Interactive Reinforcement Learning With Bayesian Fusion of Multimodal Advice

Susanne Trick, Franziska Herbert, Constantin A. Rothkopf, and Dorothea Koert, *Associate Member, IEEE*

***Abstract*—Interactive Reinforcement Learning (IRL) has shown promising results in decreasing the learning times of Reinforcement Learning algorithms by incorporating human feedback and advice. In particular, the integration of multimodal feedback channels such as speech and gestures into IRL systems can enable more versatile and natural interaction of everyday users. In this letter, we propose a novel approach to integrate human advice from multiple modalities into IRL algorithms. For each advice modality we assume an individual base classifier that outputs a categorical probability distribution and fuse these distributions using the Bayesian fusion method Independent Opinion Pool. While existing approaches rely on heuristic fusion, our Bayesian approach is theoretically founded and fully exploits the uncertainty represented in the distributions. Experimental evaluations in a simulated grid world scenario and on a real-world human-robot interaction task with a 7-DoF robot arm show that our method clearly outperforms the closest related approach for multimodal IRL. In particular, our novel approach is more robust against misclassifications of the modalities' individual base classifiers.**

***Index Terms*—Human factors and human-in-the-loop, multi-modal perception for HRI, reinforcement learning.**

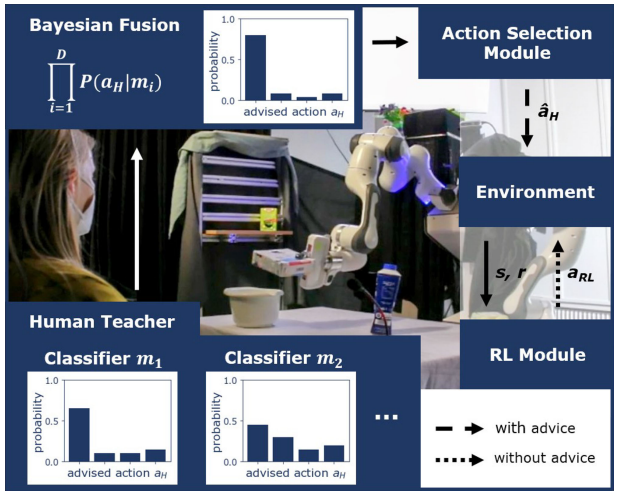


Fig. 1. We propose Multimodal IOP-Based Advice for Interactive Reinforcement Learning (MIA-IRL), using the Bayesian method Independent Opinion Pool (IOP) to combine the output distributions of the single modalities' base classifiers m_i. From the fused distribution we sample an estimated human action advice $\hat{a}_H$ to execute on the robot. When no human advice is given we use the action a_{RL} suggested by the base policy of our RL-Module.

I. Introduction

CLASSICAL industrial robots are typically designed to perform very specific and mostly repetitive tasks. In contrast, future assistive robots, which are intended to support humans in their daily lives, will be challenged by a multitude of different tasks. Since usually not all of these tasks may be known explicitly beforehand, a key component for such robots is the ability for self-improvement at runtime and adaptation to human preferences and new situations at hand.

Even though Reinforcement Learning (RL) [1], [2] offers a powerful methodology for robots to learn from direct interaction with their environment, in many practical robotic applications large state and action spaces as well as costly sample collection prevent the use of RL algorithms. This is where the novel research field of interactive RL (IRL) [3], [4] aims to improve learning speed and convergence of RL algorithms by incorporating human feedback [4] or advice [5] into the learning process.

To facilitate a beneficial interaction of everyday users with such IRL systems it is particularly important to enable ways for more natural and intuitive communication of human advice during the learning process [6]. Since humans are used to teaching other humans using natural cues such as speech, gestures, body language, gaze, or facial expressions [7], it is a central question how to best integrate such natural interaction channels into IRL algorithms. In particular, exploiting all available multimodal data can in general increase a decision's accuracy and decrease its uncertainty, which we showed in a previous work on human intention recognition [8].

Manuscript received January 19, 2022; accepted May 20, 2022. Date of publication June 13, 2022; date of current version June 28, 2022. This letter was recommended for publication by Associate Editor Mario Selvaggio and Editor Jee-Hwan Ryu upon evaluation of the reviewers' comments. This work was supported by the German Federal Ministry of Education and Research (BMBF) through Project IKIDA under Grant 01IS20045. *(Corresponding author: Susanne Trick.)*

Susanne Trick is with the Centre for Cognitive Science and Psychology of Information Processing, Technical University of Darmstadt, 64283 Darmstadt, Germany (e-mail: susanne.trick@tu-darmstadt.de).

Franziska Herbert and Dorothea Koert are with the Centre for Cognitive Science, Technical University of Darmstadt, 64283 Darmstadt, Germany, and also with the Intelligent Autonomous Systems, Technical University of Darmstadt, 64289 Darmstadt, Germany (e-mail: franziska.herbert@stud.tu-darmstadt.de; dorothea.koert@tu-darmstadt.de).

Constantin A. Rothkopf is with the Centre for Cognitive Science and Psychology of Information Processing, Technical University of Darmstadt, 64283 Darmstadt, Germany, and also with the Frankfurt Institute for Advanced Studies, Goethe University Frankfurt, 60438 Frankfurt, Germany (e-mail: constantin.rothkopf@cogsci.tu-darmstadt.de).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the ethics commitee of TU Darmstadt.

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2022.3182100, provided by the authors.

Implementation available at https://github.com/RothkopfLab/MIA-IRL.

Digital Object Identifier 10.1109/LRA.2022.3182100

Accordingly, Cruz *et al.* [9] introduced an IRL approach (termed C-IRL hereafter) which allows humans to give advice using the modalities speech and gestures. For C-IRL the authors trained an individual probabilistic classifier for each of the two advice modalities and then fused the resulting output distributions. The used fusion method reduces the decision's uncertainty if both modalities' classifiers detect non-conflicting advice and increases the uncertainty otherwise.

However, C-IRL only considers the confidence values of the predicted most likely class label and only considers probabilities above a certain threshold, thereby discarding valuable information of the single base classifiers' distributions. Additionally, in C-IRL the computation of the fused confidences is not theoretically founded but based on a heuristic tailored to exactly two modalities, and it is not discussed how to extend this computation to more modalities.

To overcome these limitations, the main contribution of this work is a new multimodal IRL algorithm that uses the Bayesian fusion rule Independent Opinion Pool (IOP) [10], [11] for combining the modalities' classifiers' output distributions (Fig. 1). As our fusion method combines individual classifiers' output distributions Bayes optimally, it reduces uncertainty correctly. Additionally, the proposed method allows straightforward generalization to more than two modalities, which is not clear in C-IRL. Because our method takes advantage of all available information of the base classifiers' distributions and computes the Bayes optimal uncertainty in the fused distribution, the action selection can be done probabilistically instead of just executing the most likely action. We evaluate our method in direct comparison to C-IRL in a simulated grid world scenario and on a real-world human-robot interaction (HRI) task, in which human participants teach a 7-DoF robot arm. The experimental evaluations show that our method clearly outperforms C-IRL, particularly in the case of partially wrong outputs of the modalities' base classifiers. Thus, we show that Bayesian fusion of modalities increases the robustness of multimodal IRL.

The rest of the letter is structured as follows. In Section II we discuss related work. Section III introduces our novel IRL approach using Bayesian fusion of multiple input modalities. In Section IV we present the experimental evaluation on theoretical corner cases, in a simulated grid world, and in a real HRI scenario. Finally, we summarize our findings and discuss future research directions in Section V.

II. Related Work

Traditionally, Interactive Reinforcement Learning allows a human trainer to give feedback on the action a robot just performed [4], [12], [13]. In contrast to this feedback-driven approach, humans also try to guide the robot on future actions by giving advice [3]. Accordingly, several IRL approaches were proposed that include human advice instead of or in addition to feedback [5], [14]–[18]. However, in many approaches the human advisors are not able to communicate their advice over natural interaction channels. In [15] the human teacher needs to use a specific programming language to interact with the learning agent. [17] proposed a computer mouse as input device for human advice, while [18] instead uses a remote control. [16] chose a graphical user interface provided on a tablet computer as input modality for advice.

More intuitive modalities for interacting with the learning agent were proposed by [14] and [5], who used speech as input source, or [19] and [20], who used facial feedback. However, humans use multiple modalities to express their intentions [21] and also their advice [9]. Accordingly, several approaches exploit multimodal input data for IRL [9], [22]–[26]. In order to teach an empathic chess partner for children, [22] combine human facial features with task-related features, e.g. if the human is winning or losing. The modalities are fused at the feature level, which however impedes generalization by exchanging or adding modalities.

In contrast, [23] propose combining the data from depth and grayscale images for a robot to learn social behavior. For both modalities, two individual Q-functions are learned, which are averaged for fusion. [24] and [25] combine facial and audio features in order to learn how to entertain people. The probability for laughing is computed individually from visual and audio cues, and the resulting probabilities are averaged for fusion. While these approaches can be straightforwardly generalized by exchanging the modalities or their respective classifiers, or by adding additional modalities, by averaging individual modalities' results, they cannot account for the uncertainty of the individual modalities' classifiers. For instance, a less certain modality has the same impact on the fused result as a more certain one and a decision's uncertainty cannot be reduced through fusion.

Cruz *et al.* [9] also use multimodal input channels for IRL, however, they explicitly consider the individual modalities' uncertainties. In their framework C-IRL, a human teacher can give advice using the two modalities speech and gestures. For each modality a separate probabilistic classifier was trained, which outputs the predicted label of the detected advice and a corresponding confidence value. The individual classifiers' outputs are combined by a heuristic fusion rule that chooses the label with the higher confidence value if the classifiers are conflicting. Furthermore, they compute a fused confidence to decrease a decision's uncertainty in case both classifiers are non-conflicting and increase it otherwise.

Although this seems to be a reasonable fusion behavior, [9] do not provide any mathematical foundation for their fusion rule, it is not sufficiently motivated why one should use exactly this function for updating the fused confidence. Moreover, their fusion discards valuable information by only considering the confidence values of the most likely classes instead of entire probability distributions and by not utilizing probabilities below a manually set threshold. Additionally, their fusion method, in particular their function for updating the fused confidence, is explicitly designed for fusing two modalities and does not straightforwardly transfer to more modalities.

In contrast to [9], we propose to use a Bayesian fusion approach. Bayesian inference was already used for inferring reward functions in inverse reinforcement learning from successive feedback [27], however not for fusing multimodal action advice for IRL. Here, we propose to use the Bayesian fusion method Independent Opinion Pool (IOP) [10], [11]. IOP provides uncertainty reduction for non-conflicting output

distributions, is theoretically founded on Bayes' rule, exploits all classifier information by considering entire probability distributions, allows to sample from the fused distribution for action selection, and is applicable to an arbitrary number of additional modalities. IOP has already been successfully applied for multimodal human intention recognition [8], and in this letter we leverage its advantages for multimodal IRL.

III. Multimodal IOP-Based Advice for Interactive Reinforcement Learning

In this work, we propose a new approach for Interactive Reinforcement Learning with multiple input modalities. Specifically, our novel method Multimodal IOP-Based Advice for Interactive Reinforcement Learning (MIA-IRL) uses the Bayesian fusion method Independent Opinion Pool (IOP) [10], [11] to incorporate multiple probabilistic base classifiers' distributions over human advice into an RL algorithm. In this section, we explain the main components of our approach, which are also illustrated in Fig. 1. We describe the agent's interaction with its environment as a Markov Decision Process (MDP) and as a core deploy a standard RL algorithm, such as Q-Learning, in our RL Module (Section III-A). We then assume a human teacher that wants to communicate intended action advice a_H to suggest to the robot which action should be performed next. The human's action advice is recognized using multiple modalities. For each modality m_i an individual base classifier is trained, which is assumed to output a categorical distribution over all possible actions $P(a_H|m_i)$ (Section III-B). Subsequently, the categorical distributions returned by all D base classifiers are fused within the Fusion Module using IOP (Section III-C). By sampling from the fused categorical distribution $P(a_H|m_1, \ldots, m_D)$ we obtain an estimate for the action proposed by the human $\hat{a}_H$, which the RL agent then executes (Section III-D). If no advice is given, the action proposed by the RL Module a_{RL} is chosen (Section III-A). For the experiments in this letter, human advice was provided in the first N episodes of learning. However, MIA-IRL could straightforwardly also incorporate distributed advice if an advisor is available over the complete learning process. Our MIA-IRL approach is also summarized in Algorithm 1.

Algorithm 1: MIA-IRL.

Require: max number of steps per episode M
init Q-table $Q[s,a]=0 \ \forall s,a$ if a possible in s, else $-\infty$
init visits per state $v[s] = 0 \quad \forall s$
init discount factor γ and exploration rate ε
init episode counter $e = 0$
while Q not converged **do**
  init steps per episode counter $j = 0$
  $s =$ random init state
  while episode not finished **and** $j <$ M **do**
    $v[s] = v[s] + 1$
    $\alpha = 1/v[s]$
    if human advice provided **then**
      for modalities $m_i = m_1, m_2, \ldots, m_D$ **do**
        $P(a_H|m_i) =$ *ModalityClassifier*(m_i)
      end for
      $P(a_H|m_1, \ldots, m_D) =$ *FusionModule*$(P(a_H|m_1), \ldots, P(a_H|m_D), Q[s])$
      $a =$ sample from distribution $P(a_H|m_1, \ldots, m_D)$
    else
      $a =$ choose ε-greedy action a from $Q[s,a]$
    end if
    execute action a, get reward r and next state s'
    $Q[s,a] = Q[s,a] + \alpha(r + \gamma \max_{a'} Q[s,' a'] - Q[s,a])$
    $s = s'$
    $j = j + 1$
  end while
  $e = e + 1$
end while

A. RL Module

The learning agent's interaction with its environment is represented as a Markov Decision Process (MDP). Thus, in a state s it takes an action a, gets a reward r, and transits to the next state s'. The agent's goal is to learn an optimal policy $\pi(s)$ in order to receive the expected maximum discounted total future reward. For the experiments in this letter, we used tabular Q-learning, which however could be replaced by other RL algorithms for different applications. The Q-function is updated according to

$$Q(s,a) \leftarrow Q(s,a) + \alpha(s)(r + \gamma \max_{a'} Q(s,' a') - Q(s,a)) \tag{1}$$

and we chose a hand-tuned discount factor $\gamma = 0.98$ and an adaptive learning rate $\alpha(s) = 1/v(s)$, which is common in literature [1], [16], where $v(s)$ is the number of times the learning agent has visited state s so far. If no human advice is given, during learning the agent chooses actions according to an ε-greedy policy with ε set to 0.1 for our experiments.

B. Classifiers for Individual Modalities

For MIA-IRL we assume base classifiers for each modality m_i that output a categorical distribution $P(a_H|m_i)$ over all possible actions a_H. For the HRI experiments in this letter, we exemplarily used two classifiers for the modalities speech and gestures. Since this work's focus is on demonstrating the benefits of applying the Bayesian fusion method IOP to IRL, these classifiers are based on off-the-shelf existing approaches. They can be straightforwardly replaced by other classifiers that return categorical distributions. In particular, adding more modalities is also possible from the mathematical formulations of the fusion method in MIA-IRL.

1) Speech: In our experiments, we chose speech as one of our modalities since it is mostly effortless and intuitive for humans to use for communicating their intentions [28]. In particular, we use keyword spotting where each keyword is assigned to an action; e.g. "milk" is the keyword for getting some milk. We use the framework Honk [29], which returns a categorical distribution over all keywords, also including the categories "silence" and "unknown". Honk is based on a Convolutional Neural Network (CNN) with two convolutional layers, one softmax layer, and Mel-Cepstrum Coefficient features as input and is implemented

in Pytorch. For training, we recorded 10 keyword utterances per word from 13 people. In addition to the 7 intended keywords (milk, flour, flower, bowl, roll, shelf, pour) we also recorded some unknown words, such as "please" or "give", that are likely to be used if people formulate their advice as a sentence. Also, noise and silence sounds were used for training. An amount of 20% of the training data for all keywords was added to the training set from the unknown words, correspondingly also 30% from the silence recordings. With a probability of 0.8, noise was added to training samples. 80% of all recorded data were taken for training, 10% each for testing and validation. Since in our experiments subjects were briefed to only use the defined keywords, before fusing the speech classifier into MIA-IRL we exclude the categories "silence" and "unknown" from the output distribution and renormalize to obtain a categorical distribution over all possible actions.

2) Gestures: Besides speech commands, humans also use nonverbal cues to communicate intentions, in particular when they refer to objects [30]. Therefore we also chose arm gestures as an advice modality. The gestures are predefined, namely pointing gestures for objects and a 2-arm symbolic gesture for the action pour. Using an RGB-D camera (Intel Realsense D435), the human skeleton is tracked based on Openpose [31]. Missing skeleton frames are interpolated using univariate splines. The tracked joint positions of arms and shoulders are aligned with the neck joint and scaled to uniform length in order to become invariant to the human-camera distance. Since we assume a gesture duration of 1 s with a skeleton tracking frame rate of 30 Hz, the resulting 30 samples of respective upper body joint positions for a gesture are transformed into a single vector as features for classification. As a classification method we chose a multiclass Support Vector Machine (SVM) with a polynomial kernel of degree 2 ($C = 1, \gamma = 0.1$), implemented using the machine learning framework Sklearn in Python. As class labels, we provide the possible actions. The trained SVM does not only return the predicted advised action but also provides a categorical probability distribution as output.

C. Fusion Module

The categorical output distributions returned by the base classifiers are fused using Independent Opinion Pool (IOP) [10], [11]. IOP fuses D categorical probability distributions $P(a_H|m_i)$ over advised actions a_H given modality data $m_i, i = 1, \ldots, D$ by multiplying them and renormalizing the resulting vector to sum to 1 in order to obtain a categorical distribution. Thus, the resulting fused distribution is

$$P(a_H|m_1, \ldots, m_D) \propto \prod_{i=1}^{D} P(a_H|m_i). \tag{2}$$

Assuming conditional independence of the categorical output distributions $P(a_H|m_i)$ returned by each modality's classifier and an uninformed prior $P(a_H)$ over actions a_H, this fusion method can be derived as probabilistically optimal by applying Bayes' rule. Its advantages are uncertainty reduction through fusion and uncertainty-dependent fusion impact [11], [32]. If the categorical base distributions to be fused are non-conflicting, the fused distribution is less uncertain than the base distributions, i.e. its entropy is lower. If instead the base distributions are conflicting, the resulting fused distribution's uncertainty is increased. Moreover, the less uncertain base distribution has a higher impact on the fused distribution than the more uncertain base distribution.

Algorithm 2: Fusion Module.

Require: classifiers' output distributions $P(a_H|m_i)$, $Q[s, :]$
// multiply distributions
$P(a_H|m_1, \ldots, m_D) = \prod_{i=1}^{D} P(a_H|m_i)$
// remove unavailable actions
for actions $a = 0, 1, \ldots$ **do**
  if $Q[s, a] == -\infty$ **then**
    remove entry $P(a_H|m_1, \ldots, m_D)[a]$
  end if
end for
renormalize $P(a_H|m_1, \ldots, m_D)$ to sum to 1
return $P(a_H|m_1, \ldots, m_D)$

Since in the defined MDP some actions are impossible in specific states, in addition to multiplying the base distributions according to IOP, the fusion module additionally excludes the probabilities of these impossible actions from the fused distribution. Then the remaining probabilities are renormalized to sum to 1. Algorithm 2 shows the complete functionality of the proposed fusion module.

D. Action Selection Module

While the proposed fusion module outputs a categorical probability distribution over all possible actions, the RL algorithm requires a discrete action to be executed. If we just chose the action with the highest probability, we would discard valuable information about the decision's uncertainty, which we intentionally wanted to consider by using probabilistic classifiers. Therefore, we propose sampling from the fused categorical distribution $P(a_H|m_1, \ldots, m_D)$ to obtain a probabilistically selected action $\hat{a}_H$ to be executed by the RL agent. If two actions' probabilities are quite similar after fusion, by sampling, each of them could be chosen to be executed instead of only the one with the slightly higher probability. Thus, we account for the system's uncertainty about the human's advice. Also, this action selection allows additional exploration, which is particularly helpful in case of imperfect base classifiers.

IV. Experimental Evaluation

In this section, we present the results of the experimental evaluation of our approach involving Bayesian fusion of multimodal advice. In Section IV-A we show the main advantages of our fusion method IOP in MIA-IRL in comparison to the fusion method in the related approach C-IRL [9] on artificial base distributions. Next, we compare the performances of MIA-IRL, non-interactive RL, and C-IRL in a simulated grid world environment (Section IV-B) and in an HRI task with a 7-DoF robot arm and 10 human subjects (Section IV-C). For all comparisons between MIA-IRL and C-IRL we replaced the fusion module and the action selection module accordingly, while using the same RL module and base classifiers.

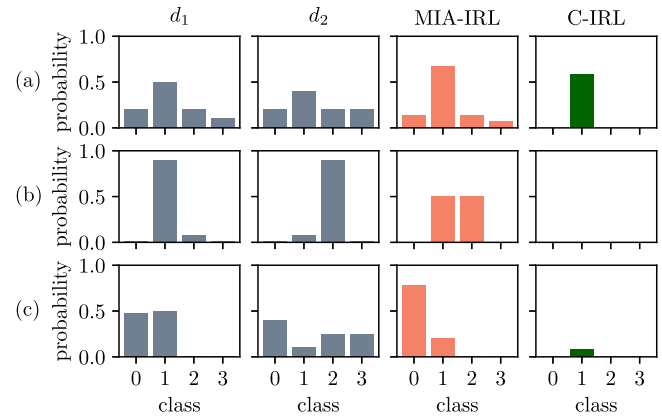


Fig. 2. Comparison of IOP in MIA-IRL and the fusion method in C-IRL [9] on exemplary base distributions d_1 and d_2. C-IRL disregards information by discarding all probabilities but the highest one (a), returns no information about which class to choose for conflicting distributions (b), or even chooses a different class than Bayes optimal IOP (c).

A. Advantages of Bayesian Fusion

As mentioned in Section II, the fusion method proposed in C-IRL [9] shares some desirable properties with our method MIA-IRL such as uncertainty reduction and uncertainty-dependent fusion impact. However, because C-IRL does not consider the complete base classifier distributions, and discards probabilities below heuristic thresholds, there exist particular situations in which the IOP fusion, which we propose for MIA-IRL, shows clear advantages. Three exemplary cases for such situations are shown in Fig. 2. In Fig. 2(a) C-IRL reduces uncertainty such as MIA-IRL but discards all probabilities apart from the highest one. However, MIA-IRL, which is Bayes optimal, assigns non-zero probabilities to all possible actions. Therefore, C-IRL's fusion method neglects the uncertainty that should be reflected in the fused distribution. It would never choose classes 0, 2, or 3, although there is a small probability that one of these classes is the correct one. Fig. 2(b) shows two conflicting base distributions. Fusion with IOP in MIA-IRL results in a distribution that assigns the same probability to classes 1 and 2. However, when fused with C-IRL all classes have a probability of 0. Thus, C-IRL disregards the fact that only classes 1 and 2 should be considered and classes 0 and 3 can be neglected. In Fig. 2(c), fusing two conflicting base distributions, C-IRL would even choose a different action than the one selected by Bayes optimal IOP in MIA-IRL. Most likely, this would lead to a misclassification by C-IRL.

These three examples highlight the theoretical advantages of the IOP fusion used in MIA-IRL compared to the fusion method in C-IRL [9]. We argue here, that these advantages lead to an increase in learning speed for IRL in particular in cases, where base classifiers may partially output wrong distributions, which we demonstrate in the following sections for a simulated grid world and a real HRI task.

B. Grid World

We first evaluate MIA-IRL in a simulated 4×4 grid world environment (Fig. 3(a)) where an agent is supposed to reach a goal while avoiding falling into one of two fires. If the agent falls into a fire, the episode ends and the agent receives a negative reward of -100. Otherwise, if the agent reaches the goal marked by the green flag, it receives a positive reward of 100. An episode may also end with a zero reward if the number of required steps in one episode exceeds 15 steps.

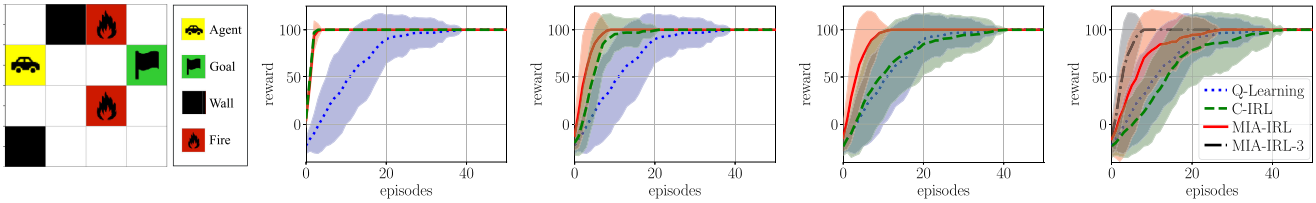
We provide simulated advice in the form of two randomly generated categorical distributions, which simulate two individual modalities' classifier outputs. This simulated advice is given during the first 10 episodes of learning.

First, we simulate correct non-conflicting categorical distributions as advice, i.e. we randomly generate two categorical distributions in which the probability for the correct action is always above 0.5. The resulting learning curves for non-interactive Q-learning, C-IRL, and MIA-IRL are shown in Fig. 3(b). Here, we plot the mean and standard deviation for the reward per episode averaged over 50 repeated runs, while for each episode we evaluate the policy 100 times and average the obtained rewards. MIA-IRL (red) as well as C-IRL (green) converge faster than standard non-interactive Q-learning (blue). A Kruskal-Wallis-Test on the convergence times of the three compared approaches showed a significant difference ($p<0.001$). The Conover-Posthoc-Test additionally provided evidence that MIA-IRL converges significantly faster than standard Q-learning ($p<0.001$). However, the convergence times of MIA-IRL and C-IRL do not differ significantly. Thus, in the case of non-conflicting correct outputs of both modalities' individual classifiers human advice speeds up learning compared to non-interactive RL, but the fusion method, either C-IRL or IOP in MIA-IRL, does not significantly influence the learning speed.

However, as real-world classifiers for human advice cannot be assumed to be always correct, next we simulate a case where one base classifier C_1 always outputs a correct distribution while the second classifier C_2 confuses the actions "right" and "left". If the correct action is "right", for C_2 a distribution with a probability above 0.5 for action "left" is randomly generated and vice versa. Fig. 3(c) shows that in this case MIA-IRL (red) converges faster than both non-interactive Q-learning (blue) and C-IRL (green). A Kruskal-Wallis significance test showed a significant difference between the convergence times of the three compared approaches ($p<0.001$). The Conover-Posthoc-Test revealed a significant difference between MIA-IRL and non-interactive Q-learning, C-IRL and non-interactive Q-learning, and MIA-IRL and C-IRL ($p<0.001$). Accordingly, MIA-IRL is more robust against partially incorrect classifier output in this case.

This effect is even stronger in a third simulated experiment, where C_2 is assumed to also confuse actions "up" and "down" in addition to "left" and "right", while C_1 is still assumed correct. Fig. 3(d) shows the corresponding learning curves. The convergence times of all approaches are significantly different (Kruskal-Wallis-Test, $p<0.001$). According to a Conover-Posthoc-Test, there is no significant difference between the convergence times of non-interactive Q-learning and C-IRL ($p = 0.34$), but a significant difference between MIA-IRL and non-interactive Q-learning ($p<0.001$) and MIA-IRL and C-IRL ($p<0.001$).

If we further modify the third simulated experiment in a way that in 20% of cases both classifiers fail, MIA-IRL still



(a) Evaluated grid world. (b) C_1 and C_2 are correct. (c) C_2 confuses right and left. (d) C_2 always incorrect. (e) In 20% both are incorrect.

Fig. 3. Learning curves of non-interactive Q-learning, C-IRL [9], and MIA-IRL for the grid world in (a) with simulated advice in the first 10 episodes. Mean rewards (lines) and standard deviations (shaded areas) over 50 runs are shown. In (b) both classifiers C_1 and C_2 advise correct actions. (c) shows the results for a correct C_1 and a C_2 confusing actions "right" and "left". In (d) C_2 additionally confuses actions "up" and "down," and in (e) in 20% of cases both classifiers fail. As (c) – (e) show, MIA-IRL clearly outperforms C-IRL if the individual modalities' classifiers partially fail. The additional curve in (e) (MIA-IRL-3) shows how performance improves by including a third classifier to MIA-IRL.

outperforms C-IRL and Q-learning significantly ($p<0.001$), as shown in Fig. 3(e). Also, since MIA-IRL is straightforwardly extendable to more than two classifiers, we can easily add a third classifier, which is correct in 60% of cases. MIA-IRL-3 using 3 advice classifiers significantly outperforms MIA-IRL and C-IRL with 2 classifiers ($p<0.001$). We expect that adding more classifiers to MIA-IRL can further increase the robustness of advice detection and by this the learning speed, depending on the quality of individual classifiers.

We conclude from the simulated experiments that if individual classifiers partially fail in detecting the correct human advice, MIA-IRL clearly outperforms C-IRL.

C. Pancake Scenario

In addition to the simulated grid world scenario, we also evaluated our approach in a real HRI scenario where human subjects can advise a 7-DoF robot arm using speech and gestures. Here, the goal of the task is that the robot should learn to assist a human in preparing a pancake batter. The task is solved successfully once the robot gets flour and milk from a nearby shelf and pours them into a bowl. The state of the robot is defined by the position of the arm, which can be AT-BOWL or AT-SHELF, the current object in the robot's hand (or the hand being empty), the positions of the objects and the current state of the bowl, which indicates if ingredients have already been poured inside. In our experiments, the objects flour, flower, and roll are always placed on the shelf, whereas the position of the milk changes between the shelf and the table between different episodes. In total, this results in 320 possible states. There are 7 actions, i.e. GO-SHELF, GET-MILK, GET-FLOUR, GO-BOWL, POUR, GET-FLOWER, GET-ROLL. The robot receives a reward of 100 if the task is solved successfully and a negative reward of -100 in case of a failure, which happens when the robot pours wrong ingredients such as flowers or the roll into the bowl or if the robot tries to get objects when already having another object in its hand. The action POUR does not only include pouring the respective ingredient into the bowl but also placing it close to the bowl on the table afterward. If the maximum number of 20 steps per episode is exceeded the episode ends with zero reward. Fig. 4 shows the full task setup. For each of the 7 actions, a speech classifier is trained to recognize a corresponding keyword and a gesture classifier to recognize a corresponding gesture. Details on the classifiers used for the experiments of this letter can be found in Section III-B1 and Section III-B2 respectively. The experimental setup was designed in a particular way to evaluate IRL in cases where classifiers may confuse intended actions. For instance, some objects are placed close to each other to cause similar pointing gestures, e.g. flour and flower, as can be seen in Fig. 4. Moreover, we chose actions with similar-sounding keywords, i.e. the keyword "roll" to get a roll (similar to "bowl") and the keyword "flower" to get a flower (similar to "flour").

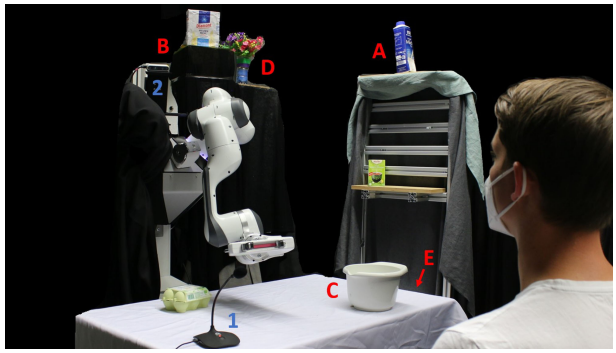


Fig. 4. The experimental setup used for evaluating MIA-IRL. A human is seated at a table to teach a 7-DoF robot arm to prepare pancake batter. The robot's task is to pour the required ingredients milk (A) and flour (B) into a bowl (C). Flowers (D) and a roll (E, not visible from shown perspective) should not be picked by the robot. The human can give advice by speech commands recorded with a microphone (1) and gestures recorded by a depth camera (2).

In the described experimental setup, we conducted experiments with 10 human participants (4 female, 6 male, 3 aged 18–25, 7 aged 26–35), who advised the robot in preparing pancake batter.[1] After a short briefing, during which the participants got familiar with the required gestures and keywords as well as the robot's movements, they carried out two experiment blocks, interrupted by a short break. In each block, the participant gave advice over the first 20 episodes, using gestures and speech commands. This choice of 20 episodes of human advice was made after preliminary experiments, and is a trade-off between performance increase and time required for each participant. In one of the blocks, MIA-IRL was used for learning and in the other block, the related method C-IRL [9] was applied. To eliminate sequence effects, 5 participants started with MIA-IRL, 5 with C-IRL.

[1]The experiments were approved by the ethics committee of TU Darmstadt on September 21, 2021 (approval code EK44/2021).

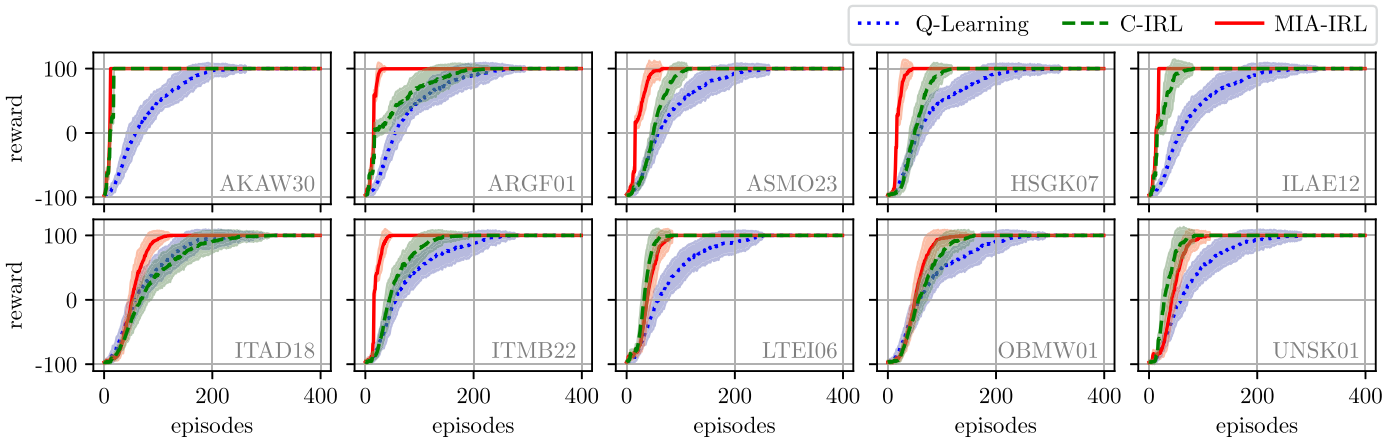


Fig. 5. Learning curves of non-interactive Q-learning, C-IRL [9], and MIA-IRL individually for all 10 participants of the pancake experiment. Each plot is labeled with the respective participant's code. Mean rewards (lines) and the corresponding standard deviations (shaded areas) over 50 runs of learning with human advice given in the first 20 episodes are shown.

For each method, after the 20 initial episodes with human advice, we let the RL agent finish the learning until convergence of the average rewards per episode. Here, we average over 50 individual runs of learning to cancel out randomness. In each episode the learned policy is evaluated 100 times, and the resulting learning curves are compared between MIA-IRL and C-IRL. In addition, we also evaluated standard non-interactive Q-learning as a baseline.

The individual resulting learning curves of all participants for MIA-IRL, C-IRL, and non-interactive Q-learning are shown in Fig. 5. For all participants MIA-IRL converges faster than non-interactive Q-learning. For 4 participants MIA-IRL and C-IRL perform similarly, while for the remaining 6 participants MIA-IRL outperforms C-IRL. The differences between participants are caused by subject-dependent variation of base classifier distributions. Particularly classifications of flour are crucial since flour is necessary for success but ambiguous for speech (similar sound flower) and pointing gestures (located next to flower). For AKAW30, LTEI06, OBMW01, and UNSK01, MIA-IRL and C-IRL perform similarly, since for all of them one classifier detects flour accurately and with high certainty while the other one is uncertain. Thus, for both methods the certain base classifier is decisive, while the fusion method, either MIA-IRL or C-IRL, has only little impact. In contrast, e.g. for ITMB22, MIA-IRL performs best, since the gesture classifier is uncertain between flour and flower with flower more likely and the speech classifier is even more uncertain with flour more likely. C-IRL favors the more certain gesture classifier and thus fails often, while MIA-IRL's fusion more often correctly chooses flour. For ARGF01 C-IRL suddenly diverges, since it only learned to solve the task from one of two starting states (milk on table). Thus, at an average reward of 50 only MIA-IRL, which learned more also for the other starting state, continues its steep increase. The base classifiers' output distributions here often match the example in Fig. 2(b), where MIA-IRL outperforms C-IRL.

In addition to the learning curves of individual participants, Fig. 6 shows the mean learning curves over all 10 participants for MIA-IRL (red), C-IRL (green), and non-interactive Q-learning (blue). MIA-IRL converges faster than both C-IRL and standard Q-learning. The Mann-Whitney-U-Test for independent samples showed a significant difference between the convergence times of MIA-IRL and standard Q-learning ($p<0.001$) and between C-IRL and standard Q-learning ($p<0.001$). The Wilcoxon-Signed-Rank-Test for dependent samples confirmed a significant difference between the convergence times of MIA-IRL and C-IRL ($p<0.001$). Thus, MIA-IRL clearly outperforms C-IRL also in real experiments with human advisors and real classifiers. In particular, the experiments show again that MIA-IRL is more robust against misclassifications of given human advice and conflicting outputs of the individual modalities' classifiers.

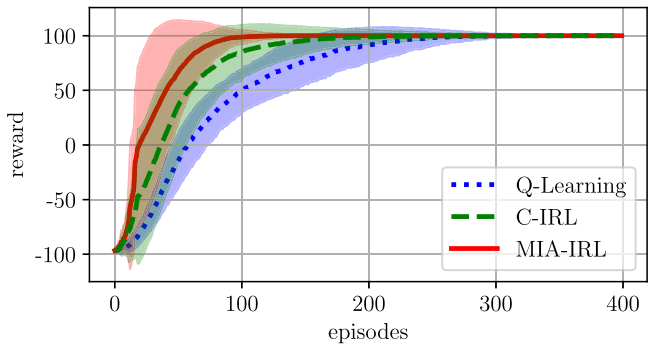


Fig. 6. Learning curves of non-interactive Q-learning, C-IRL [9], and MIA-IRL averaged over all 10 participants of the pancake experiment. Both mean rewards (lines) and the corresponding standard deviations (shaded areas) over 50 runs of all 10 participants are shown. MIA-IRL converges significantly faster than standard Q-learning and C-IRL.

V. Conclusion

In this work, we proposed MIA-IRL, a novel Interactive Reinforcement Learning approach that enables humans to advise a robot via multiple modalities, such as speech and gestures. In contrast to previous work, we fuse the modalities' classifiers' output distributions with the method Independent Opinion Pool,

which can be derived as Bayes optimal and explicitly considers the individual modalities' uncertainties correctly. Importantly, this also allows probabilistic action selection through sampling from the resulting fused distribution, instead of just choosing the most probable action, and straightforward integration of more than two modalities. In a simulated grid world scenario as well as in an HRI experiment with human participants and a real robot we showed that our approach clearly outperforms the closest related state-of-the-art approach [9]. In particular, MIA-IRL is more robust against misclassifications of the modalities' individual classifiers. Thus, MIA-IRL lays an improved solid foundation for future development of multimodal IRL.

For future work, we want to further exploit the uncertainty represented by the fused distribution. For instance, one could include an active request for additional information if the given advice is too uncertain in order to reduce the risk for catastrophic failures. Additionally, we plan to extend our fusion method to explicitly consider potential correlations between the base classifiers, evaluate MIA-IRL with additional modalities such as gaze or facial expressions, and add an additional module that learns and preserves human advice over time to enable reusing the given advice during the entire learning process. Since in the current form MIA-IRL is limited to discrete tasks, extensions for continuous tasks are also interesting to explore in future work.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[2] P. Kormushev, S. Calinon, and D. G. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, no. 3, pp. 122–148, 2013.

[3] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Real-time interactive reinforcement learning for robots," in *Proc. AAAI Workshop Hum. Comprehensible Mach. Learn.*, 2005, pp. 9–13.

[4] W. B. Knox and P. Stone, "TAMER: Training an agent manually via evaluative reinforcement," in *Proc. 7th IEEE Int. Conf. Develop. Learn.*, 2008, pp. 292–297.

[5] F. Cruz, J. Twiefel, S. Magg, C. Weber, and S. Wermter, "Interactive reinforcement learning through speech guidance in a domestic scenario," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.

[6] J. Lin, Z. Ma, R. Gomez, K. Nakamura, B. He, and G. Li, "A review on interactive reinforcement learning from human social feedback," *IEEE Access*, vol. 8, pp. 120757–120765, 2020.

[7] Y. Song, L.-P. Morency, and R. Davis, "Multimodal human behavior analysis: Learning correlation and interaction across modalities," in *Proc. 14th ACM Int. Conf. Multimodal Interact.*, 2012, pp. 27–30.

[8] S. Trick, D. Koert, J. Peters, and C. A. Rothkopf, "Multimodal uncertainty reduction for intention recognition in human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 7009–7016.

[9] F. Cruz, G. I. Parisi, and S. Wermter, "Multi-modal feedback for affordance-driven interactive reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. London, U.K.: Springer, 1985.

[11] T. R. Andriamahefa, "Integer occupancy grids: A probabilistic multi-sensor fusion framework for embedded perception," Ph.D. dissertation, Université Grenoble Alpes, Saint-Martin-d'Hères, France, 2017.

[12] F. Kaplan, P.-Y. Oudeyer, E. Kubinyi, and A. Miklósi, "Robotic clicker training," *Robot. Auton. Syst.*, vol. 38, no. 3–4, pp. 197–206, 2002.

[13] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. P. Johnson, and B. Tomlinson, "Integrated learning for interactive synthetic characters," in *Proc. 29th Annu. Conf. Comput. Graph. Interactive Techn.*, 2002, pp. 417–426.

[14] G. Kuhlmann, P. Stone, R. Mooney, and J. Shavlik, "Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer," in *Proc. AAAI Workshop Supervisory Control Learn. Adaptive Syst.*, San Jose, CA, USA, 2004, pp. 30–35.

[15] R. Maclin and J. W. Shavlik, "Creating advice-taking reinforcement learners," *Mach. Learn.*, vol. 22, no. 1, pp. 251–281, 1996.

[16] D. Koert, M. Kircher, V. Salikutluk, C. D'Eramo, and J. Peters, "Multi-channel interactive reinforcement learning for sequential tasks," *Front. Robot. AI*, vol. 7, 2020, Art. no. 97.

[17] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Reinforcement learning with human teachers: Understanding how people want to teach robots," in *Proc. ROMAN 15th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2006, pp. 352–357.

[18] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in *Proc. Int. Conf. Social Robot.*, 2013, pp. 460–470.

[19] V. Veeriah, P. M. Pilarski, and R. S. Sutton, "Face valuing: Training user interfaces with facial expressions and reinforcement learning," 2016, *arXiv:1606.02807*.

[20] G. Gordon *et al.*, "Affective personalization of a social robot tutor for children's second language skills," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, 2016, pp. 3951–3957.

[21] O. C. Schrempf and U. D. Hanebeck, "A generic model for estimating user intentions in human-robot cooperation," in *Proc. 2nd Int. Conf. Inf. Control, Automat. Robot.*, Barcelona, Spain, 2005, pp. 251–256.

[22] I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, and A. Paiva, "Modelling empathy in social robotic companions," in *Proc. Int. Conf. User Model., Adaptat., Personalization*, 2011, pp. 135–147.

[23] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Robot gains social intelligence through multimodal deep reinforcement learning," in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots*, 2016, pp. 745–751.

[24] K. Weber, H. Ritschel, F. Lingenfelser, and E. André, "Real-time adaptation of a robotic joke teller based on human social signals," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 2259–2261.

[25] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André, "How to shape the humor of a robot-social behavior adaptation based on reinforcement learning," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 154–162.

[26] F. Cruz, G. I. Parisi, J. Twiefel, and S. Wermter, "Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 759–766.

[27] H. J. Jeon, S. Milli, and A. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 4415–4426.

[28] X. Liu, S. S. Ge, R. Jiang, and C. H. Goh, "Intelligent speech control system for human-robot interaction," in *Proc. Chin. Control Conf.*, 2016, pp. 6154–6159.

[29] R. Tang and J. Lin, "Honk: A pytorch reimplementation of convolutional neural networks for keyword spotting," 2017, *arXiv:1710.06554*.

[30] G. Canal, C. Angulo, and S. Escalera, "Gesture based human multi-robot interaction," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.

[31] Z. Cao, G. HidalgoT. MartinezS. SimonWei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[32] E. Hayman and J.-O. Eklundh, "Probabilistic and voting approaches to cue integration for figure-ground segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 469–486.