Trustworthy Hand Signal Communication Between Smart IoT Agents and Humans

[†]Kanhon Kanti Podder, [‡]Jian Zhang, and [§]Shiwen Mao
[†]Department of Electrical and Computer Engineering, Kennesaw State University, Marietta, GA 30060, USA
[‡]Department of Information Technology, Kennesaw State University, Marietta, GA 30060, USA
[§]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
Email: kpodder@students.kennesaw.edu, jianzhang@ieee.org, smao@ieee.org

2024 IEEE 10th World Forum on Internet of Things (WF-IoT) | 979-8-3503-7301-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/WF-IoT62078.2024.10811287

***Abstract*—Hand signals are the most widely used, feasible, and device-free communication method in manufacturing plants, airport ramps, and other noisy or voice-prohibiting environments. Enabling IoT agents, such as robots, to recognize and communicate by hand signals will facilitate human-machine collaboration for the emerging "Industry 5.0." While many prior works succeed in hand signal recognition, few can rigorously guarantee the accuracy of their predictions. This project proposes a method that builds on the theory of conformal prediction (CP) to provide statistical guarantees on hand signal recognition accuracy and, based on it, measure the uncertainty in this communication process. It utilizes a calibration set with a few representative samples to ensure that trained models provide a conformal prediction set that reaches or exceeds the truth worth and trustworthiness at a user-specified level. Subsequently, the uncertainty in the recognition process can be detected by measuring the length of the conformal prediction set. Furthermore, the proposed CP-based method can be used with IoT models without fine-tuning as an out-of-the-box and promising lightweight approach to modeling uncertainty. Our experiments show that the proposed conformal recognition method can achieve accurate hand signal prediction in novel scenarios. When selecting an error level α = 0.10, it provided 100% accuracy for out-of-distribution test sets.**

***Index Terms*—Hand signals, Uncertainty measurement, Conformal prediction, Gesture recognition,**

I. Introduction

In manufacturing plants, construction sites, airport ramps, and the like, hand signals are the widely used and feasible, device-free communication method in noisy or voice-prohibiting environments, such as forklift hand signals in a hectic manufacturing warehouse and Air Marshaling signals [1] to direct aircrafts at an airport or helipad. Compared to other methods, hand signaling's shallow learning curve makes it ideal for supplemental communications, particularly in times of emergency, between non-technical human workers and smart IoT agents, e.g., robots, that will be extensively deployed in various workplaces in the soon-to-be-established "Industry 5.0." Many recent works [2]–[5] focus on recognizing hand signals or gestures. For example, the authors in [2] developed Convolutional Pose Machines and learning classifiers to segment and classify gestures in diverse and ambient circumstances. The authors in [3] introduced Open-Marshall, a gesture-tracking model for aircraft Marshalling signal recognition. Despite the advances made in hand gesture recognition, their predictions lack strictly guaranteed accuracy. This limitation makes it highly challenging to realize a promising future of hand signals in Industry 5.0, as two fundamental questions remain unsolved: (i) How do smart IoT agents ensure that their signal recognition is truth-worthy? (ii) How do they identify uncertainties in visual signals?

This research aims to break through the existing barrier and address the above challenges by proposing a method that builds on the theory of conformal prediction (CP) [6], [7] to provide statistical guarantees on hand signal recognition accuracy and, based on it, measure the uncertainty in such visual signaling communication process. The proposed method introduces a learning-based approach to enable IoT agents to recognize and react to hand signals, and then a calibration framework is introduced to utilize a posterior limited-size sample set to ensure that trained models provide a conformal prediction set that reaches or exceeds the truth worth and trustworthiness level under novel task conditions. Lastly, by measuring the length of the conformal prediction, an agent can identify if it faces uncertainty in this hand signal recognition process. Establishing trust between humans and artificial agents is critical in the Industry 5.0, as this trust can only be fostered based on the agents' reliable and consistent action interpretation and execution at a precise hand signal recognition and capable of addressing ambiguities [8]. To the best of our knowledge, this is the first study focused on rigorously measuring truthfulness and uncertainty in visual hand signal communication. We highlight our main contributions in this study as follows:

- This research introduces conformal recognition, an innovative approach for calibrating confidence levels within a gesture recognition framework. CP's statistical rigorousness ensures that only high-certainty gestures prompt action, while ambiguous ones trigger requests for clarification or other further processes. This feature significantly enhances the safety and reliability of human-artificial agent collaboration, especially in critical environments where precise and accurate hand signal interpretation is essential.
- Theoretical guarantees are provided to ensure that the artificial IoT system can interpret and act upon hand signals with a predefined level of confidence. CP's foundation in rigorous statistics methods significantly reduces the margin for error in command execution, providing a reliable basis for complex human-machine interactions.

979-8-3503-7301-1/24/$31.00 ©2024 IEEE

- The proposed framework utilizes a posterior calibration set to rectify predictions from a trained model to reach the accuracy marginal. It can be used with any model without fine-tuning or modifications, making it an out-of-the-box and lightweight approach to uncertainty modeling.
- We also introduce rigorous evaluation metrics to quantitatively measure the performance of the proposed conformal recognition method, combined with extensive experiments to prove that our proposed method can enable the smart IoT agent to provide 100% accuracy in recognizing hand signals in novel conditions, including new users and environments.

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 introduce the performance evaluation metrics and evaluate the proposed conformal recognition in Section IV. Section V concludes this paper.

II. Problem Statement

A completed hand signal can be presented as a data entity (O_t, l_t), where $O_t \in \mathcal{O}$ is the observation and $l_t \in \mathcal{L}$ is the ground truth label, like "Forward." Here $\mathcal{L}$ is a limited set to form the signal space that includes all support signals, and $\mathcal{O}$ is the observation space. In this work, an observation $O_t = [\mathcal{R}_t, \mathcal{K}_t]$ comprises a raw RGB video $\mathcal{R}_t = [r_{t-T}, \dots, r_t]$ and a landmark video $\mathcal{K}_t = [k_{t-T}, \dots, k_t]$. $\mathcal{R}_t$ and $\mathcal{K}_t$ are sequences of frames from a video with $T \leq t$. Here, r_t denotes an RGB frame, and a landmark frame k_t provides the skeleton image comprising the subject's relative joint positions.

a) Goal 1. Confidence guaranteed recognition: This study aims to develop a trustworthy framework for hand signal communication to ensure that smart IoT agents can guarantee to recognize the true signal from observation O_t. We will train a model $G : \mathcal{O} \to \mathcal{S}$, where $\mathcal{S}$ is an abstract embedding space. By a given observation $O_t \in \mathcal{O}$, G can extract an entity $S_t \in \mathcal{S}$, which is an abstract embedding representing a given hand signal. Based on a S_t, a subsequent classifier model $C(\cdot)$ provides a set of predicted labels $\hat{L}_t = C(S_t)$, where $\hat{L}_t$ is a subset of signal space $\hat{L}_t \subseteq \mathcal{L}$. Our goal is to ensure the models G and C to satisfy the following condition:

$$\mathbb{P}\Big[l_t \in C\Big(G(O_t)\Big)\Big] \geq (1-\alpha), \quad \forall O_t \in \mathcal{O}, l_t \in \mathcal{L}, \tag{1}$$

where $\alpha \in (0, 1)$ represents the level of error that the user can tolerate. When the classifier model C satisfies (1), it can guarantee to predict the ground truth label l_t with $(1-\alpha)$ confidence. It also guarantees that the model G extracts a correct embedding S_t corresponding to l_t with $(1-\alpha)$ confidence. This guarantee can also build confidence marginal in downstream models, such as the local actor $M : \mathcal{S} \to \mathcal{A}$, which can generate robot action $a_t \in \mathcal{A}$ from any given S_t.

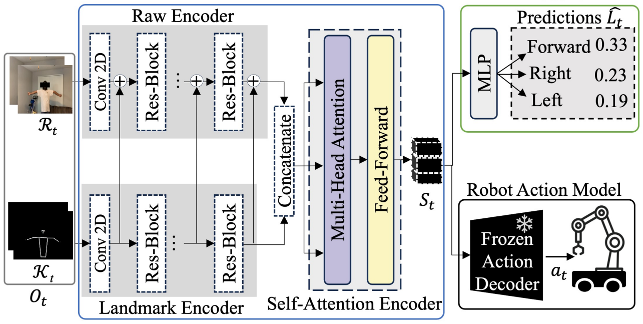


Fig. 1: The model architecture of the proposed hand signal communication between human users and smart IoT agents.

b) Goal 2. Quantify uncertainty: Once Goal 1 is achieved, we hope to quantify the uncertainty in the process of hand signal recognition. In hand signal communications, a given observation O_t should only correspond to one true label l_t. In this work, we also aim to measure the uncertainty of prediction for models C and G to avoid the potential errors propagating to downstream models. This problem can be formulated as:

$$\mathbb{U}\Big[C\Big(G(O_t)\Big)\Big] = \begin{cases} 1, & \text{if uncertainty detected} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbb{U}$ measures the uncertainty of C and G and indicates if the models are certain about the current hand signal O_t. Combining with (1), we guarantee that the models correctly and unambiguously recognize O_t as l_t, with $(1-\alpha)$ confidence.

III. Method and Materials

In this study, we propose a learning-based method to recognize hand signal from observations. Then, we will deploy conformal prediction to ensure the accuracy of the proposed method. Lastly, we will introduce the method to detect uncertainty in the recognition process.

A. Recognize hand signals

In this section, we introduce a learning-based network to recognize the hand signals from the observation O_t. A hand signal consists of several consecutive hand gestures, such as the signal of "move forward" (for short, "forward" in this work) in Air Marshaling [1]. Fig. 1 illustrates the high-level architecture of the proposed network to enable human-robots hand signal communication. It comprises three modules: model G extracts the abstract embedding S_t from the observation O_t, Model C will predict labels $\hat{L}_t$ from S_t, and an action model will based on the S_t to generate action a_t to react for O_t. The design of the robot action model is out of this work's scope, so we will ignore its technique details in this work and assume it is frozen and can always generate the right action a_t.

a) Model G: There is a prevalent of literature available, such as [9]–[11], where Convolutional Neural Networks (CNNs) were utilized for gesture detection and classification. This has inspired us to utilize CNN-based encoders to extract spatial features for hand signal recognition. Two CNN-based encoders, a raw encoder and landmark encoder, are developed to extract spatial features frame-by-frame from the raw video $\mathcal{R}_t$ and landmark video $\mathcal{K}_t$, respectively. These two encoders are latently connected to augment the sparse gesture features from the landmark encoder to the raw encoder, and we propose these techniques as cross-modality fusion. We concatenate the outputs from these two encoders to form a feature for the following network layer. The attention mechanism proposed by the transformer [12] has shown to be a method for extracting temporal information [13], [14] and representing them in an abstract embedding. Inspired by these works, we deployed a self-attention encoder to capture both spatial relations and temporal interactions from the concatenated feature extracted from the observation O_t and encompass them as an abstract embedding S_t.

b) Model C: Based on S_t, the model C deploys a multilayer perceptron(MLP) network to predict a set of labels $\hat{L}_t = [\hat{l}_1, ..., \hat{l}_n] \subseteq \mathcal{L}$. The output layer deploys a softmax activation function that associates each prediction label $\hat{l}_i$ with an uncalibrated confidence value $\hat{v}_i$.

Definition 1 (Non-conformal recognition of a hand signal)**.** we will select the one with the highest $\hat{v}_i$ in $\hat{L}_t$ as the predicted label $\hat{l}_t$ for the observation O_t, and we call $\hat{l}_t$ a non-conformal recognition.

c) Training process: We deploy a two-phase training process for models G and C. (i) *Phase 1: Pre-training*. We use a self-supervised Masked Autoencoder (MAE) [15], [16] to pre-train the raw and landmark encoders. Two auxiliary decoders will be deployed to train the two encoders in model G. By masking part of input images, the encoder is trained to efficiently extract features from visible patches, allowing the auxiliary decoder to reconstruct the complete image. Thanks to the publicly available hand gesture image datasets, such as [17]–[20], and the related landmark images that are generated by skeleton extracting tools, such as MediaPipe [21], we can train model G to gain rich knowledge about hand gestures through this MAE pre-training process. (ii) *Phase 2: Supervised hand signal training*. We use a small hand signal dataset $\mathcal{O}^h$ to train model C and the pre-trained model G. By minimizing the error of the non-conformal, predicted label $\hat{l}_t$ compared to the ground truth label l_t, we can enable the two models to learn to recognize the hand signal from its observation O_t.

B. Confidence guaranteed hand signal recognition

Conformal Prediction (CP) [6], [7] is a framework that enables any model to produce a statistically guaranteed prediction set containing ground-truth labels with a desired probability. It utilizes a calibration dataset that includes a small number of representative samples associated with corresponding ground-truth values for a novel task distribution to output a rigorous, guaranteed prediction set.

Based on the CP framework, our proposed confidence guaranteed recognition utilizes a calibration set, $\mathcal{O}^c = \{(O^c_1, l^c_1), \dots, (O^c_m, l^c_m)\}$, that includes hand signals for novel subjects and backgrounds to guarantee the trained model G and model C satisfy (1). First, we use the uncalibrated confidence value $\hat{v}$ to get the calibration set's nonconformity scores $\{q_j = 1 - (\hat{v}_j)_{l^c_j}\}^m_{j=1}$, where $(\hat{v}_j)_{l^c_j}$ is the uncalibrated confidence value that corresponds to the ground-truth prediction l^c_j at observation O^c_j. To determine $(\hat{v}_j)_{l^c_j}$, we collect the model's output for each possible label and identify the uncalibrated confidence level $\hat{v}_j$ associated with the ground-truth label l^c_j. Second, CP will define $\hat{q}$ as the $(m+1)(1-\alpha)/m$ empirical quantile of the nonconformity scores set $Q = \{q_1, \dots, q_m\}$, as:

$$\hat{q} = \mathbb{Q}\bigg(Q, \frac{(m+1)(1-\alpha)}{m}\bigg), \tag{3}$$

Here, $\mathbb{Q}$ represents the quantile function; and the fraction $(m+1)(1-\alpha)/m$ specifies the desired quantile, ensuring that $\hat{q}$ acts as a threshold to determine the uncertainty level of the predictions [7], [22]. Then, for any given new hand signal $O_t \sim \mathcal{O}^c$ (i.e., having a similar distribution as $\mathcal{O}^c$), CP will generate a calibrated prediction set, given by:

$$\hat{L}_t = \{\hat{l}_i : \hat{v}_i \geq (1 - \hat{q})\}. \tag{4}$$

We can guarantee that this prediction set will satisfy:

$$\mathbb{P}[l_t \in \hat{L}_t] \geq (1-\alpha), \forall O_t \sim \mathcal{O}^c, \tag{5}$$

where l_t is the ground-truth label of O_t. Therefore, we can achieve our goal (1) with a constraint that requires any test observation O_t to have a similar distribution of the calibration set $\mathcal{O}^c$. The proposed confidence guaranteed recognition method is presented in Algthorim 1.

Algorithm 1 Confidence guaranteed hand signal recognition

```
Input: Calibration dataset $\mathcal{O}^c = \{(O^c_1, l^c_1), \dots, (O^c_m, l^c_m)\}$, error level parameter $\alpha \in (0, 1)$
Input: Observation $O_t$
Output: $\hat{L}_t, s.t. \mathbb{P}(l_t \in \hat{L}_t) \geq (1-\alpha)$
  for $i = 1 : m$ do
      Calculate nonconformity scores $Q = \{1 - (\hat{v}_j)_{l^c_j}\}^m_{j=1}$ ;
  end for
  Calculate $\hat{q}$ as in (3) ;
  Model $C$ predicts non-conformal labels for $O_t$, $C : O_t \to \{(\hat{l}_1, \hat{v}_1), \dots, (\hat{l}_n, \hat{v}_n)\}$ ;
  Create the guaranteed prediction set: $\hat{L}_t = \{\hat{l}_i : \hat{v}_i \geq (1 - \hat{q})\}$ ;
```

Creating the calibration dataset $\mathcal{O}^c$: Eqn. (5) shows that we should collect representative samples to form $\mathcal{O}^c$ and use it to guarantee the prediction under novel conditions. We propose the *subject-wise selection*, a simple and straightforward paradigm, for creating a calibration dataset. Considering we have μ participants or users who provide hand signals with different backgrounds. For each participant, we collect $\tau \geq 1$ samples per hand signal to form the calibration set $\mathcal{O}^c$.

C. Uncertainty estimation & conformal recognition

Hand signaling is an unambiguous visual communication method; the observation O_t of a specific signal must always be interpreted as a distinctive message. The above-proposed confidence guaranteed hand signal recognition algorithm leverages model C to output a ground-truth guaranteed prediction set $\hat{L}_t$. Thus, by measuring the singleton of this prediction set, we can detect the uncertainty of the recognition process. We formulate this uncertainty measurement as follows:

$$\mathbb{U}\Big[C\Big(G(O_t)\Big)\Big] = \mathbb{1}\{len(\hat{L}_t) \neq 1\} = \begin{cases} 1, & \text{uncertain} \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathbb{1}\{len(\hat{L}_t) \neq 1\}$ is an indicator function that returns "1" if the length of $\hat{L}_t$ is not 1, and "0" while $\hat{L}_t$ is a singleton set. It empowers our method to achieve the second goal (2). Additionally, combined with (5), the above equation can guarantee a singleton prediction, namely prediction without uncertainty, with $(1-\alpha)$ confidence close to the ground-truth.

Definition 2 (Conformal recognition of a hand signal)**.** The case of $\mathbb{U} = 0$ indicates that model C produces a singleton prediction $\hat{L}_t = \{\bar{\bar{l}}_t\}$ for an observation O_t, and $\bar{\bar{l}}_t$ is regarded as the *conformal recognition.*

When $\mathbb{U} = 1$, there is uncertainty in the prediction, which indicates that maybe model G failed to extract sufficient features and embed them into S_t, and this uncertainty may also propagate to the downstream action generation or other models, and may cause unacceptable errors. For error-costly applications, (6) allows the robot, or other intelligent IoT agents, to avoid potential errors or initiate a further confirmation request to ensure execution a more proper action.

IV. Experiments and Results

A. Experiments for non-conformal recognition

1) Experimental setting: In this section, we introduce the configuration and datasets we used to train and evaluate the proposed non-conformal recognition. *Pre-training setup*: the pre-training dataset includes massive unlabeled raw images of a total of 67,554 images for training, 1,893 images for validation, and 1,716 for testing, which is collected and pre-processed from publicly available hand gesture-related image datasets [17], [18], [20] and a video dataset [19]. The related landmark dataset is generated from the raw image dataset by using hand and pose landmarks MediaPipe [21]. During pre-training, we masked out 70% of the input images, an MSE loss was used to calculate the error on masked input and reconstructed image patches, and an ADAM optimizer was used for optimization. *Non-conformal recognition setup*: To train and evaluate the performance of the proposed non-conformal recognition models, we collect two independent data sets. First, we obtained a video dataset $\mathcal{O}^h = \{(O^h_1, l^h_1), \ldots, (O^h_n, l^h_n)\}$ from $\rho = 8$ participants. It includes $|\mathcal{O}^h| = 180$ samples for three hand signals {"forward", "right," "left"}, and hereafter, if not mentioned, all the datasets are based on these three signals. Here, O^h_t is the observation and is associated with a ground-truth label l^h_t. We partitioned $\mathcal{O}^h$ into three distinct subsets for training and evaluation: a training subset $\mathcal{O}^{t-l}$, a validation subset $\mathcal{O}^{t-v}$, and a test subset $\mathcal{O}^{h-t}$. The test subset was composed exclusively of data points different from the other two subsets to ensure that it is independent but similar to the training subset, so $\mathcal{O}^{h-t}$ was an In-Distribution Test Set (*ID Test Set*). Second, we additionally collected an Out-of-Distribution Test Set (*OOD Test Set*), denoted as $\bar{\mathcal{O}}$. It was collected from $\mu = 12$ participants for testing, here $|\bar{\mathcal{O}}| = 108$, and all the μ participants were different from the ones from whom our training dataset $\mathcal{O}^h$ was collected.

We used a batch size of 4, the ADAM optimizer, and the cross-entropy loss to train model G and model C based on the samples of the training subset $\mathcal{O}^{t-l}$. They were evaluated by the non-conformal recognition results from model C using two popular metrics for classification tasks: accuracy and precision.

2) Experimental results: We evaluated the performance of the proposed hand signal recognition network, i.e., model G and model C, for non-conformal recognition on the ID test set, $\mathcal{O}^{h-t}$, and the OOD test set $\bar{\mathcal{O}}$. The experimental results are shown in Table I. For the ID test set, model C achieves 99% accuracy and 99% precision; however, when the identical model was evaluated on the OOD test set, which involved different participants and backgrounds, its performance notably decreased to 86% accuracy and 86% precision. This outcome suggests that the model while achieving a satisfactory level of performance, was unsuccessful in an unfamiliar setting and with diverse participants. Each participant executed the hand signal at varying speeds and with distinct motions. Therefore, it is necessary to perform additional calibration of the model's predictions before using it in a real-world situation. Otherwise, the model may make incorrect recognition, which may lead to dangerous actions for a smart IoT agent.

TABLE I: Non-conformal Recognition Results

Test Set	Accuracy	Precision
ID test set $\mathcal{O}^{h-t}$	99%	99%
OOD test set $\bar{\mathcal{O}}$	86%	86%

B. Experiments for conformal recognition

1) Experimental Setup: We divided the OOD Test Set $\bar{\mathcal{O}}$ collected form $\mu = 12$ participant into two subsets, $\bar{\mathcal{O}}^c$ and $\bar{\mathcal{O}}^v$, where $\bar{\mathcal{O}}^c$ is the calibration set used to conform the trained models on the OOD set, $\bar{\mathcal{O}}^v$ is the validation set to assess the overall performance and reliability of our proposed conformal recognition. We derived $\bar{\mathcal{O}}^c$ from $\bar{\mathcal{O}}$ by the proposed subject-wise selection method with $\tau = 1$ sample per hand signal from each participant, resulting in a total of $|\bar{\mathcal{O}}^c| = 36$ samples. The rest of the samples in the $\bar{\mathcal{O}}$ form the validation set $\bar{\mathcal{O}}^v$.

2) Evaluation metrics: After generating the nonconformity score of the $\mathcal{O}^c$, we evaluated uncertainty estimation on $\mathcal{O}^v$. The following three metrics are used to evaluate the performance of the proposed conformal recognition and uncertainty estimation.

a) Predictive uncertainty efficiency (PUE): PUE gauges the performance of our proposed uncertainty measurement in avoiding potential errors. For a given validate set $\mathcal{O}^v$, PUE is the rate of detected errors given by

$$\text{PUE} = \frac{|\mathbf{U} \cap \mathbf{Z}|}{|\mathbf{Z}|}, \tag{7}$$

where $\mathbf{U}$ is the set of all hand signals in $\mathcal{O}^v$ that are identified as uncertain by (6), $\mathbf{Z}$ is the set that model C generates an inaccurate non-conformal recognition $\hat{l}_t \neq l_t$, and $\mathbf{U} \cap \mathbf{Z}$ represents the errors that are detected as uncertainties by (6).

b) Conformal validated accuracy (CVA): The CVA measures the accuracy of the prediction from conformal recognition, which is defined as follows:

$$\text{CVA} = \frac{\sum_{t \in V} \mathbb{1}\{\bar{\bar{l}}_t = l_t\}}{|\mathbf{V}|}, \tag{8}$$

where $\mathbf{V}$ is the set of all hand signals in $\mathcal{O}^v$ that is NO uncertainty detected (i.e., $\mathbb{U} = 0$), and $|\mathbf{V}|$ represents its length. $\mathbb{1}\{\bar{\bar{l}}_t = l_t\}$ is the indicator function that returns "1" if the label of the conformal recognition is correct (i.e., $\bar{\bar{l}} = l_t$) and "0" otherwise.

c) Conformal validated precision (CVP): The CVP is a metric that quantifies the weighted average precision across multiple categories of hand signals (with multiple ground-truth labels), calculated from the conformal recognition set $\mathbf{V}$, in which all samples are no uncertainty detected. For hand signals with a total of L categories, CVP is computed as follows:

$$\text{CVP} = \sum_{c=1}^{L} w_c \cdot \text{Prec}_c \tag{9}$$

where $\text{Prec}_c = \frac{TP_c}{TP_c + FP_c}$, TP_c represents the true positives or the count of correct positive predictions within category c, FP_c represents the false positives or the count of wrong positive predictions within category c. $w_c \in (0, 1)$ is the weight assigned to category c, which is relative importance or frequency of each category in $\mathbf{V}$ and it satisfies $\sum_{c=1}^{L} w_c = 1$.

3) Experimental results: We used the $\bar{\mathcal{O}}^c$ to conform the well-trained models C and G and subsequently to detect uncertainties in all samples of $\bar{\mathcal{O}}^v$. The effectiveness of the proposed conformal recognition and uncertainty estimation is quantitatively assessed in Table II. Metrics including Conformal Validated Accuracy (CVA), Conformal Validated Precision (CVP), and Predictive Uncertainty Efficiency (PUE) were computed at various error levels. For comparison, we also provided "Accuracy" and "Precision" metrics; they were both based on the non-conformal recognition for the same dataset $\bar{\mathcal{O}}^v$. Since $\bar{\mathcal{O}}^v \subset \bar{\mathcal{O}}$, we find slight differences between Table I and Table II regarding these two metrics.

TABLE II: Conformal Recognition Results

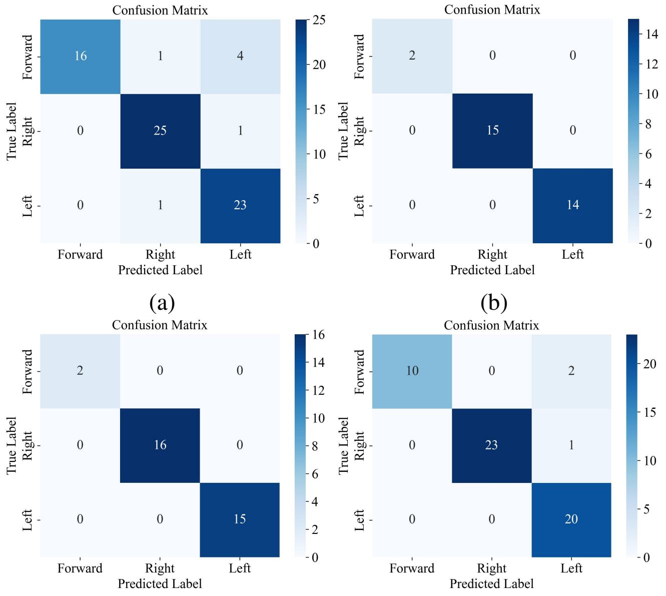
Error level α	Accuracy	Precision	CVA	CVP	PUE
0.05	0.90	0.91	1.00	1.00	1.00
0.10	0.90	0.91	1.00	1.00	1.00
0.15	0.90	0.91	0.96	0.96	0.67
0.20	0.90	0.91	0.92	0.93	0.33
0.25	0.90	0.91	0.91	0.92	0.17

The results in TABLE II, especially the CVA and CVP, illustrate that the proposed conformal recognition can achieve high performance and guarantee accurate and precise identification of hand signals at novel conditions. For the scenario of $\alpha = 0.05$ and $\alpha = 0.10$, the proposed conformal recognition achieved 100% CVA and CVP (that is, accuracy and precision after conformal validation), which avoids any potential errors. This is a significant improvement in gaining the trust of human co-workers while we deploy these smart IoT agents in open environments. One goal of the proposed conformal recognition is to effectively avoid potential errors when a learning-based IoT agent is deployed in novel conditions. The PUE metric provides a strong indicator of how efficient and effective the proposed method is on recognizing potential errors under varying uncertainty thresholds. In our experiment, under the scenarios of $\alpha = 0.05$ and $\alpha = 0.10$, it marked and avoided all potential errors in $\bar{\mathcal{O}}^v$. While increasing the level of the tolerant error α, PUE dropped as we expected since we allowed the agents to make more errors.

Fig. 2 illustrates a comparison of confusion matrices. It can be seen that as we decrease the error level α, our recognition accuracy increases in all signal categories. Notes that in the conformal recognition of Fig. 2(b), Fig. 2(c), and Fig. 2(d), no label was produced if the uncertainty was detected for an observation, thus making the number of prediction results varied in these matrices. Comparing the non-conformal results in Fig. 2(a) to other conformal results, we can also reach an interesting conclusion that the uncertainties among different categories are not consistent, and the errors correspond with uncertainties. For example, the "Forward" signal has the highest errors and also the highest uncertainties: in a total of $16 + 1 + 4 = 21$ samples, only 2 samples can be certainly recognized at $\alpha = 0.05$ and succeed in avoiding $1 + 4 = 5$ potential errors. We believe this discovery can be utilized in our future work to augment our model training process. Additionally, detected uncertainties can enable the robot to avoid potential errors and proactively seek help or clarification in a timely manner.

V. Conclusions

In this work, we presented conformal recognition that provides truth-worthy hand signal communications between humans and smart IoT agents under novel environments and with different users. It is a framework that deploys conformal prediction to measure the uncertainty in recognizing hand signals, avoiding potential errors. It utilizes a calibration set with a few representative samples in any novel conditions to enable a learned model to offer guaranteed accuracy, and subsequently, it detects the uncertainty in recognizing hand signals under a similar condition distribution. Extensive experiments showed that our conformal recognition could efficiently and effectively guarantee the accuracy of hand signal communication and avoid potential errors for out-of-distribution scenarios. We extend the theoretical discussion of uncertainty quantification found in [6] and the introductory overview in [7] by demonstrating the practical implementation and effectiveness of these methods in real-world scenarios. In summary, conformal recognition makes a promising future of enabling truth-worthy hand signal-based communication



(a)

(b)

(c)

(d)

Fig. 2: Comparison of confusion matrices: (a) non-conformal recognition on $\tilde{\mathcal{O}}^v$, (b) conformal recognition with $\alpha = 0.05$, (c) conformal recognition with $\alpha = 0.10$, and (d) conformal recognition with $\alpha = 0.15$.

to foster trust in robot-human collaboration in many voice-forbidden environments, such as construction sites and airport ramps. Our future work will focus on integrating conformal recognition in the model training process to significantly improve its generative and error resistance in novel scenarios of hand signal communication. We will also develop a framework that allows smart IoT agents to actively learn novel knowledge by detecting internal uncertainties when deployed in open environments. Specifically, we propose expanding the OOD dataset in future work to include a broader range of participants and novel environments. This extension will enable us to better assess the performance of conformal prediction in diverse and novel scenarios, further strengthening the reliability of our approach.

Acknowledgment

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

References

[1] spilve.lv, “Signals used for aircraft movement on the ramp,” accessed: 2024-04-11. [Online]. Available: www.spilve.lv/library/law/Marshaller%20Hand%20Signals.pdf

[2] 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,” *Springer Journal of Intelligent & Robotic Systems*, vol. 107, no. 3, p. 44, Mar. 2023.

[3] D. Pal, A. Singh, H. Khairnar, and A. Alladi, “OpenMarshall: Open-set recognition of aircraft marshalling signals for safe docking,” in *Proc. 3rd IEEE International Conference on Intelligent Technologies (CONIT’23)*, Hubli, India, June 2023, pp. 1–6.

[4] C. Choi, J.-H. Ahn, and H. Byun, “Visual recognition of aircraft marshalling signals using gesture phase analysis,” in *Proc. 2008 IEEE Intelligent Vehicles Symposium*, Eindhoven, Netherlands, June 2008, pp. 853–858.

[5] Y. Song, D. Demirdjian, and R. Davis, “Tracking body and hands for gesture recognition: Natops aircraft handling signals database,” in *Proc. 2011 IEEE International Conference on Automatic Face & Gesture Recognition*, Santa Barbara, CA, Mar. 2011, pp. 500–506.

[6] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY: Springer, 2005.

[7] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” *arXiv preprint arXiv:2107.07511*, Dec. 2022.

[8] N. Emaminejad, L. Kath, and R. Akhavian, “Assessing trust in construction ai-powered collaborative robots using structural equation modeling,” *Journal of Computing in Civil Engineering*, vol. 38, no. 3, Feb. 2024.

[9] 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*, vol. 31-33, no. 1, pp. 20–26, June 2020.

[10] 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,” *MDPI Sensors*, vol. 22, no. 2, p. 574, Jan. 2022.

[11] 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,” *MDPI Sensors*, vol. 23, no. 16, p. 7156, Aug. 2023.

[12] 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, Dec. 2017.

[13] 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 *Proc. IEEE GLOBECOM 2023*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 5354–5359.

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

[15] S. Kornblith, J. Shlens, and Q. V. Le, “Do better imagenet models transfer better?” in *Proc. IEEE/CVF CVPR 2019*, Long Beach, CA, June 2019, pp. 2661–2671.

[16] 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*, Mar. 2024.

[17] A. Kapitanov, A. Makhlyarchuk, and K. Kvanchiani, “HaGRID - hand gesture recognition image dataset,” *arXiv preprint arXiv:2206.08219*, Jan. 2024.

[18] 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 *Proc. 2019 IEEE Global Humanitarian Technology Conference (GHTC)*, Seattle, WA, Oct. 2019, pp. 1–7.

[19] 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 *Proc. The IEEE Winter Conference on Applications of Computer Vision*, Snowmass, CO, Mar. 2020, pp. 1459–1469.

[20] D. Biyani, N. Doohan, M. Rode, and D. Jain, “Real time sign language recognition using Yolov5,” in *Proc. 2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, Sonbhadra, India, July 2023, pp. 582–588.

[21] 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,” *arXiv preprint arXiv:1906.08172*, June 2019.

[22] R. Tibshirani, “Conformal prediction,” *UC Berkeley*, 2023.