2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

EWareNet: Emotion-Aware Pedestrian Intent Prediction and Adaptive Spatial Profile Fusion for Social Robot Navigation

Venkatraman Narayanan, Bala Murali Manoghar, Rama Prashanth RV and Aniket Bera
Department of Computer Science, Purdue University, USA

***Abstract*— We present *EWareNet*, a novel intent and affect-aware social robot navigation algorithm among pedestrians. Our approach predicts the trajectory-based pedestrian intent from gait sequence, which is then used for intent-guided navigation taking into account social and proxemic constraints. We propose a transformer-based model that works on commodity RGB-D cameras mounted onto a moving robot. Our intent prediction routine is integrated into a mapless navigation scheme and makes no assumptions about the environment of pedestrian motion. Our navigation scheme consists of a novel obstacle profile representation methodology that is dynamically adjusted based on the pedestrian pose, intent, and affect. The navigation scheme is based on a reinforcement learning algorithm that takes pedestrian intent and robot's impact on pedestrian intent into consideration, in addition to the environmental configuration. We outperform current state-of-art algorithms for intent prediction from 3D gaits.**

I. Introduction

Recent technological advancements are making human-robot collaborations increasingly important. This can have many applications, including autonomous driving, social robotics, and surveillance systems. As humans and robots co-inhabit space, designing robots that follow collision-free paths and are socially acceptable to humans is becoming increasingly important, creating several challenges. For example, in the case of a densely crowded street, the robot needs to foresee the movements of an oblivious walker who is unaware they are in its path for friendlier navigation. Understanding the emotional perceptions in such scenarios enables the robot to make more informed decisions and navigate in a socially-conscious manner.

The study of human emotions has been a much-researched subject in areas like psychology, human-robot collaboration, interaction, etc. Certain studies have sought to identify the emotion in people based on cues such as body movement (walking style, etc.) [1], [2], and verbal cues (speech tonalities and patterns) [3], [4]. There are also multimodal approaches that utilize a combination of these cues to recognize the person's emotion [5], [6], [7]. Combining the information from the robot sensors and accounting for the uncertainty of the movements of oblivious walkers remains a significant challenge.

Several approaches have addressed a socially-acceptable robot navigation problem. Many recent works like [2], [8] focus on social robot navigation in crowded scenarios. However, these algorithms suffer from the following drawbacks:

1) They rely on emotionally reactive navigation planning, which means that the perceived emotional/affective states are based entirely on historical gait sequences and do not consider gait predictions of the future.
2) The influence of the robot's behavior on the emotional response of pedestrians is not taken into account.

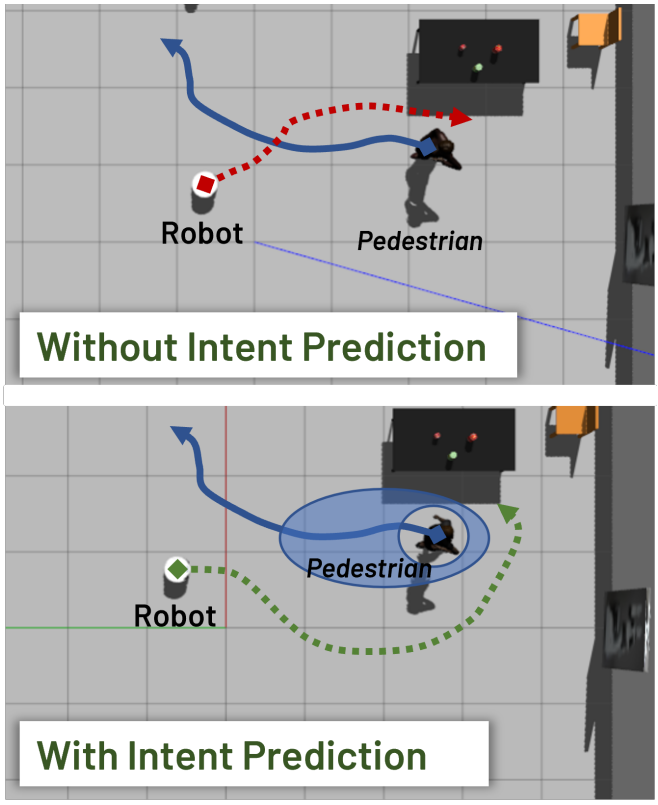


Fig. 1. ***EwareNet**: The top figure represents the path taken by the robot without intent prediction, where the trajectory intercepts the path to be taken by the pedestrian. The bottom figure shows the path taken by the robot with the intent prediction pipeline where even though there is enough space and a shorter route is available, the robot takes a longer path giving enough personal space.*

3) The proxemic constraints for pedestrian motion are heuristically computed from prior psychology studies. While this works in most cases, it does not efficiently capture each individual's uniqueness in comfort space, personal space, culture, etc.

To overcome these challenges, we propose *EWareNet*, a novel algorithm for a real-time, gait-based, intent-aware robot navigation algorithm in social scenes that takes into account the pedestrian's affect state, as depicted in figure 1. *EWareNet* is designed to work with commercial off-the-shelf RGB-D cameras and depth sensors that can be retrofitted to moving platforms or robots for navigation. The **major contributions** of our work can be summarized as follows:

- We introduce a novel approach using *transformers* to predict the full-body pedestrian intent or behavior in the form of trajectory/gaits based on historical gait sequences.

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10161504

- We also present a novel navigation planning algorithm based on deep reinforcement learning that takes into consideration the environment, pedestrian intent, and the robot's reactionary impact on pedestrian behavior.
- Our method explicitly considers pedestrian behavior in crowds and the robot's impact on the pedestrians and environment.
- Finally, we introduce an adaptive spatial density function to represent the proximal constraints for pedestrians that captures pedestrians' unique personal comfort space in terms of pose, intent, and emotion.

Pedestrian intent and affect (the scientific term for emotions) are very subjective and greatly influenced by many environmental and psychological factors. Therefore in this work, we focus only on the *perceived* emotions from the point of an external observer as opposed to *true* internal emotion. We also predict pedestrian intent as a potential full-body future trajectory based upon historical trajectories/gait patterns for a more efficient navigation strategy and may not represent actual/hidden intent. Additionally, we are only computing "walking emotion" as opposed to other forms of emotion [9].

The paper is organized as follows: Section II presents related work, section III gives an overview of our pipeline and describes each stage of our pipeline in detail, and finally, in IV we evaluate the results of our work.

II. Related Work

In this section, we present a brief overview of social-robot navigation algorithms. We also review related work on emotion modeling and classification from visual cues.

A. Social Robotics and Emotionally-Guided Navigation

A substantial amount of research focuses on identifying pedestrians' emotions based on body posture, movement, and other non-verbal cues. Ruiz-Garcia et al. [10] and Tarnowski et al. [11] use deep learning to classify different emotion categories from facial expressions. [7] use multiple modalities such as facial cues, pedestrian pose, and scene understanding. Randhavane et al. [12], [13] classify emotions into four classes based on affective features obtained from 3D skeletal poses extracted from pedestrian gait cycles. Their algorithm, however, requires a large number of 3D skeletal key points to detect emotions and is limited to single individual cases. Bera et al. [14], [15], [16], [17] classify emotions based on facial and body expressions along with a pedestrian trajectory obtained from overhead cameras. Although this technique accurately predicts emotions from trajectories and facial expressions, it explicitly requires overhead cameras in its pipeline. [2] provides an end-to-end deep learning-based emotion classification approach that takes in skeletal gaits from an arbitrary view. [18] modeled the navigation of a social robot based on human-robot trust interactions.

B. Intention Prediction

In social robotics, an accurate prediction of pedestrian trajectories plays a significant role in robot decisions in terms of reactive response, navigation planning, etc. For simplicity and ease of computation for navigational robots, the problem of pedestrian intention prediction is largely modeled as a trajectory prediction problem. Recurrent Neural Network based architectures are widely used for predicting pedestrian trajectories [19], [20], [21]. [22] use a Graph Convolution method to predict the trajectories and showed that Temporal Graph Convolutions are much better at predicting pedestrian trajectories compared to recurrent networks.

Most recurrent networks for trajectory prediction leverages some form of attention mechanism to improve the trajectories. [21], [19] uses a history of observed trajectories with either predicted future trajectory or location around pedestrians to predict the intent. [20] provide skeletal joint kinematics as attention for trajectory prediction. These models, one way or the other, solely depend on the trajectory and kinematics of joints for predicting the intent, and there is no attention given to the emotional state of the pedestrians.

C. Proxemic Constraints Modeling

Usually, robots working in pedestrian environments have used navigation algorithms where all obstacles are considered of similar relevance, including people. [14] a have studied the effects of comfort space of pedestrians from a psychological perspective and showed that comfort space varies based on the emotion of pedestrians. To avoid discomforting pedestrians, social robots must consider unique entities, evaluating the people's level of comfort with respect to the route of the robot. The navigation model of [14], [2] uses the predicted emotions and uses a constant multiplier to maintain proper comfort space with the pedestrians while planning the robot path.

Vega et al. [8] proposed a pedestrian-aware navigation strategy based on space affordances. The method is built upon using an adaptive spatial density function that efficiently clusters groups of people according to their spatial arrangement. The paper [23] discusses how an agent should learn the behavior from a reward provided by a live human trainer rather than the usual pre-coded reward function in a reinforcement learning framework. With the designed CNN-RNN model, our analysis shows that telling trainers to use facial expressions and competition can improve accuracy for estimating positive and negative feedback using facial expressions.

All these methods consider emotion as a constant metric for formulating proxemic constraints and do not adapt the navigation strategy to the change in the emotion of pedestrians.

III. Overview and Methodology

EWareNet is a novel algorithm for intent-aware and socially-acceptable navigation through crowded scenarios. The pedestrian intentions are predicted based on 3D skeletal trajectories from an onboard RGB-D camera. Our navigation algorithm then uses these trajectories for socially-acceptable navigation. Our navigation planning is adaptive to the individual comfort space constraints of pedestrians and to the uncertainty in the sensor suite of the robot. Based on a Gaussian distribution, we use dynamic obstacle representation for pedestrians to capture individuality in comfort space constraints.

The following subsections will describe our approach in detail. We discuss the details of the datasets used for training our algorithm, along with the pre-processing techniques employed. Following this, we focus on our intent prediction

routine, where we also discuss our pedestrian pose extraction from an RGB-D camera briefly. Finally, we discuss our navigation algorithm.

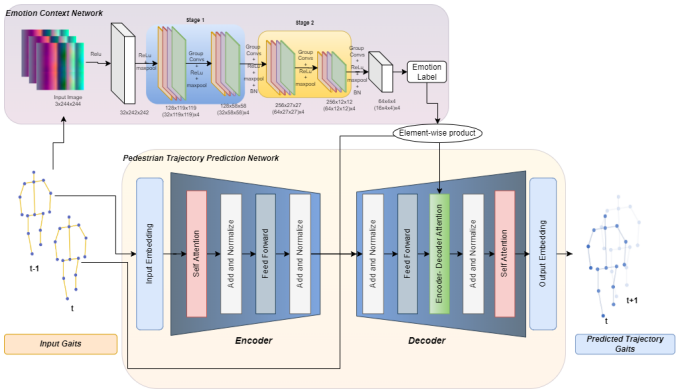


Fig. 2. ***Pedestrian Intent Prediction****: Our proposed method for predicting pedestrian intentions as full-body skeletal trajectory paths. It consists of a two-part network -* ***Emotion Context Network*** *and* ***Pedestrian Trajectory Prediction Network****. The Emotion Context Network provides an additional* ***attention mechanism*** *on top of the self-attention mechanism in transformer networks.*

A. Pedestrian Pose Extraction

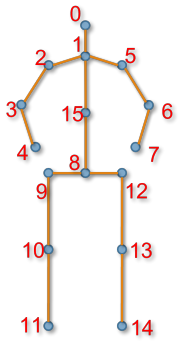
The primary objective of our work involves the effective use of pedestrian temporal skeletal poses for more harmonious navigation of a robot in crowded situations. We rely on the system, described in [24], to extract the poses from pedestrians walking in a crowded real-world scenario from an RGB-D camera. The system consists of a two-step network trained in a weakly supervised fashion. A *Structure-Aware PoseNet (SAP-Net)* provides an initial pose estimate based on spatial information of joint locations of people in the video. Later, a time-series *Temporal PoseNet (TP-Net)* corrects the initial estimate by adjusting impermissible joint angles and joint distances (based on human physiology and geometry restrictions). The temporal network also helps in tracking the pedestrian individual across the video frames. Hence, we have a set of temporally correct, *Pedestrian Joint Pose Sequence* (henceforth referred to as *gaits*) for every pedestrian in the video as an output from the system.

Similar to [2], our *Pedestrian Pose Extraction* network extracts a representation for the pedestrians as 16 skeletal joints as shown in figure 3. Every pose, $P \in \mathbb{R}^{16\times3}$ of a pedestrian consists a set of 3D positions of each joint j_i, where $i \in \{0, 1, ..., 15\}$. For any RGB-D video V, we represent the gait extracted using 3D pose estimation as G. The gait G_i, t is a set of 3D poses for pedestrian i over $P_1, P_2, ..., P_t$ where t is the frame number of the input video V.

B. Pedestrian Intent Prediction

Our Pedestrian Intent Prediction module is designed as a trajectory prediction system based on a set of skeletal gait sequences. The intent prediction system consists of three parts, *(i) an Emotion Context Network (ECN), (ii) Pedestrian Trajectory Forecasting Network*. Figure III provides a schematic representation of our Pedestrian Intent Prediction algorithm.

Joint Number	Joint Name
0	Head
1	Neck
2	Right Shoulder
3	Right Elbow
4	Right Wrist
5	Left Shoulder
6	Left Elbow
7	Left Wrist



Joint Number	Joint Name
8	Pelvis
9	Right Hip
10	Right Knee
11	Right Foot
12	Left Hip
13	Left Knee
14	Left Foot
15	Back

Fig. 3. ***Skeleton Representation****: We represent a pedestrian by 16 joints (j_i). The overall pose of the pedestrian is defined using these joint positions.*

1) Emotion Context Network: Our emotion context network extends on [2]. It is a multi-view emotion classification network. The architecture takes in the gait sequences consisting of 16 joints (depicted in figure III) for 75 time steps to classify them into one of 4 emotions *(Happy, Sad, Angry, Neutral)*. The gait sequences are processed into image embeddings to leverage faster processing abilities of *Convolution Neural Networks* [2]. The embedded images are fed to the classification network. The classification network *(ECN)* consists of multiple layers of grouped convolutions—the grouped convolutions aid in classifying the emotions for different camera angle view groups. The extracted emotion is fed as an attention mechanism to our *Transformer Network* to obtain intent-aware trajectory predictions.

$$e_i = ECN(I_i) \tag{1}$$

Each gait sequences *(i)* is assigned an emotion label, (e_i), based on equation 1, where ECN represents the emotion classification network, I_i, represents the image embedding of the gait sequences i

2) Pedestrian Trajectory Forecasting Network: The Pedestrian Trajectory Forecasting network predicts the intent-aware future full-body gait trajectories for every pedestrian based on past gaits. Our Pedestrian Trajectory Forecasting network is inspired by transformer networks depicted in [25]. Transformer networks have successfully produced state-of-the-art results involving time-series sequences. Wen et al. [26] conducted extensive analysis, which concluded that Transformers are superior in modeling long-range dependencies, training can be parallelized, and does not suffer from vanishing gradients when compared to the recurrent counterpart approaches. They have been extensively used in NLP tasks [27], [28] and time-series forecasting [29]. Guiliari et al. [25] uses transformer networks to generate state-of-the-art results on pedestrian trajectory prediction.

$$e_{i,t} = G_{i,t}{}^T.W_g \tag{2}$$

The input gait sequences for each pedestrian, $G_{i,t}$, are fed to a linear encoding layer given by equation 2. The embedded gaits inputs are provided to the transformer network. Transformer (TF) network is a modular architecture (figure III). The transformer network follows an encoder-decoder style network, each composed of 6 layers, with three building blocks each: (i) a self-attention module, (ii) a feed-forward fully-connected module, and (iii) two residual connections after each of the previous blocks. The self-attention modules

within the TF net provide the capability to capture non-linearities in the trajectory.

The encoder model creates a representation, ${E_s}^{(i,t)}$, of the input gait sequences, forming a memory module. The encoder representation is utilized by the decoder model to auto-regressively predicts future trajectories.

The *encoder-decoder attention* layer of the decoder network is fed with encoder representation and emotion context from emotion context network in III-B. The emotion probabilities, e_i, from the *ECN* are multiplied element-wise with the encoder outputs to be fed as attention to the decoder. For simplicity, we assume that the emotion context remains constant for all the timesteps predicted by the decoder. The intent-aware trajectory predictions are used in our navigation algorithm.

C. Proxemic Contraints Modeling

The personal space constraints are derived from prior works in psychology [30]. Personal space constraints are essential in social robotics and emotion-guided navigation [31]. The comfort space constraints defined in [32], [30] describe the social norms that the robot must consider not to cause discomfort to people around it. The comfort space constraints related in [30], [32], [2] do not consider the individuality of pedestrians. On the other hand, our approach is able to capture the individuality of the comfort space for pedestrians.

Our pedestrian personal space modeling is an adaptive spatial density function, similar to [8]. In an arbitrary global map in space, a pedestrian, *i*, is represented by position, orientation, and emotion (from III-B.1), $h_i = (x_i, y_i, \theta_i, C_i)^T$. Here (x_i, y_i) represents the position of the pedestrian, θ_i represents the orientation, and C_i represents the comfort space derived from emotion. The position coordinates from each pedestrian will be derived from the LiDAR coordinates after mapping detected pedestrians from the camera frame to the LiDAR frame using projective geometry. The orientation and emotion are derived from section III-B.1. The view-group angle predicted by the model is used as the approximated orientation. The gaussian expression for personal space is defined by equation 3.

$$\begin{aligned} g_{h_i} = C_i * \exp(-k_1(x - x_i)^2 \\ + k_2(x - x_i)(y - y_i) \\ + k_3(y - y_i)^2) \end{aligned} \tag{3}$$

Here, C_i is the comfort space multiplier given by,

$$C_i = \frac{\sum_{j=1}^{4} c_j \cdot \max(e_j)}{\sum_{j=1}^{4} e_j} \cdot v_g \tag{4}$$

where e_j represents a column vector of the emotion context output from section III-B.1, which corresponds to the group outcomes for each individual emotion. c_j is an individual's comfort space constant derived from psychological experiments described in [30], chosen from a set $\{90.04, 112.71, 99.75, 92.03\}$ corresponding to the comfort spaces (radius in cm) for {*happy, sad, angry, neutral*} respectively. These distances are based on how comfortable pedestrians are while interacting with others. v_g is a view-group constant chosen from a set of $\{1, 0.5, 0, 0.5\}$ based on the view group g, defined in [2].

Also, k_1, k_2, k_3 in equation 3 are the coefficients taking into account the orientation $\theta_i \in [0, 2\pi)$, defined in equations 5, 6, 7.

$$k_1(\theta_i) = \frac{cos({\theta_i}^2)}{2\sigma^2} + \frac{sin({\theta_i}^2)}{2\sigma_s^2} \tag{5}$$

$$k_2(\theta_i) = \frac{sin(2\theta_i)}{4\sigma^2} \frac{sin(2\theta_i)}{4\sigma_s^2} \tag{6}$$

$$k_1(\theta_i) = \frac{sin({\theta_i}^2)}{2\sigma^2} + \frac{cos({\theta_i}^2)}{2\sigma_s^2} \tag{7}$$

where σ_s is the variance to the sides of the pedestrian ($\theta_i \pm \frac{\pi}{2}$) and σ is the variance in the direction of the pedestrian orientation.

We understand that personal space distances depend on many factors, including cultural differences, the environment, or a pedestrian's personality. Hence we have modeled the personal space distances as an adaptive Gaussian function that is dynamically updated by our navigation algorithm (detailed in section III-D).

D. Intent-Aware Navigation

Our navigation framework is defined as a policy network trained to reach the goal while adjusting to the obstacles and personal space constraints of the pedestrians. Three consecutive LiDAR scans and tracked human poses are processed to accommodate our adaptive proxemic constraints for each detected pedestrian. Our policy network is inspired by [33], [34]. Our policy network receives the processed LIDAR scans and the intent-aware trajectory predictions for each pedestrian and outputs probabilities over the action space considered for the navigation task.

State Space: As discussed earlier, we consider three consecutive liar frames and tracked human poses. Along with these two components, the relative goal position and the robot's current velocity components constitute the state space. Each LIDAR scan is represented as $l_i \in \mathbb{R}^{512}$ and each pose is represented as $h_j \in \mathbb{R}^{16}$. Thus the state space is given by equation 8

$$s_t = \{l_0, l_1, l_2, h_0, h_1, ..., h_i\} \tag{8}$$

Since we consider three consecutive LIDAR scans and 18 consecutive traced skeleton poses, the state space is represented as $s_t \in \mathbb{R}^{(3\times512)\times(18\times16)}$

Action Space: The action space is a continuous set of permissible velocities. The action of the robot includes translational and rotational velocity. To accommodate the robot kinematics, we set the bounds for translational velocity, $v \in [0.0, 1.0]$, and rotational velocity, $w \in [-1.0, 1.0]$. We sample actions at step t using equation 9.

$$a_t = \pi(l_0, l_1, l_2, h_0, h_1, ..., h_i) \tag{9}$$

where l_0, l_1, l_2 represent the three consecutive processed LiDAR scan frames and $h_0, h_1, ..h_t$ represent the historical and predicted pedestrian trajectories concatenated together.

Policy Network: The policy network has four sub-component networks $\{f_{traj}, f_{enc}, f_{prev}, f_{act}\}$ (depicted in figure 4), where f_{traj} depicts our Pedestrian Intent Prediction network to encode the pedestrian trajectories, f_{enc} depicts our network to encode LiDAR frames from the robot, and finally f_{act} represents the network that takes the above-mentioned encodings along with robot's previous state and

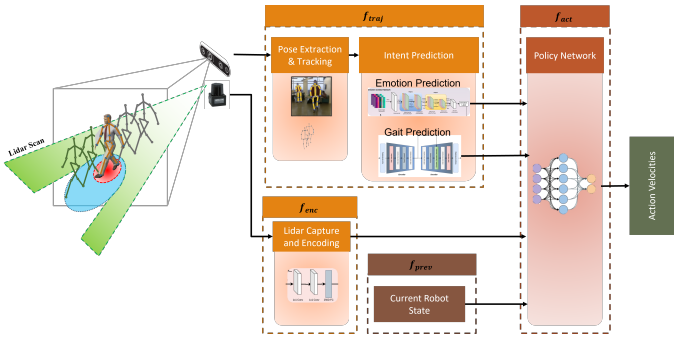


Fig. 4. *Our Navigation algorithm takes in the intent-aware pedestrian trajectories, LiDAR scans, and robot's positional information as input to a policy network. The policy network outputs the robot velocity vectors as action space.*

velocity to predict the action velocities to the robot. We embed the concatenated (historical & predicted) trajectories to skeletal gait to image embedding in [2], resize them to 244×244, and pass them to f_{traj}, which consists of four 5×5 Conv, BatchNorm, ReLU. Each Conv block is followed by a 2×2 MaxPool blocks, producing an encoded image of pedestrian trajectories $z_t^{img} = f_{traj}([h_0, h_1, h_i])$

The three consecutive lidar frames are passed through two 1×1 Conv, followed by a $256D$ fully-connected (FC) layer. The lidar frames are encoded as $z_t^{lidar} = f_{enc}([l_0, l_1, l_2])$. The f_{act} is a multi-layer perceptron (MLP) network, with 1 $128D$ FC hidden layer, and finally produces the action velocities. f_{act} takes in the encoded pedestrian trajectories, z_t^{img}, lidar encodings, z_t^{lidar}, previous velocity, v_{t-1}, goal position, s_g, and current robot position, s_t, to predict the robot velocities at time t, given by equation 10.

$$[h]v_t = f_{act}([z_t^{img}, z_t^{lidar}, v_{t-1}, s_g, s_t]) \tag{10}$$

v_t is then sent to a linear layer with softmax to derive the probability distribution over the action space, from which the action is sampled. We learn $\{f_{traj}, f_{enc}, f_{act}\}$ via reinforcement learning.

Rewards: Our reward function for the policy network extends [34]. Our goal is to find a good strategy to avoid collisions during navigation, minimize the arrival time and avoid any negative impact on the pedestrians while navigating, i.e., avoid pedestrian emotions progressing into $Angry, Sad$ while navigating (unless respective humans already display negative emotions. The reward function to achieve the mentioned goals is given in equation 11.

$$r^t = r_g^t + r_c^t + r_w^t + r_e^t \tag{11}$$

The reward r at time t is a combination of reward for reaching goal, r_g, reward for avoiding collisions, r_c, reward for smooth movement, r_w and reward for avoiding negative impact on pedestrians, r_e. The reward for reaching the intended goal/target is given by equation 12.

$$r_g^t = \begin{cases} r_{arrival}, & \text{if } ||p^t - g|| < \xi. \\ w_g(||p^{t-1} - g|| - ||p^t - g||), & \text{Otherwise.} \end{cases} \tag{12}$$

The penalty for colliding with obstacles is given by equation 13.

$$r_c^t = \begin{cases} r_{collision}, & \text{if robot collides.} \\ 0, & \text{Otherwise.} \end{cases} \tag{13}$$

For ensuring smooth navigation, the penalty for large rotational velocities is given by equation 14.

$$r_w^t = \begin{cases} w_w|w^t|, & \text{if } |w^t| > 0.7. \\ 0, & \text{Otherwise.} \end{cases} \tag{14}$$

To ensure the robot does not alarm any pedestrian, we penalize the robot for any change in detected pedestrian emotion into negative emotions, $Angry, Sad$, with equation 15. The penalty is only applied when the pedestrian emotion, e_i, changes from non-negative emotion to negative emotion. In other words, we do not penalize the robot when pedestrian emotion remains unchanged.

$$r_e^t = \begin{cases} r_e, & \text{if } e_0, .., e_i \in \{Angry, Sad\}. \\ 0, & \text{Otherwise.} \end{cases} \tag{15}$$

In our implementation, we use $r_{arrival} = 15, w_g = 2.5, r_{collision} = -15, w_w = -0.1, r_e = -2.5, \xi = 0.1$

IV. Experiments and Results

A. Pedestrian Intent Prediction

1) Metric: Since our pedestrian intent prediction is modeled as a trajectory prediction module, we evaluate the network with metrics commonly used for full-body trajectory/pose prediction networks. We use Mean Squared Error (MSE), which is the average l_2 distance between the ground-truth and the predicted poses at each time step for all pedestrians present in the frame.

2) Datasets: In this work, we leverage datasets designed for specific tasks of our approach.

Pedestrian Emotion Datasets: Our intent prediction system is reinforced with emotion labels extracted from the historical gaits. For this purpose, we use two labeled datasets by Randhavane et al. [35] and Bhattacharya et al. [36]. It contains 3D skeletal joints of 342 and 1835 pedestrian gait cycles each (2177 gait samples in total).

Pedestrian Intent Prediction Datasets: Our intent prediction system is modeled as a trajectory prediction system based on past trajectory/gait sequences reinforced with pedestrian emotions for the same. We use two social datasets, created from NTU RGB+D 60 [37] and PoseTrack [38] for training and evaluating our and demonstrate its superior performance against several relevant baselines. The NTU dataset contains both single-person actions and mutual actions. The PoseTrack is a large-scale multi-person dataset based on the MPII Multi-Person benchmark. The dataset covers a diverse variety of interactions, including person-person and person-object, in crowded, dynamic scenarios. For our experiments, we consider only the actions involving pedestrians *walking and/or running*. Hence we only select samples that fall under these categories.

3) Implementation Details: For training, our dataset (IV-A.2) has a train-validation split of 90%-10%. We perform training using an ADAM [39] optimizer, with decay parameters of ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The experiments were run with a learning rate of 0.009 and with 10% decay every 250 epochs. The models were trained with MSE loss, $\mathcal{L}$. The training was done on 2 Nvidia RTX 2080 Ti GPUs having 11GB of GPU memory each and 64 GB of RAM.

B. Intent-Aware Navigation

1) Metric: We evaluate our navigation algorithm using three metrics based on prior literature (i) Total distance traveled from source to goal, (ii) Time taken to reach the goal, (iii) Average personal space deviation (Δ_{ps}^{avg}), i.e., the difference between expected personal space and actual personal space provided by the robot while passing the pedestrian averaged across all pedestrian encounters.

2) Implementation Details: We rely on simulation environments to train our policy network. We use Stage Mobile Robot Simulator [40] to generate multiple scenarios (see in figure 5) with obstacles to train our policy network. We perform a multi-stage training of our policy network to aid policy convergence. We initially train the network on random scenarios from figure 5 - (A), without any obstacles. The map is divided into a 6×6 grid. A random point from two different grids is chosen as the source and target location for training the random policy. Once the robot converges with a policy to reach a goal without obstacles, we further train on random samples with obstacles. Sample maps are shown in figure 5 - (B to E). Furthermore, we place a random number of pedestrians, ranging anywhere from 5 to 50, with emotions and trajectories from the datasets by Randhavane et al. [35], and Bhattacharya et al. [36] for training our policy network. Thus, we train the robot to navigate with pedestrian proxemic constraints using RMSProp [41] with a learning rate of 0.00004 and $\epsilon = 0.00005$.

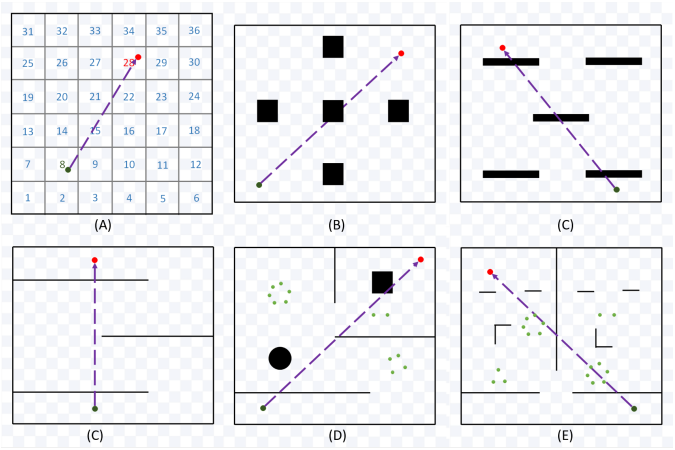


Fig. 5. *The various scenarios used to train our navigation algorithm. As mentioned in section IV-B.2, several configurations of parameters within a scenario are tweaked to generate multiple training/testing scenarios based on figures A-E.*

3) Analysis on Experiment Results: We tabulate the results from our experiment in Table I. We compare our implementation with [34] and [8]. We use scenarios 5 and 6 from figure 5 to perform our comparison studies. We also perform 20 iterations of the testing for each scenario under multiple environmental configurations (initial and goal positions, number and position of obstacles, pedestrian and emotion parameters). It can be seen that our implementation has the best Average Personal Space Deviation across all our experiments. Figure 6 showcases the experimental results of our comparison studies.

Method	Total Distance (m)	Avg. Time (min)	Δ_{ps}^{avg} (m)
Scenario 5			
Crowdmove [34]	152.5	2.6	25.3
RRT [8]	210.4	3.5	27.4
PRM [8]	217.6	3.6	26.8
Ours	172.3	2.9	**10.2**
Scenario 6			
Crowdmove [34]	224.5	7.2	57.8
RRT [8]	262.8	9.5	63.4
PRM [8]	245.6	8.7	61.8
Ours	237.6	8.7	**32.8**

TABLE I

COMPARISON OF *EWareNet* WITH VARIOUS BASELINES USING SCENARIOS 5 AND 6 FROM FIGURE 5. WE REPORT THE METRICS MENTIONED IN SECTION IV-B.1 COMPARED WITH [34] AND [8]. WE CAN SEE THAT OUR *EWareNet* HAS THE **BEST *Average personal space deviation*** (Δ_{ps}^{avg}) ACROSS ALL OUR EXPERIMENTS.

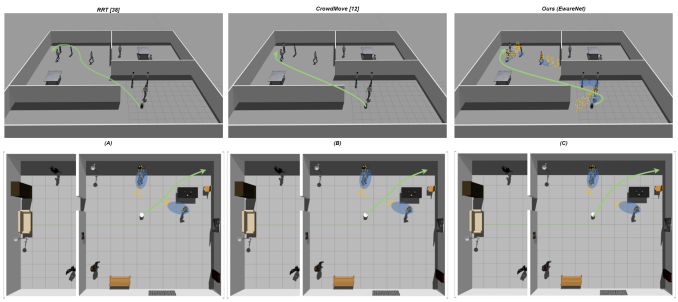


Fig. 6. ***Experimental Demo****: We showcase the performance of our EWareNet with other algorithms mentioned in Table I. Here the trajectory taken by the robot is depicted in Green, along with intent-aware trajectories as gaits in Yellow and adaptive personal/comfort space area around a pedestrian in Blue. Figures A & D represent the experiment with RRT [8], B & E represent the experiment with CrowdMove [34] and C & F represents EWareNet.*

V. CONCLUSION AND FUTURE WORK

We introduce a novel approach using *Transformer Attention Networks* to predict the full-body human intent or behavior in the form of trajectory/gaits based on historical gait sequences. We also present a novel navigation planning algorithm based on deep reinforcement learning that takes into consideration the environment, human intent, and the robot's reactionary impact on human behavior. Our method explicitly considers pedestrian behavior in crowds and the robot's impact on the environment and the uncertainty in the robot's sensor suite. Finally, our adaptive spatial density function to represent the proximal constraints for pedestrians captures their unique, personal comfort space in terms of pose, intent, and emotion. In the future, we plan to develop a platform to model pedestrian gaits that embody emotion to help benchmark emotion-guided social robot navigation algorithms. We also intend to extend our work to include other markers of emotion (such as facial expressions), specifically to address static pedestrians.

REFERENCES

[1] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, Jan 2013.

[2] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha, and A. Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," *IROS*, 2020.

[3] Tin Lay Nwe, Foo Say Wei, and L. C. De Silva, "Speech based emotion classification," in *Proceedings of TENCON 2001*, Aug 2001.

[4] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *2010 20th International Conference on Pattern Recognition*, Aug 2010, pp. 4605–4608.

[5] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," 2019.

[6] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *ICMI*. ACM, 2004.

[7] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 234–14 243.

[8] A. Vega, L. J. Manso, D. G. Macharet, P. Bustos, and P. Núñez, "Socially aware robot navigation system in human-populated and interactive environments based on an adaptive spatial density function and space affordances," *Pattern Recognition Letters*, vol. 118, pp. 72 – 84, 2019, cooperative and Social Robots: Understanding Human Activities and Intentions. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518303052

[9] J. M. Montepare, S. B. Goldstein, and A. Clausen, "The identification of emotions from gait information," *Journal of Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, 1987.

[10] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Deep learning for emotion recognition in faces," in *Artificial Neural Networks and Machine Learning*, A. E. Villa, P. Masulli, and A. J. Pons Rivero, Eds. Cham: Springer International Publishing, 2016, pp. 38–46.

[11] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175 – 1184, 2017, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[12] T. Randhavane, A. Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha, "Identifying emotions from walking using affective and deep features," *arXiv preprint arXiv:1906.11884*, 2019.

[13] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "The liar's walk: Detecting deception with gait and gesture," *arXiv preprint arXiv:1912.06874*, 2019.

[14] A. Bera, T. Randhavane, and D. Manocha, "The emotionally intelligent robot: Improving socially-aware human prediction in crowded environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[15] A. Bera, T. Randhavane, R. Prinja, K. Kapsaskis, A. Wang, K. Gray, and D. Manocha, "How are you feeling? multimodal emotion learning for socially-assistive robot navigation," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, 2020, pp. 894–901.

[16] A. Bera, S. Kim, T. Randhavane, S. Pratapa, and D. Manocha, "Glmp-realtime pedestrian path prediction using global and local movement patterns," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5528–5535.

[17] A. Bera, T. Randhavane, R. Prinja, K. Kapsaskis, A. Wang, K. Gray, and D. Manocha, "How are you feeling? multimodal emotion learning for socially-assistive robot navigation," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 894–901.

[18] V. S. Dorbala, A. Srinivasan, and A. Bera, "Can a robot trust you? a drl-based approach to trust-driven human-guided navigation," *arXiv preprint arXiv:2011.00554*, 2020.

[19] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6262–6271.

[20] D. Pavllo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," *arXiv preprint arXiv:1805.06485*, 2018.

[21] A. Abuduweili, S. Li, and C. Liu, "Adaptable human intention and trajectory prediction for human-robot collaboration," *arXiv preprint arXiv:1909.05089*, 2019.

[22] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 424–14 432.

[23] G. Li, H. Dibeklioğlu, S. Whiteson, and H. Hung, "Facial feedback for reinforcement learning: A case study and offline analysis using the tamer framework," 2020.

[24] R. Dabral, A. Mundhada *et al.*, "Learning 3d human pose from structure and motion," in *ECCV*, 2018.

[25] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," *arXiv preprint arXiv:2003.08111*, 2020.

[26] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," *ArXiv*, vol. abs/2202.07125, 2022.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners. arxiv 2020," *arXiv preprint arXiv:2005.14165*.

[29] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *arXiv preprint arXiv:1912.09363*, 2019.

[30] G. Ruggiero, F. Frassinetti, Y. Coello, M. Rapuano, A. S. Di Cola, and T. Iachini, "The effect of facial expressions on peripersonal and interpersonal spaces," *Psychological research*, 2017.

[31] R. Kirby, "Social robot navigation," 2010.

[32] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726 – 1743, 2013.

[33] J. Yang, Z. Ren, M. Xu, X. Chen, D. Crandall, D. Parikh, and D. Batra, "Embodied visual recognition," *arXiv preprint arXiv:1904.04404*, 2019.

[34] T. Fan, X. Cheng, J. Pan, D. Manocha, and R. Yang, "Crowdmove: Autonomous mapless navigation in crowded scenarios," *arXiv preprint arXiv:1807.07870*, 2018.

[35] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Learning perceived emotion using affective and deep features for mental health applications," in *2019 ISMAR*, Oct 2019.

[36] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, ser. AAAI'20. AAAI Press, 2020, p. 1342–1350.

[37] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[38] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and S. B., "PoseTrack: A benchmark for human pose estimation and tracking," in *CVPR*, 2018.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] R. Vaughan, "Massively multi-robot simulation in stage," *Swarm intelligence*, vol. 2, no. 2, pp. 189–208, 2008.

[41] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, no. 8, 2012.