Inferring Human Intentions from Predicted Action Probabilities

Lei Shi
University of Stuttgart
Germany
lei.shi@vis.uni-stuttgart.de

Paul-Christian Bürkner
University of Stuttgart
Germany
paul.buerkner@gmail.com

Andreas Bulling
University of Stuttgart
Germany
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

Inferring human intentions is a core challenge in human-AI collaboration but while Bayesian methods struggle with complex visual input, deep neural network (DNN) based methods do not provide uncertainty quantifications. In this work we combine both approaches for the first time and show that the predicted next action probabilities contain information that can be used to infer the underlying user intention. We propose a two-step approach to human intention prediction: While a DNN predicts the probabilities of the next action, MCMC-based Bayesian inference is used to infer the underlying intention from these predictions. This approach not only allows for the independent design of the DNN architecture but also the subsequently fast, design-independent inference of human intentions. We evaluate our method using a series of experiments on the Watch-And-Help (WAH) and a keyboard and mouse interaction dataset. Our results show that our approach can accurately predict human intentions from observed actions and the implicit information contained in next action probabilities. Furthermore, we show that our approach can predict the correct intention even if only a few actions have been observed.

KEYWORDS

Bayesian Inference, Deep Neural Network, Intention, MCMC

1 INTRODUCTION

A hallmark of human cognition is Theory of Mind (ToM), i.e. our ability to attribute mental states to others, such as thoughts, beliefs, or feelings. A critical requirement is the ability to understand others' intentions, i.e. their commitment to carrying out a particular action in the future [18]. This understanding enables us to anticipate others' actions [12] and is thus essential for us to engage in social communication and to interact naturally, effortlessly, and seamlessly with each other. In contrast, despite its importance for the research on human-computer interaction (HCI) and human-AI collaboration, current AI agents still lack the ability of ToM and fail to understand users' attention, predict their intentions, and anticipate their needs and actions. This limits the agents to operating after users' actions, thereby drastically restricting the naturalness, efficiency, and user experience of current interactions.

To allow AI agents to have the ability to predict the user's intention, previous works focused on predicting intentions based on Bayesian methods [1, 2, 27] and Deep Neural Networks (DNNs) [6, 10]. Bayesian-based methods can provide the uncertainty of the prediction but have the disadvantages of handling complex input data form (e.g. images) and gearing the probabilistic models for the domain they are trained in. DNN-based methods, on the other hand, are excellent at handling complex input data forms but cannot easily quantify the epistemic uncertainty in the prediction. A model that combines DNN-based and Bayesian-based methods together could have the advantages of both. This can benefit practical applications in two aspects. First, a collaborative AI agent needs to operate in the real world and it needs to deal with data with high (visual) complexity. Deploying DNNs can easily adapt to complex input. Second, the uncertainty quantification about the prediction of intention can help better provide decisions on future actions and reduce the risk of a potential wrong prediction.

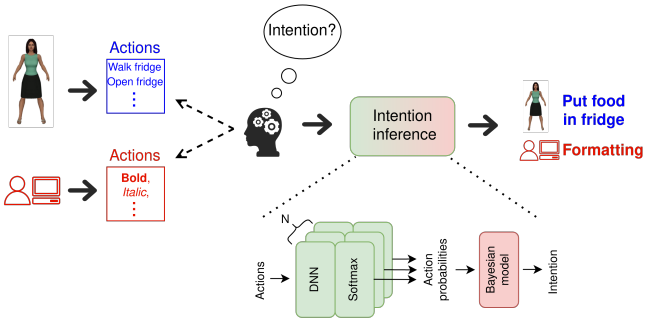

Figure 1: Overview of our proposed method to predict human intentions. An agent observes the human actions and tries to infer the human's intention. Deep Neural Networks (DNNs) together with a Bayesian model infers the human intention.

In this work, we propose a novel two-step procedure to infer human intentions from sequences of actions (Figure 1). Our approach combines DNNs to obtain the probabilities of the next action with MCMC-based Bayesian inference for inferring intentions. Specifically, given action data from N different intention, we train N DNN models for next action prediction. At test time, one action sequence data is fed to all N models to obtain the action probabilities, the action sequence represents the true intention. The output from N DNNs represents the probabilities assuming N intentions are applied. Next, we use Markov Chain Monte Carlo (MCMC) sampling to train a Bayesian model with all action probabilities from all DNNs to infer the intentions. Our two-step method decouples the next action prediction (DNNs) and actual intention prediction (Bayesian model). We do not have any requirements on the DNN input format and network architecture. They can be modified and optimised according to different tasks. The Bayesian model is independent of the DNNs, it only takes the action probabilities from DNNs and predicts the intentions with uncertainties. We demonstrate the effectiveness of our method through experiments on two datasets: Watch-And-Help (WAH) [19] and keyboard and mouse interaction dataset [36]. Our results show that our method can correctly predict the intentions of users. We further evaluate the performance of our method with 10% to 100% of observed actions in one action sequence on both datasets. The results show that using 20% of actions in a sequence, is often sufficient for the true intention

to have clearly higher posterior probability than all other intentions, although with substantial uncertainty. This demonstrates that our method can infer the users' intention already in an early stage where only a few actions have been observed.

The main contribution of this work is the two-step method to infer human intention. We use action probabilities of next action prediction from DNNs with a Bayesian model to infer intentions. To our best knowledge, we are the first ones to propose the joint use of DNNs and Bayesian models to decouple the next action prediction and intention prediction. Our method has three advantages. First, the DNNs and the Bayesian inference are decoupled. The inference of intention does not depend on the DNN architecture. One can optimise the DNN architecture for classifying the next action separately. Second, training the Bayesian model requires less time and Bayesian inference provides a fast prediction on intention. Third, our method can predict the intention correctly and efficiently when using a few observed actions in the series of actions.

2 RELATED WORK

Human-AI collaboration has attracted increasing interest recently. Several works focused on developing computational agents for collaboration in virtual environments [14, 20, 31]. The virtual environments possess near-realistic scenes and objects and support different types of actions. The importance of intention prediction in collaboration has been shown in virtual environments [19] and real-life scenarios [34]. Correctly predicted human intentions lead to more effective collaborations in robotic shared autonomy [13], Human-Robot handover [33] and cooperative assembly [17]. Several prior works have focused on action anticipation based on videos, i.e., the task of predicting future actions based on observed behaviour in the past [8, 9, 21]. Different types of models have been used , e.g. two-stream CNN [9], LSTM [8], video transformer [11], or graph neural networks [35]. In [4], the authors further used label smoothing technique to improve the work [8]. Other works have also used goals in anticipating future actions [7, 23].

In [12], the intention was defined as the intended ingredients for a sandwich. SVMs were used to predict the intention from human gaze data. SVM was also used in intention prediction during human interactions [3]. Other approaches include MDP [15], probabilistic graphical model [29], k-nearest neighbour (kNN) [22]. In [36], the authors investigated the task of predicting user intents from mouse and keyboard input as well as gaze behaviour. In another line of work, gaze behaviour was also identified as a rich source of information for predicting users' search intents [24–26] and even visually reconstructing it [30]. Perhaps the works in [16, 28] are the most similar to ours. In [28], a Bayesian model to infer intentions from ontic actions and gaze action. The ontic actions are the actions that change the state of the world and the gaze actions are the regions where an agent is looking at with regard to the ontic action. The work in [16] further introduced a deceptive component into the Bayesian model for the scenario where the human might perform ambiguous actions on purpose. Although Bayesian models were used for intention prediction, the actions were not predicted by neural networks. Rather they were pre-processed and then used in the Bayesian models.

3 METHOD

For each of the N possible intentions, we train a separate DNN on data where the ground-truth intention is known. The task of the DNNs is to predict the next action from all previous actions. Since our method works with arbitrary DNN architectures that perform this prediction task, we are not focusing on the specific architecture of the DNN here. It is important, however, that each DNN has a final Softmax (or equivalent) layer to obtain the predicted next-action probabilities. All action probabilities are then used to train a Bayesian model to predict the intention from the set of predicted next-action probabilities (see below for details). At test time, the data of each intention is forwarded to all DNNs to obtain N next-action probabilities representing N intentions. We refer the action sequence forwarded to the DNNs with the known intention label as the true intention. The N DNNs are trained with N intentions and we interpret each DNN as an assumed intention, i.e. given one action sequence, the DNNs do not know the true intention, the ith DNN assumes it is from the ith intention. The Bayesian model then uses all action probabilities jointly to infer the posterior distribution over the N assumed intention. Specifically, the Bayesian only uses the action probabilities from one action sequence to infer user intention. All DNNs can be trained separately as they do not need to share weights for our procedure to work.

Formally, an intention $\mathcal{I}$ consists of a series of actions,

$$\mathcal{I}_{ij} = [a_0, ..., a_L], 0 < i < N, 0 < j < M, 0 < k < L, \tag{1}$$

where a is action, M is the number of action series belonging to i^{th} intention and L is the total number of actions in $\mathcal{I}_{ij}$. I_{ij} represents an action series instance in the ith intention. In the training of ith DNN, we use all instances in I_i as training data. The input of for the DNN are the $\mathcal{I}_{ij}$, whereas the ground-truth next action $y_i = a_{k+1}$ constitutes the target variable. The loss function is then

$$\mathcal{L} = f(y, \hat{y}), \tag{2}$$

where $\hat{y}$ is the DNN prediction and $f(\cdot)$ is a cross entropy loss. After all networks are trained, each intention data is passed to all N networks and obtains N Softmax outputs. The i^{th} Softmax outputs produced by the i^{th} DNN represent the action probability assuming i^{th} intention is applied. To infer the intention of humans based on the series of actions and their DNN predictions serving as a surrogate likelihood, we set up the following Bayesian model:

$$\begin{aligned}
&a_k \sim \text{categorical}(\theta_k), \forall k = 1, ..., L, \\
&\theta_{km} = \sum_{i=1}^{N} P(a_{km}|\mathcal{I}_i) P(\mathcal{I} = \mathcal{I}_i), \forall m = 1, ..., M, \\
&P(\mathcal{I}) \sim \text{Dirichlet}(\alpha),
\end{aligned}$$

where $P(\mathcal{I} = \mathcal{I}_i)$ is the i^{th} element of the intention probability $P(\mathcal{I})$ to be inferred by the model, and $\alpha \in \mathbb{R}_+^N$ is the concentration vector of the Dirichlet prior on $P(\mathcal{I})$, which we set to $\alpha = 1$ to obtain an uninformative prior. The action probabilities $P(a_{km}|\mathcal{I}_i)$ of the mth possible action to occur at the kth position in the sequence are obtained from the output of DNNs. To predict the intention $P(\mathcal{I})$, we use the probabilistic programming language Stan [5], which

employs a state-of-the-art Markov-chain Monte-Carlo (MCMC) sampler to $P(\mathcal{I})$.

4 EXPERIMENT

4.1 Datasets

4.1.1 Watch-And-Help Dataset. WAH is a dataset for social intelligence and human-AI collaboration [19]. In the dataset, an AI agent Bob helps another human-like agent Alice perform household activities. The world is a 3D virtual environment. There are two stages of collaboration, i.e., the *Watch* stage and the *Help* stage. In the *Watch* stage, Bob observes Alice demonstrating an activity and Bob helps Alice with the same activity in the *Help* stage. In this work we only consider the *Watch* state given that we are interested in inferring the intentions of Alice. We understand the activities are defined by a set of sub-goals represented by predicates. Both agents can perform different actions to accomplish their goals. An activity is accomplished once the states of all sub-goal predicates are reached. In total, there are five types of activities with each activity having two to eight sub-goals. The dataset has one training and two test sets. To evaluate our method we only need information on the activity and actions and thus leave the sub-goals aside. Furthermore, to keep the activity category consistent, we focus only on those types of activities that are present in the training set and test set 1. We treat the activity as the intention and predict the intentions from the sequence of actions. Since we use the DNN to predict the next action in an action sequence, we modify the original action sequences for the use of next action prediction. For an action sequence $[a_0, ..., a_L]$, when a new action is observed, we create a new action sequence. The dataset does not have action sequences from different users, to evaluate from a user perspective, we create 92 artificial users and randomly assign action sequences to the users.

4.1.2 Keyboard and Mouse Interaction Dataset. To complement the household activities performed in the virtual environment in the WAH dataset, we also evaluated our method on keyboard and mouse interaction dataset introduced in [36]. 16 participants were asked to format text according to several formatting rules (the interaction intentions). The evaluation task on this dataset was to predict these interaction intentions from mouse and keyboard input. The text consisted of titles, subtitles and paragraphs and a rule contained instructions on how to format it using the mouse and keyboard (e.g. "make the title bold"). Participants could perform seven different actions for formatting the text. The dataset contains data from two types of formatting tasks: First, participants were asked to perform formatting according to seven predefined formatting rules. Each rule was repeated five times. Second, each participant was asked to create a custom rule themselves and to format the text according to this rule. We only used data from the first part of the dataset for our experiment since there is only one intention for each participant in the second part. We used the data from participants one to 11 for training and the data from participants 12 to 16 for testing.

4.2 Experimental Settings

We performed two experiments on the WAH dataset and the keyboard and mouse interaction dataset. We first evaluated our method on inferring users' intentions using the full action sequences. In the second experiment, we used 10% to 90% of the actions in an action sequence with a 10% step to infer the intentions. Since the WAH dataset is created in a virtual environment and the action sequences do not belong to any user, we created virtual users by randomly grouping the data in test set 1 and test set 2. As a result, test set 1 had 92 artificial users, each user had one action sequence in *put fridge*, two action sequences in *put dishwasher*, and three action sequences in *read book*. Test set 2 had nine users, each user had one action sequence in *put fridge*, five action sequences in *put dishwasher*, and five action sequences in *read book*. We show the results on test set 1 in section 5, but we conducted experiments on both test set 1 and test set 2. We observed similar outcomes, we only show results on test set 1 due to the limit of space. The architecture of our DNN model is based on the one in [20].

To train the DNNs on the WAH dataset, we used 2,000 epochs, a batch size of 32 and a learning rate of $3e^{-4}$. For the keyboard and mouse interaction dataset, we trained for 100 epochs, the batch size was eight, and the learning rate was $1e^{-4}$. To train the Bayesian model we used the same training strategy for both datasets. For each action sequence, we performed Bayesian inference via four MCMC chains, each with 2,000 iterations of which the first 1,000 were discarded as warmup. All Bayesian models converged well according to standard convergence criteria [32].

5 RESULTS

5.1 User Intention Prediction

Figure 2 shows the result of user intention prediction on test set 1 of the WAH dataset. We report the posterior mean and 90% credible intervals (CIs) of the probabilities of all assumed intentions. The top, middle and bottom plot shows the results when the true intention is *put fridge*, *put dishwasher* and *read book*. For the true intention *put fridge*, for most users our method can predict the correct intention, meaning that the assumed intention with the highest posterior mean probability is the same as the true intention. In a few cases, the posterior mean probability of put fridge is close to put dishwasher or read book. We can see that the posterior mean of put fridge and put dishwasher for user 49 is 0.43 and 0.39. When the true intention is *read book*, our method can also predict the correct intentions of most users with a few exceptions,i.e. user 13, 23, and 33. For user 13, the posterior mean of read book 0.43, only 0.02 higher than put fridge. For user 23 and 33, the posterior mean of put fridge is slightly higher than read book. For the true intention *put dishwasher*, although the posterior mean of put dishwasher are the highest in most users predictions, the difference between put dishwasher and the other two assumed intentions are smaller compared to the cases in true intention *put fridge* and *read book*. For user 6, 19, 20, 44, 49, 51, 53, the difference between the posterior mean of put dishwasher and the posterior mean of read book are around 0.1. For user 15, 27, 30, 46, 50, 56, 65, 71, and 84, the differences are below 0.07. Overall, our model can predict the correct true intention *put fridge*, however the prediction for user 49 is rather uncertain. For true intention *read book*, predictions for most users are correct but more predictions are more uncertain. The model can predict users' true intention *put fridge* and *read book* better than *put dishwasher*.

Figure 3 shows the posterior mean and 90% CIs of the predicted intentions in the keyboard and mouse interaction dataset. The

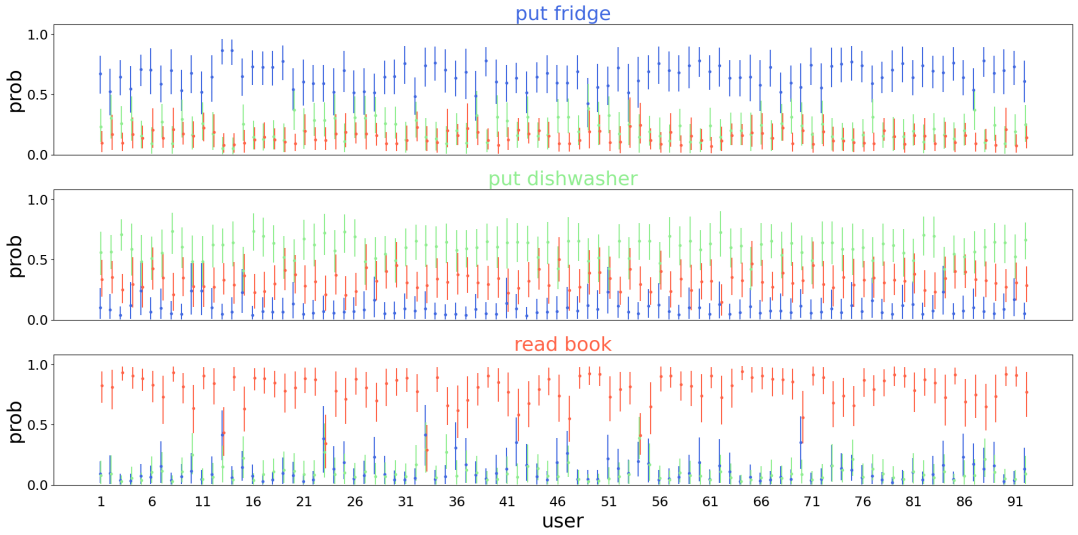

Figure 2: Result of intention prediction of users on test set 1 in WAH dataset.

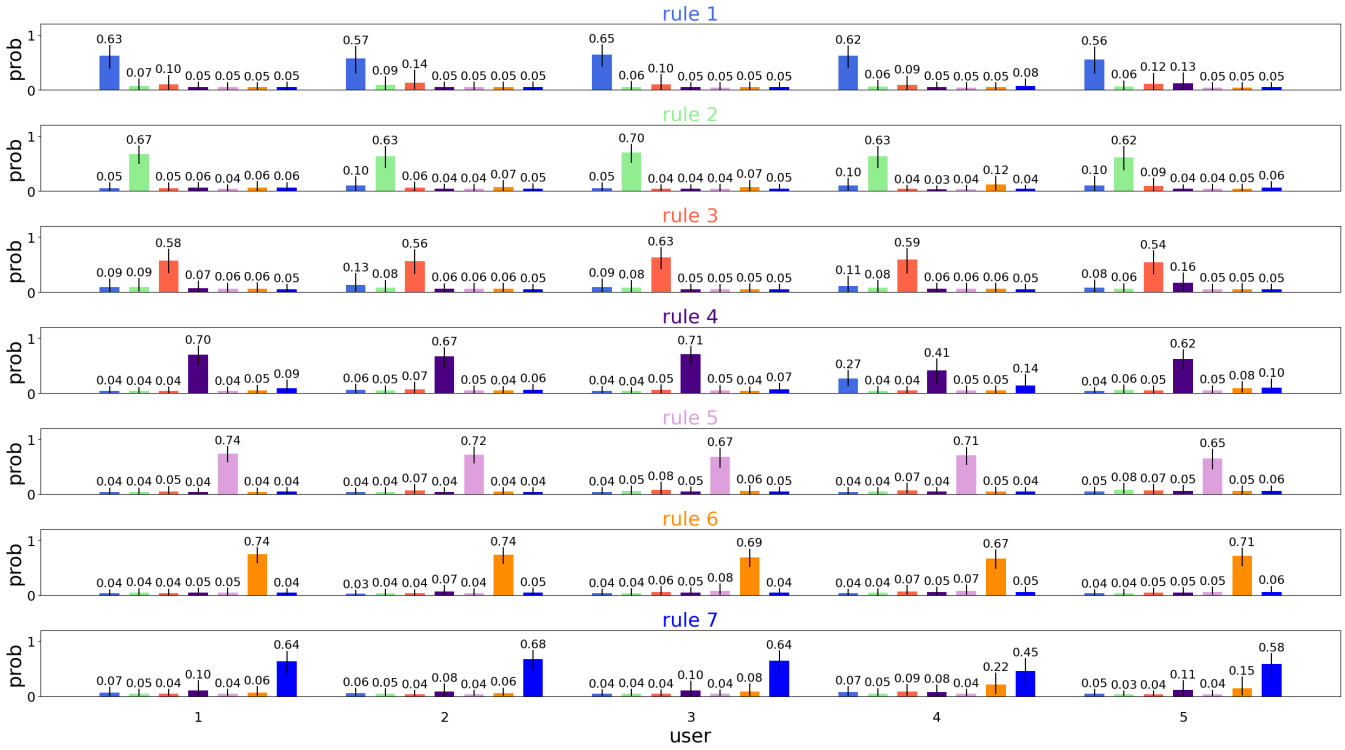

Figure 3: Result of intention prediction of users in keyboard and mouse interaction dataset.

prediction on all user data on all rules are correct in terms of the highest posterior mean of the assumed intention being the true intention. For user 1, 2,3 and 5, the differences of the posterior mean between the correctly predicted intention and the rest intentions in all true intentions are quite large. For user 4, the posterior of the correct intention for true intention *rule 4* and *rule 7* are more uncertain than the other true intentions. For true intention *rule 4*, the posterior mean of rule 4 is 0.41 while the posterior mean of rule 1 is 0.27. For true intention *rule 7*, the posterior mean probabilities of rule 7 and rule 6 are 0.45 and 0.22 respectively.

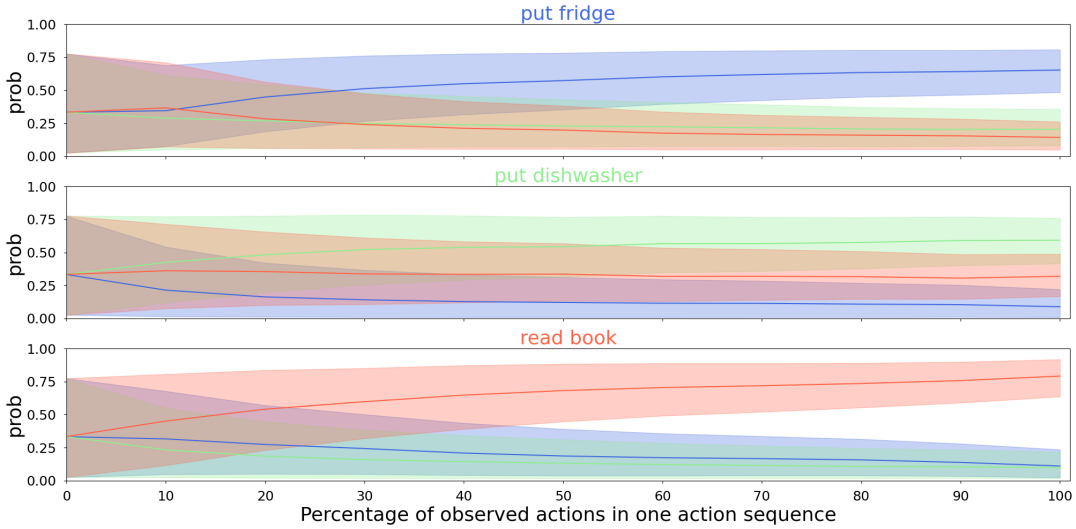

Figure 4: Posterior mean probabilities and CI bounds when different percentages of observed actions in an action sequence are used for inference. The results on test set 1 in WAH dataset are shown.

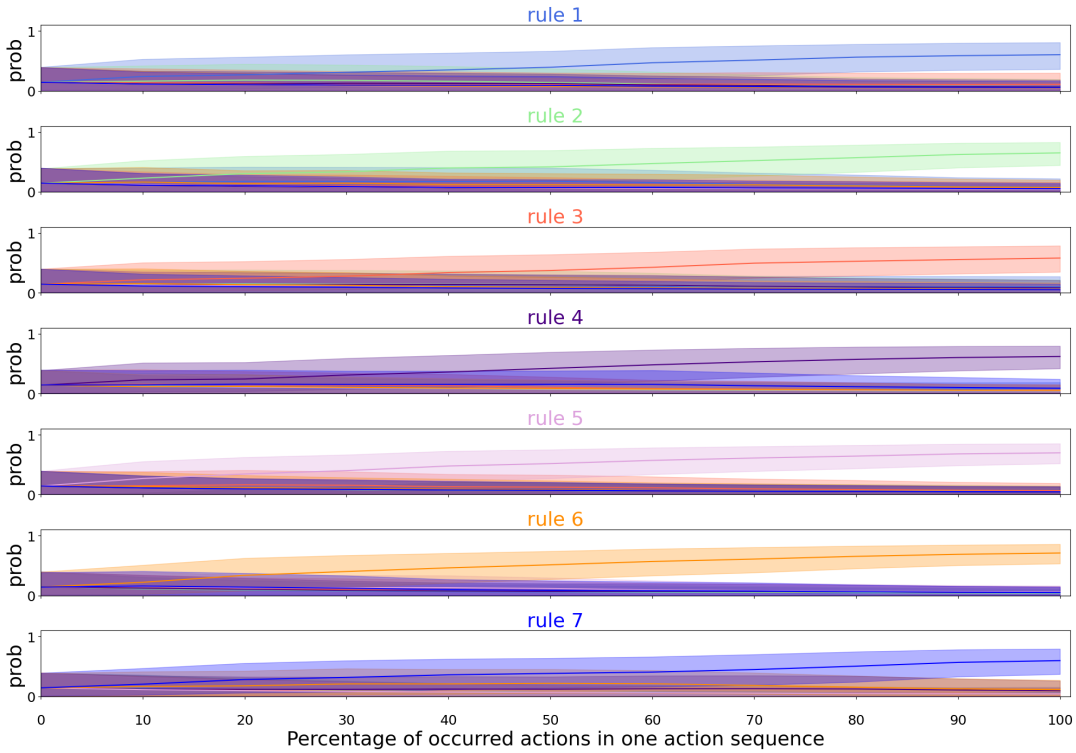

Figure 5: Posterior mean probabilities and CI bounds when different percentages of observed actions in an action sequence are used for inference. The results in keyboard and mouse interaction dataset are shown.

5.2 Different Lengths of Observed Actions in Action Sequences

Figure 4 shows the results when different percentages of observed actions in one action sequence are used for inferring intention on test set 1 in WAH dataset. The percentage of the action in a sequence is varied from 10% to 100% in steps of 10%. For instance, 50%, we extract the first 50% of the actions in the sequence and use it to infer the intention. We plot the average posterior mean

probabilities and 90% CIs, i.e. at each percentage of observed actions in one sequence, the values of posterior mean and CI bounds are the average values from all users. For all true intentions, the posterior mean probabilities of the correct intentions are already relatively higher than the other assumed intentions when 20% of action in a sequence has been observed. The posterior mean probabilities increase with the increase of observed actions in action sequences. The CI bounds decrease as the percentage of observed actions increases. This shows that the more actions have been observed in one action sequence, the more certain the Bayesian model is about its intention predictions.

When the true intention is *put dishwasher*, the predictions of Bayesian models are more uncertain than the other two true intentions. This can be observed when predicting user intention using full action sequence (Figure 2) and partially observed action sequence (Figure 4). We interpret that it is due to the noisier distribution of the actions in action sequences and the predictions of the DNNs. That is, the actions in the action sequences are not representative enough. By representative actions we mean the actions from which the intention can be easily interpreted. For instance, *open fridge* is a representative action for the intention *put fridge*.

Figure 5 shows the result on the keyboard and mouse interaction dataset. We show the average posterior mean probabilities and the 90% CIs of all seven intentions when different percentages of actions in one action sequence are used for inference. For all true intentions, the posterior mean probabilities of the correct intentions increase with more actions having been observed in the action sequences. At 10%, the assumed intention with the highest posterior mean probability is the same as the true intention for all rules but the differences are small. The predictions are still quite uncertain. At 20%, the differences in true intention *rule 2*, *rule 5*, and *rule 6* become larger, but the CI bounds remain at wide ranges. For true intentions *rule 5* and *rule 6*, the probabilities of correct intentions are close to 0.5, however, the CIs have not decreased a lot. At 50%, the CIs of the correct intention already no longer overlap with the CIs of the other intentions.

6 DISCUSSION

In the evaluation of WAH dataset, we manually created artificial users by assigning action sequences to users and most intentions of users were inferred correctly. In the keyboard and mouse interaction dataset, all predictions of all users for all true intentions were correct. We used one action sequence from one user to perform Bayesian inference. This shows the Bayesian model is efficient for inferring intention in terms of the number of observations of action sequence. It does not have to see multiple action sequences to infer the correct intention, only seeing one action sequence is adequate.

We were also interested in how the Bayesian model performs when fewer actions have been observed in the action sequences. An intuition is that the model is more confident about the inferred intention when more actions have been observed. This is confirmed by experiments in two aspects. First, the posterior mean probabilities increase when more actions are observed. Second, the ranges of CI bounds become smaller meaning the Bayesian model is more certain about its predictions. Additionally, the Bayesian model can predict the true intentions correctly even at an early stage in the action sequence. Being able to predict human/agent intention in an early stage can benefit agent-agent and human-agent interaction. For instance, in the WAH scenario, an agent can help the other agent finish a task by completing other actions in the same task. In the scenario of keyboard and mouse interaction, the computer/agent can optimise the user interface or give suggestions while the human is formatting the text. It is necessary that the agent has enough time to plan and deploy collaboration and interaction. To be able to predict intentions when only partial actions in an action sequence allows the agent to have sufficient time for planning. It is worth noting that the uncertainties of the predicted intention in early stages are relatively high and this should be taken into consideration when designing the interaction with a human.

When the number of intentions scales up, we can train the DNN in a multi-task setting, i.e. each DNN representing an intention is jointly trained and they do not share weights. This is equivalent to training separate DNNs but saves the time of training. As for the Bayesian inference, more intentions would not affect the computational time significantly.

7 CONCLUSION

In this work we proposed a two-step procedure to infer human intentions from a series of actions based on DNNs and Bayesian inference. First we trained DNNs to obtain the probabilities of predicted next action in a sequence. Then we used MCMC-based Bayesian inference to infer the human intention from the predicted next-action probabilities. We performed experiments on the WAH and keyboard and mouse interaction datasets to validate our approach. The results show that we can accurately infer the intentions even when only one action sequence from one user is available at inference time. This suggests that the implicit information contained in the next action probabilities generated by DNNs can be used to infer the intention using a Bayesian model. In addition, we demonstrated that our approach still provides correct predictions even if only a few actions have been observed.

8 ACKNOWLEDGEMENT

Lei Shi is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. Andreas Bulling is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 801708.

REFERENCES

[1] David W Albrecht, Ingrid Zukerman, and An E Nicholson. 1998. Bayesian models for keyhole plan recognition in an adventure game. *User modeling and user-adapted interaction* 8 (1998), 5–47.

[2] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[3] Roman Bednarik, Hana Vrzakova, and Michal Hradis. 2012. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the symposium on eye tracking research and applications*. 83–90.

[4] Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni Maria Farinella, and Lamberto Ballan. 2021. Knowledge distillation for action anticipation via label smoothing. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 3312–3319.

[5] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.

2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76, 1 (2017), 1–32.

[6] Sergio Casas, Wenjie Luo, and Raquel Urtasun. 2018. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*. PMLR, 947–956.

[7] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. 2020. Procedure planning in instructional videos. In *European Conference on Computer Vision*. Springer, 334–350.

[8] Antonino Furnari and Giovanni Maria Farinella. 2020. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4021–4036.

[9] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2019. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5562–5571.

[10] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. 2019. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent vehicles symposium (IV)*. IEEE, 969–974.

[11] Rohit Girdhar and Kristen Grauman. 2021. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13505–13515.

[12] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6 (2015), 1049.

[13] Siddarth Jain and Brenna Argall. 2019. Probabilistic human intent recognition for shared autonomy in assistive robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–23.

[14] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474* (2017).

[15] Hema S Koppula, Ashesh Jain, and Ashutosh Saxena. 2016. Anticipatory planning for human-robot teams. In *Experimental robotics*. Springer, 453–470.

[16] Thao Le, Ronal Singh, and Tim Miller. 2021. Goal Recognition for Deceptive Human Agents through Planning and Gaze. *Journal of Artificial Intelligence Research* 71 (2021), 697–732.

[17] Tingting Liu, Erli Lyu, Jiaole Wang, and Max Q-H Meng. 2021. Unified Intention Inference and Learning for Human–Robot Cooperative Assembly. *IEEE Transactions on Automation Science and Engineering* 19, 3 (2021), 2256–2266.

[18] Bertram F Malle, Louis J Moses, and Dare A Baldwin. 2001. *Intentions and intentionality: Foundations of social cognition*. MIT press.

[19] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. 2020. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890* (2020).

[20] Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. 2023. NOPA: Neurally-guided Online Probabilistic Assistance for Building Socially Intelligent Home Assistants. *arXiv preprint arXiv:2301.05223* (2023).

[21] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. 2021. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[22] Chen Qu, Liu Yang, W Bruce Croft, Yongfeng Zhang, Johanne R Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 25–33.

[23] Debaditya Roy and Basura Fernando. 2022. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2745–2753.

[24] Hosnieh Sattar, Andreas Bulling, and Mario Fritz. 2017. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2740–2748. https://doi.org/10.1109/ICCVW.2017.322

[25] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep Gaze Pooling: Inferring and Visually Decoding Search Intents From Human Gaze Fixations. *Neurocomputing* 387 (2020), 369–382. https://doi.org/10.1016/j.neucom.2020.01.028

[26] Hosnieh Sattar, Sabine Müller, Mario Fritz, and Andreas Bulling. 2015. Prediction of Search Targets From Fixations in Open-world Settings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 981–990. https://doi.org/10.1109/CVPR.2015.7298700

[27] Murtuza N Shergadwala, Zhaoqing Teng, and Magy Seif El-Nasr. 2021. Can we infer player behavior tendencies from a player's decision-making data? integrating theory of mind to player modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 17. 195–202.

[28] Ronal Singh, Tim Miller, Joshua Newn, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. 2018. Combining planning with gaze for online human intention recognition. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. 488–496.

[29] Dan Song, Nikolaos Kyriazis, Iason Oikonomidis, Chavdar Papazov, Antonis Argyros, Darius Burschka, and Danica Kragic. 2013. Predicting human intention in visual observations of hand/object interactions. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 1608–1615.

[30] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural Photofit: Gaze-based Mental Image Reconstruction. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 245–254. https://doi.org/10.1109/ICCV48922.2021.00031

[31] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems* 34 (2021), 251–266.

[32] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-normalization, folding, and localization: An improved $\widehatR$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis* 16, 2 (2021), 667–718.

[33] Weitian Wang, Rui Li, Yi Chen, Yi Sun, and Yunyi Jia. 2021. Predicting human intentions in human–robot hand-over tasks through multimodal learning. *IEEE Transactions on Automation Science and Engineering* 19, 3 (2021), 2339–2353.

[34] Zhikun Wang, Katharina Mülling, Marc Peter Deisenroth, Heni Ben Amor, David Vogt, Bernhard Schölkopf, and Jan Peters. 2013. Probabilistic movement modeling for intention inference in human–robot interaction. *The International Journal of Robotics Research* 32, 7 (2013), 841–858.

[35] Xinxiao Wu, Jianwei Zhao, and Ruiqi Wang. 2021. Anticipating Future Relations via Graph Growing for Action Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2952–2960.

[36] Guanhua Zhang, Susanne Hindennach, Jan Leusmann, Felix Bühler, Benedict Steuerlein, Sven Mayer, Mihai Bâce, and Andreas Bulling. 2022. Predicting Next Actions and Latent Intents during Text Formatting. In *Proceedings of the CHI Workshop Computational Approaches for Understanding, Generating, and Adapting User Interfaces*. 1–6.

Received 22 February 2024; accepted 21 March 2024