Meanwhile, M_t and M_f are the original CENGeR models that used transfer learning and fine-tuning. We also trained and validated them with the same three-fold setup introduced for learning efficiency in section IV-B2, but we only selected the best accuracy to represent the performance of each model. The comparison of their accuracy is shown in TABLE IV.

TABLE IV
ABLATION STUDY

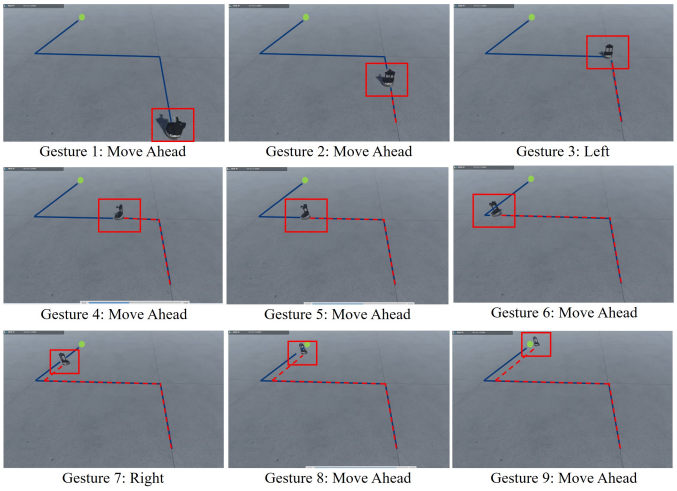
	M_t^r	M_f^r	M_t^l	M_f^l	M_t^n	M_f^n	M_t	M_f
seen	63%	39%	96%	90%	90%	90%	94%	97%
unseen	52%	35%	89%	85%	89%	80%	93%	95%

Results in TABLE IV tell that our design of the raw encoder, landmark encoder, and cross-modality fusion improve the robustness of our system, especially under unseen scenarios. At first glance, they may indicate that the landmark encoder contributes more to hand signal detection, but our further analysis revealed when the landmarks are not reliable, like

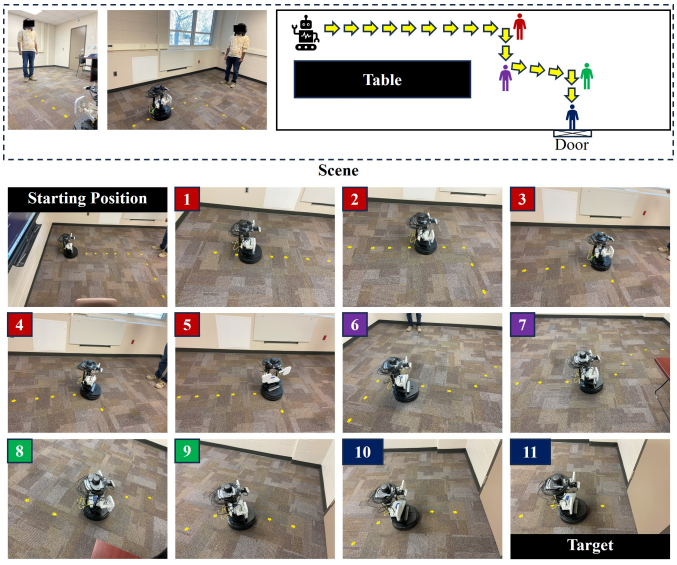
under poor lighting scenarios, the raw encoder and the cross-modality fusion help to enhance the accuracy of the system.

5) Robots experiments: To evaluate the proposed IHSR's feasibility in facilitating robot-human coordination by hand signals, we deployed the trained model M_{f2} in a virtual environment introduced in [1] and a real-world experiment using a LoCoBoT, which is presented in Fig. 7. In both experiments, the robotic action decoder took the output from M_{f2} to generate an action a_t to navigate the robots in a step-by-step mode. We designed experimental scenarios that require a user navigated mobile robots to follow given routes, blue lines in Fig. 7(a) and yellow arrows in Fig. 7(b), by combining three hand signals that M_{f2} has been trained for: "move ahead", "right," and "left."

In the virtual environmental experiment, we deployed an RGB camera face to a user to collect raw video $\mathcal{R}_t$ and convert it to landmark video $\mathcal{K}_t$, then inputted to IHSR to generate a_t for the robot. Fig. 7(a) shows it is capacity to follow a blue-lined route towards a green-circled target in the virtual environment. This demonstrates the IHSR's capability in real-time gesture and pose recognition for robotic navigation. We deployed our real-world experiment in a conference room with a size of $5 \times 5\ m^2$. A user stood in front of the robot and used air Marshalling gestures to maneuver it toward a planned course to move from one side of a table to the door, as shown in Fig. 7(b). The successful navigation results in both experiments demonstrate the IHSR's ability to enable robotic navigation via gesture recognition, providing a promising method for intuitive human-robot collaboration. These results also show the IHSR's robustness, gesture interpretation accuracy, and suitability for various robotic platforms.



(a) Virtual environment experiment: the user navigated the robot, marked as a red block, to follow the blue route by Marshalling signals. A red dot line illustrates the robot's trajectory, which tracks the designed blue route well.



(b) Real-world environment experiment: The top left figures show a user standing before the robot to provide hand signals. The top right figure illustrates our experimental design, in which a user navigates the robot to follow the arrows on the floor, and the human symbols indicate the users' positions while the robot navigates. The rest of the figures show photos of the robot navigating and following the designed route to reach the target.

Fig. 7. The robot experiments in (a) virtual and (b) Real-world scenarios.

V. Conclusions

In this work, we presented IHSR, a learning-based framework that empowers robots to communicate with humans using hand signals. IHSR comprises CENGeR, an elaborated network architecture, and delicate training methods to rapidly learn novel hand signals from very few samples in complex environments. The proposed CENGeR captures the latent spatial relation between the spare joint-based landmark and context-rich raw frames. Its self-attention encoder captures the latent relations in the temporal gesture patterns and implants them with spatial information into abstract embedding for downstream robotic tasks. Additionally, IHSR proposed a pre-training on massive and tasks unrelated datasets to embed CENGeR with elemental knowledge in detecting various hand gestures and poses, which will facilitate the rapid training for hand signal recognition. Extensive experiments showed that our IHSR could efficiently learn a novel hand signal in less than 50 samples and offer strong out-of-distribution robustness in novel scenarios. In summary, IHSR makes a promising future of enabling hand signal-based communication to foster robot-human collaboration in many voice-forbidden environments, such as construction sites and airport ramps. Our future work will focus on the robotic action decoder, which could fuse environmental observations and CENGeR's hand signal embeddings to form an intelligent task policy for complex scenarios. For example, it could compensate its human co-workers when it finds potential failures/mistakes in received hand signals. Meanwhile, we also found that misclassification of gestures, even if it is highly unlikely, may lead to the robot performing incorrect actions. To overcome this limitation, one of our future research directions will focus on using formal methods to prevent incorrect hand signal recognition and downstream action performed by robots.

Acknowledgment

This work is supported in part by the NSF under Grants CCSS-2245607 and CCSS-2245608.

REFERENCES

[1] Y. Wu, J. Zhang, S. Wu, S. Mao, and Y. Wang, "Cmrm: A cross-modal reasoning model to enable zero-shot imitation learning for robotic rfid inventory in unstructured environments," in *IEEE Global Communications Conference 2023*, 2023.

[2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[3] M. Á. de Frutos Carro, F. C. LópezHernández, and J. J. R. Granados, "Real-time visual recognition of ramp hand signals for uas ground operations," *Journal of Intelligent & Robotic Systems*, vol. 107, no. 3, p. 44, 2023.

[4] D. Pal, A. Singh, H. Khairnar, and A. Alladi, "Openmarshall: Open-set recognition of aircraft marshalling signals for safe docking," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, 2023, pp. 1–6.

[5] C. Choi, J.-H. Ahn, and H. Byun, "Visual recognition of aircraft marshalling signals using gesture phase analysis," in *2008 IEEE Intelligent Vehicles Symposium*. IEEE, 2008, pp. 853–858.

[6] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: Natops aircraft handling signals database," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 500–506.

[7] K. K. Podder, M. E. Chowdhury, A. M. Tahir, Z. B. Mahbub, A. Khandakar, M. S. Hossain, and M. A. Kadir, "Bangla sign language (bdsl) alphabets and numerals classification using a deep learning model," *Sensors*, vol. 22, no. 2, p. 574, 2022.

[8] K. K. Podder, M. Chowdhury, Z. B. Mahbub, and M. Kadir, "Bangla sign language alphabet recognition using transfer learning based convolutional neural network," *Bangladesh J. Sci. Res*, pp. 31–33, 2020.

[9] K. K. Podder, M. Ezeddin, M. E. Chowdhury, M. S. I. Sumon, A. M. Tahir, M. A. Ayari, P. Dutta, A. Khandakar, Z. B. Mahbub, and M. A. Kadir, "Signer-independent arabic sign language recognition system using deep learning model," *Sensors*, vol. 23, no. 16, p. 7156, 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] E. Shabaninia, H. Nezamabadi-pour, and F. Shafizadegan, "Transformers in action recognition: A review on temporal modeling," *arXiv preprint arXiv:2302.01921*, 2022.

[12] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.

[13] H. Chen, J. Wang, A. Shah, R. Tao, H. Wei, X. Xie, M. Sugiyama, and B. Raj, "Understanding and mitigating the label noise in pre-training on downstream tasks," *arXiv preprint arXiv:2309.17002*, 2023.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[15] A. Kapitanov, A. Makhlyarchuk, and K. Kvanchiani, "Hagrid - hand gesture recognition image dataset," *arXiv preprint arXiv:2206.08219*, 2022.

[16] A. M. Rafi, N. Nawal, N. Bayev, L. Nima, C. Shahnaz, and S. A. Fattah, "Image-based bengali sign language alphabet recognition for deaf and dumb community," in *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, 10 2019, pp. 1–7.

[17] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1459–1469.

[18] D. Biyani, N. Doohan, M. Rode, and D. Jain, "Real time sign language recognition using yolov5," in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, 07 2023, pp. 582–588.

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.