Predicting Human Intent to Interact with a Public Robot: The People Approaching Robots Database (PAR-D)

Sydney Thompson
sydney.thompson@yale.edu
Yale University
New Haven, United States

Alexander Lew
a.lew@yale.edu
Yale University
New Haven, United States

Rohan Phanse
rohan.phanse@yale.edu
Yale University
New Haven, United States

Alex Huang
alex.huang@yale.edu
Yale University
New Haven, United States

Elizabeth Stanish
elizabeth.stanish@yale.edu
Yale University
New Haven, United States

Yifan Li
yli287@ur.rochester.edu
University of Rochester
Rochester, United States

Marynel Vázquez
marynel.vazquez@yale.edu
Yale University
New Haven, United States

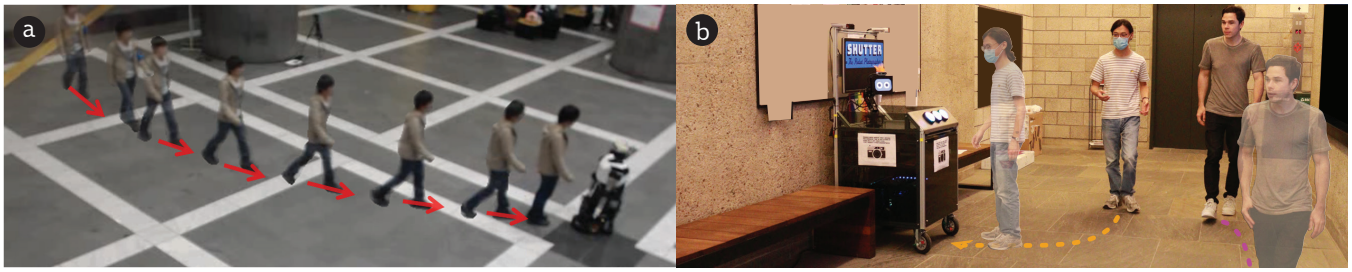


Figure 1: The People Approaching Robots Database (PAR-D) consists of a subset of the ATC Approach Trajectory dataset [28] with augmented ground truth labels (a) and two new datasets captured with a robot photographer, illustrated in (b). The lines and arrows drawn on the images represent future human trajectories. Image (a) is from https://dil.atr.jp/sets/approach_robot/.

ABSTRACT

This work studies the problem of predicting human intent to interact with a robot in a public environment. To facilitate research in this problem domain, we first contribute the People Approaching Robots Database (PAR-D), a new collection of datasets for intent prediction in Human-Robot Interaction. The database includes a subset of the ATC Approach Trajectory dataset [28] with augmented ground truth labels. It also includes two new datasets collected with a robot photographer on two locations of a university campus. Then, we contribute a novel human-annotated baseline for predicting intent. Our results suggest that the robot's environment and the amount of time that a person is visible impacts human performance in this prediction task. We also provide computational baselines for intent prediction in PAR-D by comparing the performance of several machine learning models, including ones that directly model pedestrian interaction intent and others that predict motion trajectories as an intermediary step. From these models, we find that trajectory prediction seems useful for inferring intent to interact with a robot in a public environment.

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685706

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Interactive systems and tools**.

KEYWORDS

Human-Robot Interaction; Human Behavior Forecasting

ACM Reference Format:
Sydney Thompson, Alexander Lew, Rohan Phanse, Alex Huang, Elizabeth Stanish, Yifan Li, and Marynel Vázquez. 2024. Predicting Human Intent to Interact with a Public Robot: The People Approaching Robots Database (PAR-D). In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685706

1 INTRODUCTION

Humans often indicate their social intentions through nonverbal signals, including their motion trajectory, walking speed, and gaze patterns [29]. By understanding and responding to these signals, robots can more appropriately interact with users. For example, a robot could identify users who may want to interact with it based on their motion and then proactively greet them [23], address users who are waiting their turn to approach it [10], or end an interaction in a socially appropriate manner when it seems like people want to leave [6]. Reasoning about human intention is particularly useful for robots that are deployed in dynamic environments, like transit stations [56], museums [27, 58], malls [14, 26] or universities [43, 50]. In such public settings, people typically behave in more varied ways than in controlled laboratory environments [25].

In this work, we study the problem of predicting intention to interact with a public robot. Given observations of the humans nearby a robot, the goal is to classify individual pedestrians into one of two categories: as *intending to interact* or *not intending to interact* with the robot in the near future.

We present the People Approaching Robots Database (PAR-D), a new collection of datasets for predicting a human's intent to interact with a robot in public environments, as in the scenarios of Figure 1. This database includes an augmented version of the ATC Approach Trajectory dataset [28], and two new datasets of interactions with a robot photographer on a university campus.

We investigate the complexity of the intent prediction problem on PAR-D by first establishing a human baseline for intent prediction. To this end, we conducted an online survey where participants were shown clips of examples from the database and annotated whether they thought people intended to interact with a robot. We then explored using a variety of computational approaches for modeling interaction intent on the PAR-D database. Two types of models predict intent from the observations of a pedestrian directly, while a third type predicts future motion trajectories as an intermediary step to inferring intent.

Our results suggest human accuracy on predicting interaction intent is influenced by both the environment of the robot and the length of time that the person is observed. We also found that predicting future motion trajectories as an intermediate step is more effective then predicting intent directly in more nuanced settings.

In summary, our work has three main contributions:

(1) The PAR-database, a new database for predicting human intent with a public robot which provides a richer observation space than existing public datasets.[1]
(2) A human baseline for interaction intent prediction for PAR-D.
(3) An evaluation of machine learning models for predicting interaction intent on PAR-D, including open code and models.

2 RELATED WORK

Human Engagement with Robots. The concept of interaction intent is tightly coupled with the notion of user engagement, which has been studied in Human-Robot Interaction (HRI) for many years. For example, past work has focused on defining and measuring engagement in different contexts [5, 13, 17, 32, 44], understanding how robot behaviors influence human engagement [51, 59], adapting robot behavior based on current engagement [1, 12], and identifying user disengagement [3, 6, 8, 11, 36]. See Oertel et al. [46] for a recent review on engagement in Human-Agent Interaction.

Interaction Intent in HRI. While definitions of engagement are helpful for identifying when a person has started an interaction, we are primarily interested in whether someone *intends* to engage (or interact) with a robot in the near future. In human-human interactions, intent to engage in social encounters is often expressed through proxemic behavior [29, 49], i.e., how people move and position themselves relative to other people. Prior HRI research has suggested that humans utilize similar patterns in human-robot interactions [21, 62], motivating intent prediction models that leverage historical motion information to make intention predictions [10, 18, 23, 33, 69], especially in one-on-one interaction settings. For instance, prior work has used heuristics [18] and a wide variety of machine learning models for this task [9, 10, 18, 33, 38, 60]. The latter type of approaches have demonstrated the value of using of pose features [60], proxemic features [69], and facial features [38] for intent prediction. Further, recent work has highlighted the usefulness of neural networks for this inference task [23, 60, 69].

Despite this progress, it remains an open question how computational model performance compares to human performance on intent prediction in HRI. Our work aims to address this gap by establishing human baselines for interaction intent prediction on several datasets. These baselines can help better understand the performance of machine learning models on this problem domain.

HRI Datasets for Intent Prediction. There is a limited number of open datasets to study spontaneous, public human-robot interactions. One particularly relevant dataset is the ATC Approach Trajectory dataset, which we augment in our work due to limited ground truth labels (as detailed in Sec. 3.1). To the best of our knowledge, UE-HRI dataset [7] is the only other open dataset that could be used to study our problem of interest in public environments; however, pedestrians were instructed that only one person had to be close to the robot at a time, which may have influenced groups of people approached the robot. Also, in the UE-HRI dataset, lines on the ground indicated to humans where they should stand to interact with the robot. This limited the observed human behavior and made the data collection less naturalistic than the ATC dataset.

As part of this work, we contribute two datasets in PAR-D that were collected with Shutter, a robot on a photography kiosk. The new dataset includes interactions in two public settings and contains rich pose features for pedestrians. One of our motivations to collect new data for predicting human intent to interact with a robot is a growing interest in data-driven modeling in multi-party HRI [4]. Additionally, we wanted to study the problem of interaction intent with a different robot, type of interaction, and cultural context than those considered for the ATC dataset. This effort aimed to enrich the existing corpus of HRI data and open up doors to evaluate the generalizability of different machine learning approaches across environments and interaction scenarios.

3 THE PAR DATABASE (PAR-D)

This work contributes the People Approaching Robots Database, which consists of 3 datasets for intent prediction in Human-Robot Interaction. One dataset is a subset of the ATC Approach Trajectory dataset [28], which shows Robovie in a shopping mall in Japan, with augmented labels for human intent prediction in multi-party settings. The other two datasets are new. They were collected with Shutter, a robot photographer, in two university buildings in the United States. There are three major differences between the ATC data and the Shutter data. First, the ATC data provides fewer features to describe human behavior than the Shutter data. The ATC dataset only includes position, orientation, and velocity of the head and body of people and the robot and the relative orientation of the person's head and body to the robot's body orientation. The Shutter datasets provide position, orientation, and velocity for both the

[1] https://interactive-machines.com/resources.html

robot and each human body joint as well as relative orientation and distance of the person's head and body to the robot's head and cart (see Sec. 1.2 and Table 3 in Appendix for details on dataset features). This makes the Shutter data richer in terms of human observations. Second, the visibility of pedestrians varies across the datasets. The ATC dataset was recorded with a number of environmental sensors, providing a prolonged and wide view of people around the robot, with minimal occlusions. Meanwhile, the Shutter data provides a more limited view of pedestrians, resulting in shorter observations of people nearby the robot and more occlusions. Third, the datasets differ in terms of their deployment locations and, thus, constraints on human spatial behavior. The ATC dataset was collected in an open area of a mall. There is almost no interference on peoples' trajectories from physical obstacles nearby the robot. By contrast, the Shutter data was collected in building lobbies with specific entry and exit points. The robot was placed near walls, such that it would not block pedestrian walkways. This placement is representative of non-mobile robots in public spaces, such as robots that serve as receptionists [35] or information providers [9, 10]. The next sections further describe the datasets.

3.1 Augmented ATC Dataset

The ATC Approach Trajectory dataset [28] shows people interacting with Robovie in a shopping mall. The robot had two behavioral conditions in the ATC dataset: proactively approaching a pedestrian, or passively waiting for an interaction to begin. In this work, we only considered the latter examples because we were interested in predicting human intent to interact with the robot, not exploring the effect of robots displaying intent to interact.[2] A brief description of dataset features is found in Table 4 in Appendix.

To identify scenarios when the robot passively waited for an interaction to begin, we checked the robot's velocity. In particular, we filtered out examples were the velocity was nonzero at any point in time during a scenario. This resulted in a dataset with 63 scenarios where nobody interacts with the robot and 64 scenarios which end in an interaction with the robot.

3.1.1 Data Collection Platform. Robovie is a mobile robot with a height of 120cm, 4 degrees of freedom arms and 3 degrees of freedom head. During data collection, people were tracked using environment sensors, beyond the robot's field of view. An example scenario where a person approached Robovie is shown in Fig. 1(a).

3.1.2 Data: Input Features and Ground Truth Labels. The ATC dataset includes position, velocity, and head orientation for Robovie and each pedestrian in a scenario. Up to 56 people can be observed in a single frame of data. For each scenario, the ATC dataset included labels for when a pedestrian started a conversation with the robot. Unfortunately, only one pedestrian was labeled as interacting with the robot in each positive scenario where the robot passively waited for interactions, even though other people seemed to interact with the robot as well in several cases. Thus, we augmented the original ATC interaction labels.

An annotator watched videos of a generated top-down view of the scene and labeled all people whom they considered to interact with Robovie. Interaction was defined as displaying similar behavior to the person who had already been labeled as interacting with the robot. To check that the annotation procedure was repeatable, a second annotator then followed the same procedure described above. The annotations resulted in a Cohen's Kappa score of 0.835, indicating high reliability.

In the remainder of this work, we used the original labels in the ATC dataset and the labels from the first annotator to generate ground truth targets for intent prediction. This choice maximized the internal consistency of the target values used for training machine learning models (Sec. 5). Overall, with the new labels, the dataset contains observations for 112 people who interact with the robot and 4245 people who do not interact with the robot. See the supplementary video for example visualizations of the ATC data.

3.2 Shutter Interaction Datasets

We collected two new datasets with a robot placed in the lobby of two different university buildings. We further refer to one dataset as Shutter-Lobby1 and to the other as Shutter-Lobby2.

3.2.1 Data Collection Platform. The robot, called Shutter, served as a photographer to passersby for both datasets. Shutter is an open-source social robot built on a 4 degree of freedom arm [57]. The robot's head is a screen with rendered eyes (Fig. 1(b)). The robot was mounted on a cart, which housed a desktop computer controlling the robot beneath the cart's surface. Also, the cart helped gather human input via physical buttons and record observations of the environment and interactions with two forward-facing Microsoft Azure Kinect sensors that provided body tracking data. Shutter interacted with users using motion, gaze, text-to-speech, and the buttons on the front of the cart. A monitor behind the robot communicated visual information, such as photos taken of users.

3.2.2 Interaction with the Robot. The behavior of the robot was managed with a behavior tree [16, 37]. In particular, Shutter maintained a resting pose until a person entered the field of view of either Kinect sensor. Then, it randomly selected one of three greeting behaviors: looking away from the person, tracking the person with gaze and motion, or tracking the person with gaze and motion while verbally greeting them. The interaction advanced if the person pressed any of the cart's buttons. In this case, the robot explained its role as a photographer, took a photo of the user with a camera attached to its head and showed the picture on the screen behind it, praising it verbally. Interactions could be halted by users at any time by exiting the robot's field of view, after which the robot returned to a resting pose.

3.2.3 Data Collection Procedure. Our data collection process took place on 13 days over three weeks in August 2022, resulting in a total of 5 hours of data with 1059 unique recordings. In these recordings, 2429 people passed by Shutter, 393 of whom interacted with the robot. Deployments occurred at different times to observe varied pedestrian patterns throughout a working day. On each day of recording, we placed the robot in its deployment location (either Lobby1 or Lobby2) and monitored the platform from an adjacent hallway to remove the influence of robot handlers on users [55]. The system was restarted if an operational failure arose.

The layout of the two environments where Shutter was deployed are shown in Figure 2. The robot was placed close to building lobbies,

[2] See https://dil.atr.jp/ISL/sets/approach_robot/ for more details about the original data.

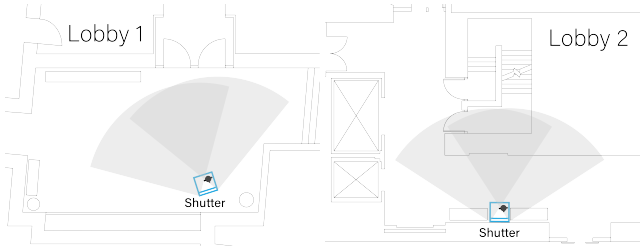


Figure 2: The two deployment locations for the collection of the Shutter Interaction dataset. The gray areas represent the field of view of each of Shutter's cameras.

nearby primary entry and exit points of the area and oriented against a wall to ensure all approaching people were visible. In Lobby1 (Figure 2, left), people could approach the robot frontally or from either side. In Lobby2 (Figure 2, right), people approached the robot from the sides. Both lobbies were accessible from public streets, so campus visitors, students, and staff passed the robot in the course of its deployment. See the supplementary video for example clips that show how the robot was deployed in both lobbies.

3.2.4 Data: Input Features and Ground Truth Labels. A frame of a scenario includes position and orientation information about each pedestrian and the robot. All coordinates and angles are in a global coordinate frame specific to each environment because the robot's position varied slightly during each deployment. More details on the input data can be found in the supplementary Appendix.

Each person in the Shutter Interaction dataset was labeled as *interacting* or *not interacting* with the robot based on two criteria: 1) whether a button was pressed and the person was standing in close proximity to the robot, and 2) whether the person was moving very slowly, facing the robot, and in close proximity to the robot. Close proximity was defined as 3.6 meters, based on Hall's proxemic zones [20]. Slow walking speed was defined as less than .65 meters/second, which is half of a typical adult walking speed [53]. To reduce noise in the response variable, if either of the criteria were true for at lease three consecutive frames, the person was labeled as interacting with the robot.

3.3 Intent Prediction with PAR-D

In PAR-D and the rest of this paper, we view intent prediction as mapping a history of d features of a pedestrian's behavior over T time-steps to whether or not the person intends to interact with a robot within 4 seconds in the future. We chose a 4 second cut-off based on the work by Bohus and Horvitz [9]. It is important to acknowledge that human intent is a non-observable value. We approximate this value under the assumption that if someone interacts with the robot – as annotated in the ATC and Shutter datasets – it is because they intended to do so in the near past.

With the machine learning models analyzed in Sec. 5, we approximate the above mapping with a function $f : \mathbb{R}^{d \times T} \rightarrow \{0, 1\}$, trained with a labeled dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{X}^i, y^i)\}_{i=1...N}$ of input features $\mathbf{X}^i \in \mathbb{R}^{d \times T}$ and corresponding targets $y^i \in \{0, 1\}$ from PAR-D. The input, $\mathbf{X}^i$, is a sequence of features for an individual pedestrian: $\mathbf{X}^i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_T^i]$, where each $\mathbf{x}_t^i$ includes at least the position and velocity of the person i at time t. As in prior work [9], the target y^i is computed based on whether the pedestrian i indeed interacts with the robot in the near future. Before we discuss machine learning methods to compute f, we present a human baseline for this prediction problem in the next section.

4 HUMAN PREDICTION BASELINES

We created a human baseline for predicting intent in a subset of PAR-D as a reference to better understand the performance of computational models. To this end, we recruited 84 annotators through Prolific,[3] a tool for crowd-sourcing responses to online surveys. The annotators watched video representations of scenes from one of the three PAR-D datasets, and predicted whether a specific pedestrian would interact with the robot within 4 seconds after the video ended, per our problem formulation for intent prediction in Sec. 3.3.

In the annotation survey, each video clip showed a specific, highlighted person for 0.2 to 4 seconds from a given sequence, an interaction scenario where people pass by or interact with a robot in PAR-D. Multiple people from the same sequence were often used as separate examples. This was an important consideration because humans can affect the way other pedestrians interact socially [20, 42] and we wanted to see, for example, whether people that shared physical space nearby a robot were annotated in accordance to the ground truth values for interaction intent in PAR-D.

4.1 Annotation Survey

Each survey included 20 clips from a single dataset in a randomized order and was completed by 7 different annotators via Prolific. The annotators were compensated at a rate of $16 USD per hour for completing the survey. Further, at the start of the survey, they were incentivized to perform accurately with a performance-based bonus of $0.05 USD for each correct intent prediction.

4.1.1 Training: The survey first introduced the robot and showed example clips from the dataset. Then, it asked annotators to answer two practice questions before annotating clips for our human baseline. These practice questions served to clarify the annotation task and reduce errors. One practice question showed a positive example of a person intending to interact with a robot, and the other showed a negative example. Both examples included an explanation of the correct annotation, which the participant could see after they answered the question.

4.1.2 Annotations: Each participant was shown a series of clips, each with a single person highlighted. For each clip, they were asked *"Will the highlighted person interact with the robot?"* and could choose among two answers. One answer was *"Yes, the highlighted person will interact with the robot within 4 seconds of the end of the clip"*; the other was *"No, the highlighted person will not interact with the robot within 4 seconds of the end of the clip"*.

The next two sections describe in detail how we created the specific clips that were used for computing the human baselines in PAR-D. Then, we present the results from this annotation effort.

[3] https://prolific.com

4.2 Annotated Data Samples for ATC

For the ATC dataset in PAR-D, a total of 80 examples were selected for annotation. To select these examples, we first split the dataset into 5 folds because we were interested in evaluating the performance of machine learning models across different partitions of the data with cross-validation (as later described in Sec. 5. Preliminary results on the folds helped us identify the easiest ("Easy") and hardest ("Hard") folds in the data, from which we then chose a pool of 40 examples each – 20 positive and 20 negative – for the human baseline. The specific procedure used to select these examples is in the Appendix in the supplementary material.

For the annotation survey for the ATC dataset, each video clip was exactly 20 frames (4 seconds long). The start time for each clip was chosen uniformly between the first time the highlighted person was visible and 4 seconds before the last time they were visible. The clips were created as 2D visualizations, similar to the examples shown in the supplementary video but without revealing the ground truth labels for the examples.

4.3 Annotated Data Samples for Shutter

Similar to the ATC dataset, we selected 80 examples from the Shutter-Lobby1 and 80 examples from the Shutter-Lobby2 datasets for the human baseline. However, due to a more limited field of the environment and more occlusions in the Shutter data than the ATC data, fewer people were visible in a single sequence for Shutter and the sequences were much shorter. As a result, we selected the examples for the human baselines for the Shutter datasets by choosing examples with special consideration for how long a person was visible from the Kinect sensors on Shutter's cart.

We balanced the example pools for the Shutter-Lobby datasets by both positive and negative examples and by the window of time that a person was visible. We grouped the length of time that a person is visible into four time windows of $(0, 1]$, $(1, 2]$, $(2, 3]$, and$(3, 4]$ seconds. In particular, for a given Shutter dataset, we selected 10 positive and 10 negative examples for each time window. The specific procedure used is described in Section 3 of the Appendix.

For the annotation surveys for the Shutter-Lobby datasets, each clip could be between 1 frame (0.2 seconds) and 20 frames (4 seconds). The start time for each clip was chosen uniformly from the time that the highlighted person was visible. The clips were created as 3D visualizations, similar to the examples shown in the supplementary video but without revealing the ground truth labels.

4.4 Human Prediction Performance

4.4.1 Prediction Metrics. For a given example, a prediction label for the human baseline was determined by the mode of annotators' responses. Then, we compared the predictions with the ground truth labels from the datasets, and evaluated performance with standard classification metrics: accuracy, F1 score, Precision, and Recall. We also measured the average entropy of the annotations for each dataset, which was calculated across participants for each clip. Entropy provides information about how much agreement there was in the annotations for a given clip.

4.4.2 Results. Table 1 shows the performance of human annotators at predicting pedestrians' intent to interact with Robovie and Shutter. Human prediction performance was best on the easiest ATC fold ("Easy Fold" in Table 1), reaching 0.88 in Accuracy and 0.86 in terms of F1 score. Performance on the hardest fold for the ATC dataset was also high, with an F1 score of 0.83. Meanwhile, the human baseline performance on the Shutter-Lobby datasets was lower in terms of both Accuracy and F1 scores. For example, in Lobby1 and Lobby2, the F1 score was 0.72 and 0.65, respectively.

Dataset	*Acc.*	*F1*	*P*	*R*	*Avg. Entropy*
ATC - Easy Fold	0.88	0.86	0.94	0.8	0.27
ATC - Hard Fold	0.85	0.83	0.79	0.88	0.33
Shutter-Lobby1	0.67	0.72	0.63	0.83	0.39
Shutter-Lobby2	0.65	0.65	0.65	0.65	0.42

Table 1: Human baseline performance on the ATC, Shutter-Lobby1 and Shutter-Lobby2 datasets. "Acc." is Accuracy, "P" is Precision, "R" is Recall, and "Avg." is Average.

Shutter-Lobby1 Dataset					
Visibility (secs.)	*Acc.*	*F1*	*P*	*R*	*Avg. Entropy*
$(0, 1]$	0.80	0.82	0.75	0.9	0.40
$(1, 2]$	0.5	0.62	0.50	0.80	0.40
$(2, 3]$	0.75	0.76	0.73	0.80	0.39
$(3, 4]$	0.65	0.70	0.62	0.80	0.38
Shutter-Lobby2 Dataset					
Visibility (secs.)	*Acc.*	*F1*	*P*	*R*	*Avg. Entropy*
$(0, 1]$	0.5	0.5	0.5	0.5	0.5
$(1, 2]$	0.75	0.67	1.0	0.5	0.42
$(2, 3]$	0.85	0.87	0.79	1.0	0.40
$(3, 4]$	0.50	0.55	0.50	0.60	0.35

Table 2: Human baseline performance based on the time a person was visible in the Shutter-Lobby datasets. "Acc." is Accuracy, "P" is Precision, "R" is Recall, and "Avg." is Average.

The F1 scores of the human baselines correlated with agreement among the annotators, e.g., as the easiest ATC fold had the lowest average entropy (indicating highest agreement).

Table 2 breaks down the results by the length of time that a person was visible by Shutter. Disagreement among annotators reduced from 0.4 to 0.38 and from 0.5 to 0.35 as the duration of pedestrian observations increased for both Shutter-Lobby1 and Shutter-Lobby2, respectively. In terms of Accuracy and F1 score, there seemed to be a complex relationship between performance, how long a person is visible, and specifics of the Shutter datasets. For example, in Shutter-Lobby1, Shorter observations with a length of (0,1] seconds led to highest performance and a length of (2,3] seconds led to second-best performance. But a length of (1,2] seconds had worst performance. The results are different for Lobby2 where, for instance, the (0,1] length led to worst prediction performance.

4.5 Discussion of Human Baselines

We suspected that the Shutter-Lobby datasets could result in higher human prediction performance because they provide much richer observations of pedestrians than ATC, allowing us to visualize the data in 3D rather than 2D. However, humans did better at predicting intent in the ATC examples, suggesting that the examples in the Shutter datasets were more complex or difficult for humans to label.

There are many factors that could have contributed to the difference in performance between the Shutter and ATC data. At first glance, one may think that the difficulty stems from the length of time people were visible for in the Shutter data; however, this is unlikely. The reason is that the F1 scores for the longest-duration clips in the Shutter-Lobby datasets was much lower than the scores for the ATC datasets (compare Tab. 1 with Tab. 2). Other factors that could have led to lower performance in the Shutter data are the type of visualizations and noise in the pedestrian poses. In terms of visualizations, prior research in HRI suggests that different human view-points can affect the legibility of robot motion [45]. Thus, the specific type of visualizations (3D vs. 2D) and camera angle that we used to render the clips for the annotation surveys could have affected the legibility of human behavior in PAR-D. Also, there are more environmental constraints on spatial behavior in the Shutter-Lobby datasets than the ATC dataset, but none of these constraints were rendered in the visualizations. Perhaps the lack of this information led annotators to largely agree on the interpretation of an example, given the similar average entropy for Shutter and ATC; but that interpretation was correct less often Shutter than ATC. Finally, the increased noise in the pedestrian observations in the Shutter data could have negatively affected human prediction performance, especially because the human poses were generated from fusing observations from two Kinects. The latter challenges with noise in human observations are common for real-world robot deployments; despite the lower human performance, we believe they make the Shutter-Lobby datasets valuable to the community.

Another important finding from the human baselines is that the duration for which a person is visible can impact the predictability of their intent to interact with a robot, although we did not observe a simple relationship between human prediction performance and visibility duration. The complexity of this relationship could be due to specifics of the environment in which the Shutter-Lobby datasets were recorded. However, for our human baseline, annotators visualized pedestrian behavior similar to how our machine learning models make predictions in Sec. 5, focusing on human behavior relative to the robot and without awareness of obstacles. In the future, considering more spatial constraints can potentially increase human baseline performance, given that these constraints can impact how people behave spatially around a robot [9].

5 COMPUTATIONAL MODELS OF INTENT

We investigate the performance of a variety of machine learning models on PAR-D. We classify the models into two main types (Random Forest models, and Encoder-Decoder models) based on their computational structure. All models were all implemented following the problem formulation for intent prediction in Sec. 3.3. For training and evaluation, the datasets were resampled to 5Hz.

5.1 Random Forest Models

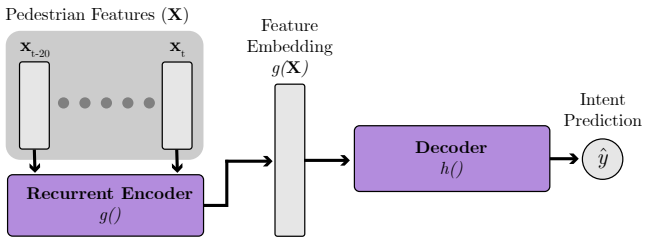
We use random forest (RF) models to directly map observations of pedestrians to intent targets. In order to better understand the value of a variety of features, we considered three different feature combinations for these models: position, heuristic, and all features. The position features included only the position of the person over time, with no information about the robot. The heuristic features were those that were critical to create the heuristic labeling procedure for the Shutter-Lobby datasets. These features were the person's position, orientation, and velocity. Finally, the set with all the features considered all the pedestrian features available in the ATC and Shutter datasets, respectively, as explained in Sec. 3.

5.2 Encoder-Decoder Models

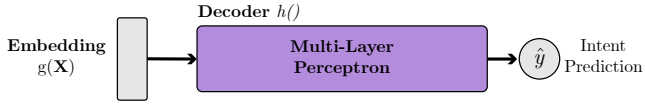
We considered two encoder-decoder architectures to predict interaction intent. This type of architecture has gained popularity for sequence-to-sequence prediction [30, 47, 48, 63] and data compression [31, 41, 64] with deep learning [19].

In general, the models' structure can be described as a composition of two functions: $f = h \circ g$, where g is an encoder function that computes a feature embedding for the input features $\mathbf{X}$ and h is a decoder function that transforms the output of g into an intent prediction. Figure 3a illustrates this encoder-decoder structure. The encoder function in all the models described below corresponds to a Gated Recurrent Unit (GRU) neural network [15] because such recurrent models are well suited to summarize sequence data into a feature embedding [19]. We denote the encoder simply by $g(\mathbf{X})$, which outputs the last hidden state $\mathbf{s}^{(T)}$.

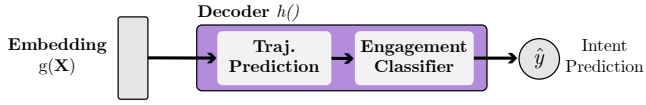
We consider two decoder architectures in our work. The first model is an *MLP-based model* (GRU-MLP), which has a multi-layer perceptron (MLP) as the decoder (Figure 3b). This model was motivated by the effectiveness of neural networks as general function approximators [34, 65], the simplicity of multi-layer perceptrons, and their popularity in machine learning. The GRU-MLP model



(a) Encoder-decoder structure for intent prediction models.



(b) Decoder with multi-layer perceptron.



(c) Decoder with recurrent trajectory model and engagement classifier.

Figure 3: Structure of encoder-decoder models. As in (a), the models take pedestrian features X as input, compute an embedding with an encoder $g()$, and then predict human intent to engage with a robot with a decoder $h()$.

predicts interaction intent directly, mapping the past behavior of a pedestrian to their intent, similar to the RF models.

The second model is a *trajectory-based model* (GRU-Traj), as in Figure 3c. This model uses a decoder composed of a recurrent trajectory forecasting model (implemented as a neural network) and a current-engagement classifier (implemented as a Random Forest). GRU-Traj was motivated by work in trajectory prediction in computer vision [52] and the success of recurrent encoder-decoder architectures for this task [2, 22, 24]. Intuitively, a person's motion trajectory can tell us about their interaction intent. If we could forecast their motion, we could infer their intent by reasoning about their future spatial behavior (e.g., as in [43]).

Neural network models were implemented using PyTorch and the RF models were implemented using Scikit-Learn. A detailed description of the structure, training, and best hyperparameters for these models is provided in the Appendix to facilitate reproducibility efforts and benchmarking with PAR-D.

5.3 Evaluation Setup

We focused our evaluation of the computational models on two main research questions:

RQ1) *How well can the different models predict intent in PAR-D?*
RQ2) *How does the performance of the machine learning models compare with the human baselines?*

For the first question in particular, we considered the RF models that used different features for intent prediction as different models, although they were of the same classifier type.

We created two types of test sets for each of the PAR-D datasets to investigate the research questions. One type considered all possible examples in unseen partitions of the data. The other type considered the same examples used to compute the human baselines in Sec. 4. From now on we refer to these two test sets as "Test" and "Annotated", respectively.

For the ATC dataset, we used 5-fold cross-validation to create five different partitions of the data, as described in Section 4.2. For each fold, 20% of the data was used for testing models and the rest was used for training and validation. We aggregate results by averaging performance over the test folds, which from now on we refer to as the "Test" data for ATC. For the human baselines, we gathered annotations on a subset of the data of two folds. We evaluated performance on each of these folds separately. We further refer to them as the "Easy Fold" and the "Hard Fold" for ATC.

For each of the Shutter datasets, we used the examples that were not considered for the human-performance baseline for training (90% of the data) and validation (10%).

Because interaction with the robot is relatively rare, the PAR-D datasets are highly imbalanced. Only 2% of training examples are positive in the ATC Approach Trajectory dataset, 3.7% in the Shutter-Lobby1 dataset, and 5.7% in the Shutter-Lobby2 dataset. Thus, we evaluate the trained models based on F1 Score, which captures the trade-off between precision and recall, and is not biased by the majority class like the Accuracy metric.

5.4 Results

5.4.1 ATC. Figure 4 shows the F1 scores for the models trained on the ATC Approach Trajectory [28] dataset. Detailed results for all

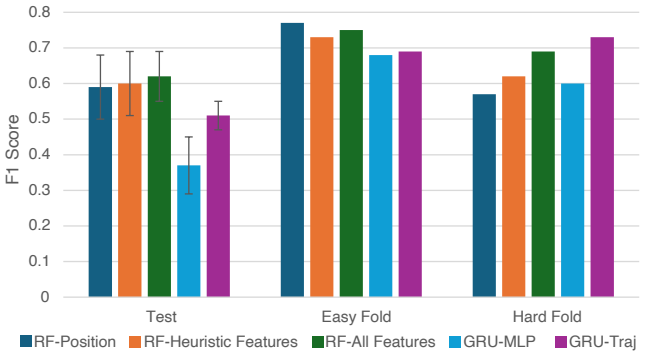


Figure 4: F1 scores for machine learning models on the ATC test data. The "Easy Fold" and "Hard Fold" sets are the same used for our human baselines. Best viewed in color.

models are in the Appendix in the supplementary material. The RF model using all features performed best in terms of average F1 score on Test, reaching a score of 0.62. Slightly lower performance was obtained with the RF with heuristic and position features, which had F1 scores of 0.6 and 0.59, respectively. The GRU-Traj model followed with an F1 score of 0.51. The GRU-MLP was the worst-performing model on Test (0.37 F1 score).

Interestingly, the RF model with only position features was highest performing model type for the "Easy Fold" and the lowest performing on the "Hard Fold", reaching F1 scores of 0.77 and 0.59, respectively. By contrast, the GRU-Traj model was second-lowest on the "Easy Fold" with an F1 score of 0.69. However, it had the highest score on the "Hard Fold", with an F1 score of 0.73.

5.4.2 Shutter-Lobby1 Dataset. F1 scores for the models trained on the Shutter-Lobby1 dataset are shown in Figure 5 (top). For Shutter-Lobby1, the GRU-Traj model performed best on both the "Test" and "Annotated" sets, achieving an F1 score of 0.18 and 0.67, respectively. Meanwhile, the RF model with all features performed worst on "Test" (F1 score of 0.06) and scored second-highest on "Annotated" (F1 score of 0.5). Though the GRU-MLP was second-highest on "Test", it did the worst on the "Annotated" data by a wide margin, reaching an F1 score of 0.14 and 0.04, respectively.

5.4.3 Shutter-Lobby2 Dataset. F1 scores for models trained on the Shutter-Lobby2 dataset are shown in Figure 5 (bottom). The GRU-MLP and GRU-Traj models performed best on both the "Test" and "Annotated" sets. The GRU-MLP model reached an F1 score of 0.32 on Test and 0.66 on the Annotated whereas the GRU-Traj model had an F1 score of 0.36 on "Test" and 0.62 on the "Annotated". There was no consistent ordering for the performance of the RF models between the "Test" and "Annotated" sets.

5.5 Discussion

5.5.1 RQ1: Prediction Performance. The results on the ATC and Shutter-Lobby datasets suggest that the GRU-Traj models can facilitate predicting interaction intent on more nuanced data than other model types. In the ATC dataset, the trajectory models have middling performance on the "Test" and "Easy Fold", but outperform all other model types on "Hard Fold". In the Shutter-Lobby datasets,

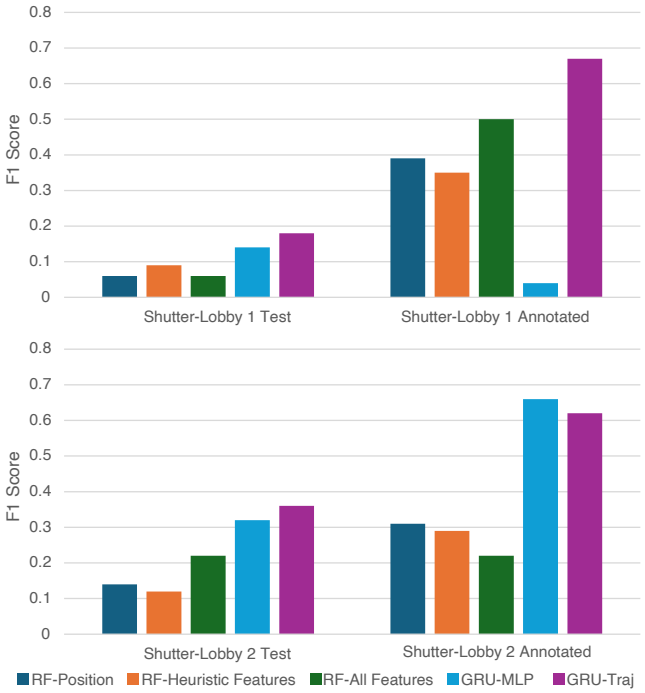


Figure 5: F1 scores for machine learning models on Test data from Shutter-Lobby1 and Shutter-Lobby2. The "Annotated" sets are the same used for our human baselines.

the GRU-Traj model performs well, outperforming the RF models by a wide margin. These findings support the idea that separating behavior forecasting from behavior interpretation can be beneficial in this problem domain.

Although the RF models are performant on the ATC dataset, they cannot model pedestrian behavior on the Shutter-Lobby datasets as well as the encoder-decoder models often can. The data representation of the Shutter-Lobby datasets does not seem to be the cause for this difference in performance. Making predictions on the Shutter-Lobby data with position or heuristic features only did not always improve the RF model's performance. Rather, we suspect that the RF models had more difficulty than the encoder-decoder models in hard test sets because of the nuanced of the data and the input feature space, which the neural networks could more easily transform into other useful representations than the RF models.

Finally, we observed that the performance of the GRU-MLP model on the Shutter-Lobby1 Annotated dataset was surprisingly low. Further inspection of the model's predictions indicated that the model was predicting mostly negative intent to interact on this specific dataset. We suspect this happened because of a distribution shift between the training and validation data in the Shutter-Lobby 1 dataset (which were highly imbalanced) and the annotated samples (which had the same number of positive and negative examples). An exploratory, additional analysis that supports this idea is presented in the last section of the Appendix.

5.5.2 RQ2: Human and Models Comparison. Performance in the Shutter-Lobby2 dataset was similar between computational models and humans (F1 score of 0.66 vs. F1 score of 0.65, respectively). However, for the ATC and Shutter-Lobby1 datasets, human annotators outperformed the computational models. This may be due to a major difference between how the humans and models saw the data: models never saw any scene context while humans could see all pedestrians at once (similar to the supplementary video but without intent labels). Perhaps having more knowledge of the social context of the interaction could have helped computational models.

6 LIMITATIONS & FUTURE WORK

Our work has limitations that point to interesting future research directions. First, we considered two definitions for interaction in this work. One was from the ATC dataset. The other one was a heuristic-based definition, which was inspired by [9] and aimed to facilitate scalability of the labeling process with the Shutter data. However, there are alternative ways to define interaction with a robot in public environments. For example, an interaction could be defined based on whether a person is paying continuous attention to the robot, even if they have not started to explicitly interact with each other. Future work could extend our PAR database with labels for alternative definitions of interactions.

Second, we only evaluated a limited set of computational models in this work. For example, we explored using a trajectory forecasting model in conjunction with a current-engagement classifier, which were trained independently in this work. But future work could explore training such models jointly by making the engagement classifier a neural network as well. Further, other models could use transformer architectures, which provide state-of-the-art results in trajectory prediction (e.g., [39, 40, 54, 66–68]).

Another important limitation of our current work is that our models do not consider the layout of the environment where people encounter the robot. As described by Bohus and Horvitz [9], the spatial constraints of an environment can have significant impact on how people approach a robot. In future work, we plan to address this shortcoming by incorporating information about the environment into intent prediction models, e.g., using occupancy maps as in [61].

7 CONCLUSION

In this work, we first introduced the PAR Database for researching pedestrian behavior and intent to interact with robots in public spaces. Then, we contributed human baselines for predicting intent and discussed several factors that can affect prediction performance, including spatial constraints in the robot's environment and the length of time for which a person is observed. Further, we conducted an initial evaluation of computational models for intent prediction on the PAR-database. We found that forecasting a person's motion as an intermediary step to infer intent can help make correct intent predictions, especially with more complex data. We hope that the HRI community takes advantage of our open-source contributions and the insights from this work to build more robust intent prediction models for HRI in the future.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (NSF), Grant No. (IIS-2143109). The findings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Ahmed A Abdelrahman, Dominykas Strazdas, Aly Khalifa, Jan Hintz, Thorsten Hempel, and Ayoub Al-Hamadi. 2022. Multimodal Engagement Prediction in Multiperson Human–Robot Interaction. *IEEE Access* 10 (2022), 61980–61991.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.

[3] Florian Alt, Daniel Buschek, David Heuss, and Jörg Müller. 2021. Orbuculum-Predicting When Users Intend to Leave Large Public Displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–16.

[4] Elisabeth André, Ana Paiva, Julie Shah, and Selma Šabanovic. 2020. Social agents for teamwork and group interactions (dagstuhl seminar 19411). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[5] Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7, 4 (2015), 465–478.

[6] Atef Ben-Youssef, Chloé Clavel, and Slim Essid. 2019. Early detection of user engagement breakdown in spontaneous human-humanoid interaction. *IEEE Transactions on Affective Computing* 12, 3 (2019), 776–787.

[7] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[8] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-Fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics* 11, 5 (2019), 815–828.

[9] Dan Bohus and Eric Horvitz. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference*. 244–252.

[10] Dan Bohus and Eric Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference, The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 10.

[11] Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. 2–9.

[12] Dan Bohus, Chit W Saw, and Eric Horvitz. 2014. Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. Citeseer, 637–644.

[13] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. 2012. Detecting engagement in HRI: An exploration of social and task-based context. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 421–428.

[14] Zhichao Chen, Yutaka Nakamura, and Hiroshi Ishiguro. 2022. Android as a receptionist in a shopping mall using inverse reinforcement learning. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7091–7098.

[15] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[16] Enrique Coronado, Xela Indurkhya, and Gentiane Venture. 2019. Robots Meet Children, Development of Semi-Autonomous Control Systems for Children-Robot Interaction in the Wild. In *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*. 360–365. https://doi.org/10.1109/ICARM.2019.8833751

[17] Lee J Corrigan, Christopher Peters, Ginevra Castellano, Fotis Papadopoulos, Aidan Jones, Shweta Bhargava, Srini Janarthanam, Helen Hastie, Amol Deshmukh, and Ruth Aylett. 2013. Social-task engagement: striking a balance between the robot and the task. In *Embodied Commun. Goals Intentions Workshop ICSR*, Vol. 13. 1–7.

[18] Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. 2017. Automatically classifying user engagement for dynamic multi-party human–robot interaction. *International Journal of Social Robotics* 9, 5 (2017), 659–674.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[20] Edmund T Hall and Edward T Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.

[21] Brandon Heenan, Saul Greenberg, Setareh Aghel-Manesh, and Ehud Sharlin. 2014. Designing social greetings in human robot interaction. In *Proceedings of the 2014 conference on Designing interactive systems*. 855–864.

[22] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6272–6281.

[23] Koichiro Ito, Quan Kong, Shota Horiguchi, Takashi Sumiyoshi, and Kenji Nagamatsu. 2020. Anticipating the start of user interaction for service robot in the wild. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9687–9693.

[24] Boris Ivanovic and Marco Pavone. 2019. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2375–2384.

[25] Malte Jung and Pamela Hinds. 2018. Robots in the wild: A time for more robust theories of human-robot interaction. , 5 pages.

[26] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2010. A communication robot in a shopping mall. *IEEE Transactions on Robotics* 26, 5 (2010), 897–913.

[27] Daphne Karreman, Geke Ludden, and Vanessa Evers. 2015. Visiting cultural heritage with a tour guide robot: A user evaluation study in-the-wild. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*. Springer, 317–326.

[28] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May i help you?-design of human-like polite approaching behavior. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 35–42.

[29] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[30] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[31] Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.

[32] Melissa Kont and Maryam Alimardani. 2020. Engagement and mind perception within human-robot interaction: A comparison between elderly and young adults. In *International Conference on Social Robotics*. Springer, 344–356.

[33] Seongyong Koo and Dong-Soo Kwon. 2009. Recognizing human intentional actions from the relative movements between human and robot. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 939–944.

[34] Věra Kůrková. 1992. Kolmogorov's theorem and multilayer neural networks. *Neural networks* 5, 3 (1992), 501–506.

[35] Min Kyung Lee and Maxim Makatchev. 2009. How do people talk with a robot? An analysis of human-robot dialogues in the real world. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 3769–3774.

[36] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. 2015. Comparing models of disengagement in individual and group interactions. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 99–105.

[37] Alexander Lew, Sydney Thompson, Nathan Tsoi, and Marynel Vázquez. 2023. Shutter, the Robot Photographer: Leveraging Behavior Trees for Public, In-the-Wild Human-Robot Interactions. *arXiv preprint arXiv:2302.00191* (2023).

[38] Kang Li, Shiying Sun, Xiaoguang Zhao, Jinting Wu, and Min Tan. 2019. Inferring user intent to interact with a public service robot using bimodal information analysis. *Advanced Robotics* 33, 7-8 (2019), 369–387.

[39] Lihuan Li, Maurice Pagnucco, and Yang Song. 2022. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2231–2241.

[40] Yao Liu, Lina Yao, Binghao Li, Xianzhi Wang, and Claude Sammut. 2022. Social Graph Transformer Networks for Pedestrian Trajectory Prediction in Complex Social Scenarios. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1339–1349.

[41] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 64–73.

[42] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. 2023. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–39.

[43] Marek P Michalowski, Selma Sabanovic, and Reid Simmons. 2006. A spatial model of engagement for a social robot. In *9th IEEE International Workshop on Advanced Motion Control, 2006.* IEEE, 762–767.

[44] Jauwairia Nasir, Barbara Bruno, Mohamed Chetouani, and Pierre Dillenbourg. 2022. What if Social Robots Look for Productive Engagement? *International Journal of Social Robotics* 14, 1 (2022), 55–71.

[45] Stefanos Nikolaidis, Anca Dragan, and Siddhartha Srinivasa. 2016. Viewpoint-based legibility optimization. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 271–278.

[46] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92.

[47] Seong Hyeon Park, ByeongDo Kim, Chang Mook Kang, Chung Choo Chung, and Jun Won Choi. 2018. Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1672–1678.

[48] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A Comparison of sequence-to-sequence models for speech recognition.. In *Interspeech*. 939–943.

[49] Raúl Quintero, Ignacio Parra, David Fernández Llorca, and MA Sotelo. 2015. Pedestrian intention and pose prediction through dynamical models and behaviour classification. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 83–88.

[50] Ely Repiso, Anaís Garrell, and Alberto Sanfeliu. 2020. People's adaptive side-by-side model evolved to accompany groups of people by social robots. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2387–2394.

[51] Eduardo Rodriguez-Lizundia, Samuel Marcos, Eduardo Zalama, Jaime Gómez-García-Bermejo, and Alfonso Gordaliza. 2015. A bellboy robot: Study of the effects of robot behaviour on user engagement and comfort. *International Journal of Human-Computer Studies* 82 (2015), 83–95.

[52] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research* 39, 8 (2020), 895–935.

[53] Michaela Schimpl, Carmel Moore, Christian Lederer, Anneke Neuhaus, Jennifer Sambrook, John Danesh, Willem Ouwehand, and Martin Daumer. 2011. Association between walking speed and age in healthy, free-living individuals using mobile accelerometry—a cross-sectional study. *PloS one* 6, 8 (2011), e23299.

[54] Ze Sui, Yue Zhou, Xu Zhao, Ao Chen, and Yiyang Ni. 2021. Joint intention and trajectory prediction based on transformer. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7082–7088.

[55] Yuwei Sun and Daniel W Carruth. 2018. How Does Presence of a Human Operator Companion Influence People's Interaction with a Robot?. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 136–142.

[56] Mikael Svenstrup, Thomas Bak, Ouri Maler, Hans Jørgen Andersen, and Ole B Jensen. 2009. Pilot study of person robot interaction in a public transit space. In *Research and Education in Robotics—EUROBOT 2008: International Conference, Heidelberg, Germany, May 22-24, 2008. Revised Selected Papers*. Springer, 96–106.

[57] Sydney Thompson, Austin Narcomey, Alexander Lew, and Marynel Vázquez. 2024. Shutter: A Low-Cost and Flexible Social Robot Platform for In-the-Wild Deployments. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 94–96.

[58] Sebastian Thrun, Maren Bennewitz, Wolfram Burgard, Armin B Cremers, Frank Dellaert, Dieter Fox, Dirk Hahnel, Charles Rosenberg, Nicholas Roy, Jamieson Schulte, et al. 1999. MINERVA: A second-generation museum tour-guide robot. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, Vol. 3. IEEE.

[59] Bekir Berker Türker, Zana Buçinca, Engin Erzin, Yücel Yemez, and T Metin Sezgin. 2017. Analysis of Engagement and User Experience with a Laughter Responsive Social Robot.. In *Interspeech*. 844–848.

[60] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. 2016. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems* 75 (2016), 4–16.

[61] Marynel Vázquez, Alexander Lew, Eden Gorevoy, and Joe Connolly. 2022. Pose Generation for Social Robots in Conversational Group Formations. *Frontiers in Robotics and AI* (2022), 341.

[62] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. 2016. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 36–43.

[63] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.

[64] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, 12 (2010).

[65] Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2021. How neural networks extrapolate: from feedforward to graph neural networks. In *International Conference on Learning Representations (ICLR)*.

[66] Hai-Yan Yao, Wang-Gen Wan, and Xiang Li. 2022. End-to-end pedestrian trajectory forecasting with transformer network. *ISPRS International Journal of Geo-Information* 11, 1 (2022), 44.

[67] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*. Springer, 507–523.

[68] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9813–9823.

[69] Zhijie Zhang, Jianmin Zheng, and Nadia Magnenat Thalmann. 2021. Engagement Intention Estimation in Multiparty Human-Robot Interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 117–122.

Received 10 May 2024; accepted 18 July 2024