Intention Recognition in Human Robot Interaction Based on Eye Tracking

Carlos Gomez Cubero, Matthias Rehm

Intention Recognition in Human Robot Interaction Based on Eye Tracking

Carlos Gomez Cubero and Matthias Rehm

Technical Faculty of IT and Design, Aalborg University, Aalborg, Denmark
{cgcu|matthias}@create.aau.dk

Abstract. In human robot interaction any input that might help the robot to understand the human behaviour is valuable, and the eyes and their movement undoubtedly hold valuable information. In this paper we propose a novel algorithm for intention recognition using eye tracking in human robot collaboration. We first explore how the Cascade Effect hypothesis and a LSTM-based machine learning model perform to classify intent from gaze. Second, an algorithm is proposed, which can be used in a real time interaction to infer intention from the human user with a small uncertainty. A data collection with 30 participants was conducted in virtual reality to train and test the algorithm. The algorithm allows to detect the user intention up to two seconds before any user action with a success rate of up to 75%. These results open the possibility to study human robot interaction, where the robot can take the initiative based on the intention recognition.

Keywords: Human-robot interaction · Intention recognition · Eye tracking

1 Introduction

When robots start collaborating in close proximity with humans it becomes necessary for the robot to be able to predict human behavior for allowing safe interactions. On a low level this would mean predicting trajectories for human movement to prevent path overlaps and collisions with the human worker and thus ensure safety when working in the same space [12]. On a higher level this means recognizing the intent of the user in relation to the current task in order to predict the user's next action in the shared work space, e.g. predicting the need for a specific tool and being ready to give it to the user when s/he needs it. Intention recognition for such pick and place or give and take actions is both relevant for social scenarios [20] and industrial settings [2]. The reason for this is that the collaboration between robot and human will be more natural, fluent and effective, when the robot is able to predict, which object the human user will require in the next step [11].

Three sources have been identified by Liu and colleagues [11] for inferring intentions in interactions: motion intention, object arrangement (environmental layout), and daily activities (task semantics). On the other hand, studies of

human intention recognition show that especially eye gaze and head orientation are instrumental for intention recognition [6]. In this paper, we thus investigate a fourth source for inferring intentions, by utilizing eye tracking data for training a deep learning model for intention recognition in human robot collaboration.

2 State of the Art

While frequently used in collaborative robotics, intention is not clearly defined. From a cognitive point of view, intention is a high level concept that allows real-time synchronization of tasks and actions to achieve the intention. Vernon and colleagues for instance describe shared intention for cognitive robotics [21]. Schlenoff and colleagues ([17]; [16]) present an approach based on ontologies where intentions can best be described as the overall task a user wants to perform and the state of the work space is used to infer the corresponding intention based on the model. On the other side we can find low level concepts of intention that are often concerned with user movements, such as utilizing the force applied to joints of a robot arm to predict where the user wants the arm to go to [23] or where the user wants to go [22], predicting upper limb motion using FMG [1], hand motor intention with EMG [8], or computer vision to infer if an approaching user intents to interact with a robot [13]. Others focus on the manipulated objects in collaborative tasks and bind intentions to the manipulations possible with the given objects in object intention networks [7]. A similar approach is described in [10] that also integrate NLP and derived semantics in the object-guided intention recognition.

As we have seen above, different sensors have been used for predicting human intention such as EMG [8], FMG [1], EEG [5], or Computer Vision [13]. In this paper, we are investigating eye tracking as a source for intention recognition. Bader and colleagues [3] have used fixations and saccades as features for gaze-based intention classification and provide a first theoretical account of the causal relations between gaze behavior and interaction context. Singh and colleagues [19] show the advantage of gaze as a predictive cue to intention recognition. In their case, the goal is to predict a player's intention in a board game and they can show that the integration of gaze analysis in the prediction model allows to recognize the players intention ca. 90 seconds before the necessary actions take place. In contrast to a board game, collaboration in robotic scenarios is usually more fast paced but it nevertheless indicates that gaze might be a useful method for intention recognition in close proximity interactions. The Cascade Effect hypothesis was introduced by Shimojo and colleagues [18]; it states that in a selection task it is more likely that an object is selected the more attention it accumulates before the selection. It was also shown that this effect can be used to influence the selection by presenting a certain item for a longer period. Bird and colleagues [4] revised this experiment by changing the position where the items were presented obtaining similar results. The Cascade Effect thus relies on only one variable – the amount of attention – but does not take other features into account as saccades or fixations that occur when using the gaze in a cognitive

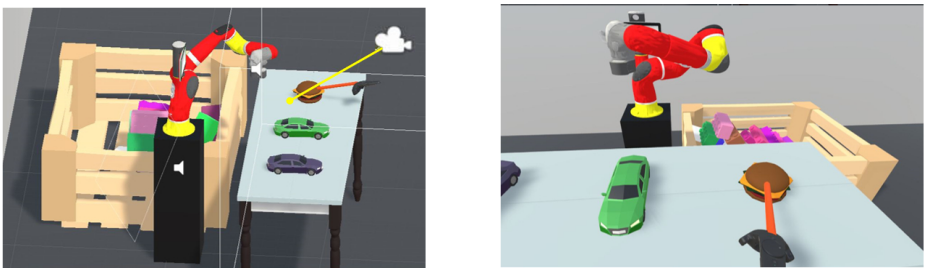

Fig. 1. Screenshots of the Virtual Reality scenario for the data collection. It contains a Sawyer robot and a table with three objects. The participant selects an item using the VR controller. Left image: an eagle eye view of the interaction, the camera icon represent the position of the headset, the yellow line shows the ray cast from the headset to the position of gaze on the table. Right image: the participant's point of view.

process. The gaze can be treated as a time series which makes it feasible to apply machine learning for predicting user intention based on the gaze behavior.

In the following we describe the data collection and training of a LSTM model for intention prediction from eyetracking data.

3 Data Collection

The data collection was designed in Virtual Reality (VR), consisting of a simple human-robot collaborative task. The task is to select an item between three possibilities on a table while a Sawyer robot is at the other side. The participant is asked to take their time and select the object that they like the most, without choosing an item at random. The robot then takes the selected item and places it in a box after which three new items appear. The participant has to perform this task several times for four different stages. In stage one, the participant has to select one of three identical bricks, in stage two, the bricks have different colors, in stage three, the bricks have different shapes, and in stage four, the participant has to decide between different objects (see Figure 1). The number of rounds in the first and second stages is smaller, as the task is more simple, in order to spend more time in the last two stages and keep the attention of the participants.

The hardware used for the data collection was a HTC Vive Headset together with a Pupil Labs VR/AR add-on eye tracker system. The computer used was a VR ready computer tower capable of running the VR experiment with around 140 frames per second. The participants were 30 students and staff from the university, 22 male and 8 female, age between 22 and 32 years and with different study backgrounds, non of them related to Robotics.

The collected data consists of the values of the position of gaze (POG) over the table, being the X axis the long side of the table and Y axis the short side, and the item selected in the end. The data was recorded with the eye tracking

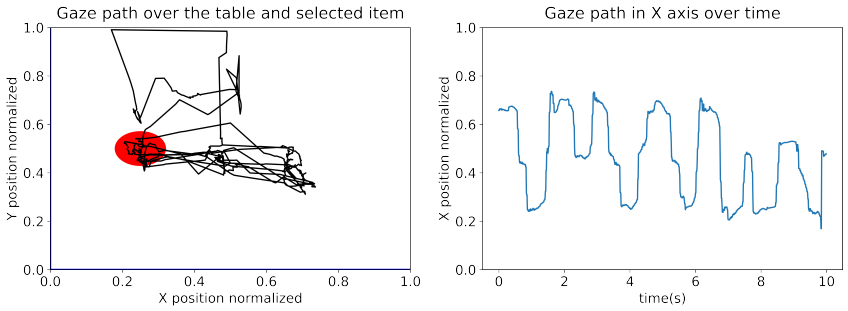

Fig. 2. Graphic representation of a series. Left image: in black a representation of the path travelled by the position of gaze over the table during an entire action, the red circle represents the position of the item selected. Right image: the plot of the X position (long side of the table) of the position of gaze of the same series over the time.

system at maximum frame rate, this is around 140Hz but it fluctuates, therefore the data-set was re-sampled to 100Hz as a common sampling rate using the nearest neighbour method. The data is then normalized to values between 0 and 1 according to the size of the table. In Figure 2 a series from the data is shown, plotting the path of the POG on the table and the position on the X axis over the time.

The duration of the selection process for the different stages varies between three and ten seconds. Figure 3 presents the histogram of duration. We collected 885 series, 162 for stage one, 189 for stage two, 272 for stage three and 261 for stage 4. The number of items selected for each of the three possibilities is evenly distributed in the dataset.

Russo and Leclerc [14] define three stages in a decision process based on gaze:

1. Orientation: occurs when the task starts and the participant orients the gaze towards the region that contains the items.
2. Evaluation: occurs when the participant gaze is scanning the items and its attention is engaged in the task. The POG jumps from one item to another in a move called saccade and between saccades the POG is hold for some instants which are called fixation [15]. During this stage the participant is performing a cognitive process that will result in a final selection.
3. Verification: is the last stage, the participant has already made a decision and now is about to signal the decision. During this stage the POG is steady on the item while the participant is moving the hand, in this case, to reach the item.

Figure 3 shows the percentage of series in which the attention is fixed in the item selected along the time. It is expected that during the verification most of the participants will fix the attention on the item as they try to reach for it. We can approximate from Figure 3 that the verification stage in this specific task takes around 1,5 seconds since there is a steady percentage of participants staring

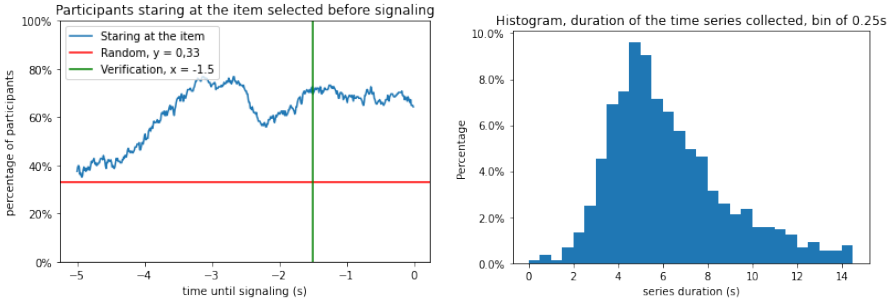

Fig. 3. Left image: for the series longer than 5 seconds, the percentage of participants staring at the item signaled withing the last 5 seconds. Used to estimate the start of the verification stage (green line). Right image: histogram of the duration of the series collected.

at the selected item. Identifying the verification stage is key to understand the amount of time that passes since the participant makes the decision until it is signaled to the system. In Section 5.3 we observe an increase in the accuracy of the trained models by removing the verification stage in the dataset.

4 Cascade Effect

According to the Cascade Effect hypothesis it is expected that the item selected on average would have more time of attention than the others. To calculate the accuracy of the Cascade Effect the table was divided in three equally sized zones in the horizontal plane, with an item placed in the middle of each zone. The attention for each item is calculated by accumulating the time the POG dwells in each zone. This parameter is stored for each zone and then compared with each other to infer the selection. Figure 4 also shows the accuracy of using this method over time, with around 75% at the time when the item is signaled and around 65% three seconds before.

5 Machine Learning

We use Long Short-Term Memory Neural Networks (LSTM) [9] as the machine learning approach. LSTMs are a type of Recurrent Neural Networks (RNN), a deep learning architecture designed to handle time series. LSTMs are widely used for both classification and regression of time series. The data used is the raw POG, which present a time series suitable for processing with a LSTM, unlike other previous work presented in the state of the art, which usually uses static analysis of fixation and saccades.

Using a LSTM-based machine learning model it is expected to achieve a classification of the series if there are features that differentiate them. As seen before, the amount of attention is already a feature, the model could easily learn

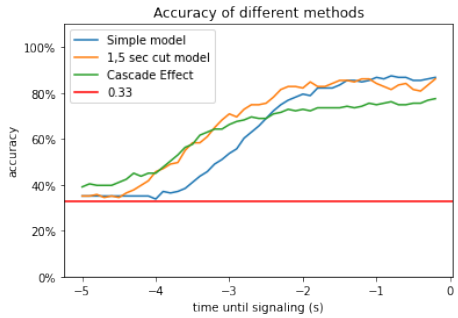

Fig. 4. Accuracy of the different methods to infer intention over the time. In blue the LSTM model trained with the dataset unedited, in orange the LSTM model trained without the verification stage (last 1.5 seconds), in green the inference based on cascade effect, in red the random inference (accuracy = 0.33).

it. But the great advantage of using machine learning is that other features can be extracted during training that otherwise would go unnoticed. The process to obtain a good model is described in the following paragraphs. In case of a different scenario this process can be repeated to find a model that suits better.

5.1 Accommodating the Data

The data collected was duplicated in two datasets, one of which was truncated the last 1,5 seconds minding the verification stage. For both datasets the series shorter than three seconds and longer than 14 seconds were discarded. Then the series were truncated to the last 5 seconds, in order to train a model with a size large enough to analyze up to 5 seconds at the same time, series shorter than that were padded with zeros at the beginning. This leave us with a data set of 755 series split in 60% for training, 20% for validation and 20% for testing.

5.2 Model Topology and Training Method

The topology of the model consists of two LSTM layers where the node size of the first layer is twice as big as the second layer, and followed by a fully connected neural network of output three for classification. The LSTM layers perform the feature extraction and the fully connected layer the classification based on the features. A training battery was conducted with different node sizes in order to find out, which what number fits the data better. To prevent overfitting and increase the accuracy of the model dropout layers were placed in between, with a dropout rate of 0,2. The loss function used was the Categorical Cross-Entropy, which is used for multi-class classification.

The model with the best fit is presented in Figure 5, a sequential model with a first LSTM layer with node size of 18, a second LSTM layer with size 9 and a fully connected layer with output 3.

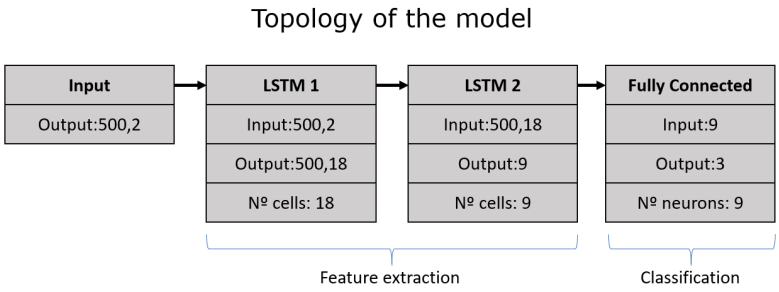

Fig. 5. Topology of the machine learning model trained for intention recognition. Consist on a sequence of two LSTM layers and a Fully Connected Neural Network. Shape and size of each layer is given with each layer.

5.3 Testing the Model

As a first test, the model input consists of series from the test dataset, obtaining a single output that corresponds to the most likely item to be selected. Both models output an accuracy around 80% in this test. To test the model's performance during and interaction (and not only at the end), the series are tested for each sample that has been logged. This is done by starting with a series full of zeros and entering the samples by the tail one by one, then the series are given to the model with each new sample. This simulates a real time operation of the model and gives a set of results dependent on the time.

The results for this test are shown in Figure 4. Both models have an identical accuracy at the end but the model trained without the last 1,5 seconds has a better accuracy. This can be explained as a consequence of training with the verification stage, because as the attention is fixed on the object it is easier for the model to classify it at the last moment without taking the rest of the series into account, thus causing overfitting. We can conclude that the model trained without verification stage performs better, classifying around 80% of the series two seconds before signaling and 70% of the series three seconds before signaling.

6 Proposed Algorithm Combining both Methods

The LSTM approach has an acceptable accuracy to infer the intention of the participant with around 80%. It can be used in scenarios where the signaling time is fixed. But it fails to work in a more relaxed interaction where the signal time is not known, as it outputs a great amount of false positives the rest of the time and therefore would not be reliable to use in most of the cases. To overcome this problem an algorithm is proposed that combines both methods. It consists in comparing the output of the LSTM model with the cascade effect plus a threshold. If the classification of the LSTM model matches with the item that accumulates more attention and this accumulated attention is higher than the threshold, then the prediction is valid.

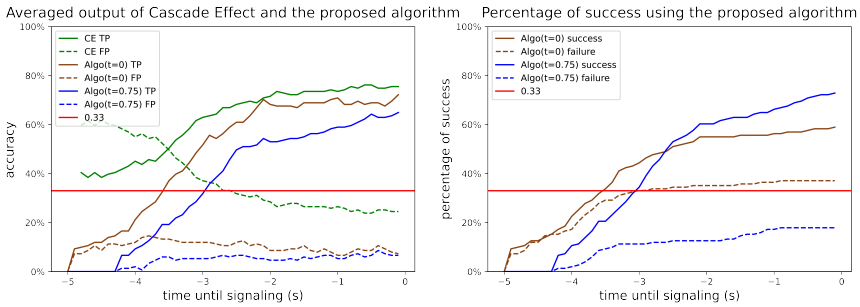

Fig. 6. Left image: comparison of the output from using Cascade Effect method (CE) and the proposed algorithm (Algo) using threshold t = 0 seconds and t = 0,75 seconds. The graph shows the True Positives (TP) with solid line, and False Positives (FP) with dotted line. Right image: the success or failure when taking the very first output of the algorithm to infer the result. The solid line represents the success, the dotted line the failures.

This drastically decreases the number of false positives but also the accuracy. Applying a threshold of 0 seconds the accuracy of the prediction at the end of the actions is close to 70% while the false positives move from 10% to 5%. If the threshold is adjusted to 0.75 seconds the prediction decreases but the false positives also decrease to a steady 5%. These results are presented and compared to the the cascade effect in Figure 6. Filtering out the false positives leads to an increase in the number of successful inferences when taking the first output of the algorithm. This can be seen in Figure 6, where the algorithm with higher threshold has an overall lower accuracy but a better percentage of success.

7 Conclusion

We can conclude that the cascade effect and the LSTM model on their own have a good performance to infer intention when the interaction has been concluded, but lack robustness for a real time interaction, due to its uncertainty the rest of the time. To decrease this uncertainty, the algorithm proposed here is based on both methods and can be used during the interaction. Tweaking the parameters allows adjusting the number of false positives and increasing reliability.

This makes it possible to study the effect of a proactive robot that could infer user intention based on the user's gaze behavior while the interaction is taking place. In such a scenario, the robot can take the initiative and reach for the item before the user has made his/her final decision. It remains to be shown if such a proactive behavior is increasing the task efficiency and trust in the robot or has a detrimental effect, e.g. due to a feeling of loss of control.

References

1. Anvaripour, M., Khoshnam, M., Menon, C., Saif, M.: Fmg- and rnn-based estimation of motor intention of upper-limb motion in human-robot collaboration. Frontiers in Robotics and AI **7** (2020). https://doi.org/10.3389/frobt.2020.573096
2. Awais, M., Henrich, D.: Human-robot collaboration by intention recognition using probabilistic state machines. In: 19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010). pp. 75–80 (2010). https://doi.org/10.1109/RAAD.2010.5524605
3. Bader, T., Vogelgesang, M., Klaus, E.: Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In: Proceedings of the 2009 International Conference on Multimodal Interfaces. p. 199–206. Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1647314.1647350
4. Bird, G.D., Lauwereyns, J., Crawford, M.T.: The role of eye movements in decision making and the prospect of exposure effects. Vision Research **60**, 16–21 (2012)
5. Buerkle, A., Eaton, W., Lohse, N., Bamber, T., Ferreira, P.: Eeg based arm movement intention recognition towards enhanced safety in symbiotic human-robot collaboration. Robotics and Computer-Integrated Manufacturing **70** (2021)
6. Duarte, N.F., Raković, M., Tasevski, J., Coco, M.I., Billard, A., Santos-Victor, J.: Action anticipation: Reading the intentions of humans and robots. IEEE Robotics and Automation Letters **3**(4), 4132–4139 (2018). https://doi.org/10.1109/LRA.2018.2861569
7. Duncan, K., Sarkar, S., Alqasemi, R., Dubey, R.: Scene-dependent intention recognition for task communication with reduced human-robot interaction. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. pp. 730–745. Springer International Publishing, Cham (2015)
8. Feleke, A.G., Bi, L., Fei, W.: Emg-based 3d hand motor intention prediction for information transfer from human to robot. Sensors **21**(4) (2021). https://doi.org/10.3390/s21041316
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (12 1997). https://doi.org/10.1162/neco.1997.9.8.1735
10. Li, J., Lu, L., Zhao, L., Wang, C., Li, J.: An integrated approach for robotic sit-to-stand assistance: Control framework design and human intention recognition. Control Engineering Practice **107**, 104680 (2021). https://doi.org/https://doi.org/10.1016/j.conengprac.2020.104680
11. Liu, T., Lyu, E., Wang, J., Meng, M.Q.H.: Unified intention inference and learning for human-robot cooperative assembly. IEEE Transactions on Automation Science and Engineering pp. 1–11 (2021). https://doi.org/10.1109/TASE.2021.3077255
12. Luo, R., Mai, L.: Human intention inference and on-line human hand motion prediction for human-robot collaboration. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5958–5964 (2019). https://doi.org/10.1109/IROS40897.2019.8968192
13. Pattar, S.P., Coronado, E., Ardila, L.R., Venture, G.: Intention and engagement recognition for personalized human-robot interaction, an integrated and deep learning approach. In: 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM). pp. 93–98 (2019). https://doi.org/10.1109/ICARM.2019.8834226
14. Russo, J.E., Leclerc, F.: An Eye-Fixation Analysis of Choice Processes for Consumer Nondurables. Journal of Consumer Research **21**(2), 274–290 (1994)

15. Salvucci, D., Goldberg, J.: Identifying fixations and saccades in eye-tracking protocols. pp. 71–78 (01 2000). https://doi.org/10.1145/355017.355028
16. Schlenoff, C., Kootbally, Z., Pietromartire, A., Franaszek, M., Foufou, S.: Intention recognition in manufacturing applications. Robotics and Computer-Integrated Manufacturing **33**, 29–41 (2015), special Issue on Knowledge Driven Robotics and Manufacturing
17. Schlenoff, C., Pietromartire, A., Kootbally, Z., Balakirsky, S., Foufou, S.: Ontology-based state representations for intention recognition in human–robot collaborative environments. Robotics and Autonomous Systems **61**(11), 1224–1234 (2013), ubiquitous Robotics
18. Shimojo, S., Simion, C., Shimojo, E., Scheier, C.: Gaze bias both reflects and influences preference. Nature Neuroscience **6**(12), 1317–1322 (2003)
19. Singh, R., Miller, T., Newn, J., Velloso, E., Vetere, F., Sonenberg, L.: Combining gaze and ai planning for online human intention recognition. Artificial Intelligence **284**, 103275 (2020)
20. Trick, S., Koert, D., Peters, J., Rothkopf, C.A.: Multimodal uncertainty reduction for intention recognition in human-robot interaction. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7009–7016 (2019). https://doi.org/10.1109/IROS40897.2019.8968171
21. Vernon, D., Thill, S., Ziemke, T.: The Role of Intention in Cognitive Robotics, pp. 15–27. Springer International Publishing, Cham (2016)
22. Wang, Y., Wang, S.: A new directional-intent recognition method for walking training using an omnidirectional robot. Journal of Intelligent Robot Systems **87**, 231–246 (2017)
23. Ye, L., Xiong, G., Zeng, C., Zhang, H.: Trajectory tracking control of 7-dof redundant robot based on estimation of intention in physical human-robot interaction. Science progress **103**(3) (2020)