Human-Robot Proxemics using Recurrent Neural Networks

Yuan Gao, Sebastian Wallkötter, Mohammad Obaid, Ginevra Castellano

Abstract— **In this paper, we investigate the applicability of deep learning methods to adapt and predict comfortable human-robot proxemics. Proposing a network architecture, we experiment with three different layer configurations, obtaining three different end-to-end trainable models. Using these, we compare their predictive performances on data obtained during a human-robot interaction study. We find that our long short-term memory based model outperforms a gated recurrent unit based model and a feed-forward model. Further, we demonstrate how the created model can be exploited to create customized comfort zones that can help create a personalized experience for individual users.**

I. Introduction

The advent of the ”AI era” machine learning, especially in the form of supervised deep learning [1], is helping to accelerate the state-of-the-art in numerous fields such as computer vision [2], self-driving cars [3] and natural language processing [4]. A driving idea is to automatically discover structure, namely input-output correlations or features, in a given dataset that can be used to predict the output for previously unseen input configurations. To achieve this, algorithms have to identify which inputs will improve their predictive ability and learn to become invariant against the unimportant ones. Considering this ability, machine learning techniques can be a suitable tool to be utilized within the Human-Robot Interaction (HRI) domain, in particular learning from users how to predict and respond to their social behaviours in an adequate way.

In this context, human-robot proxemics (interpersonal distances) is one of the topics that has attracted the attention of researchers for over a decade, as it influences how users perceive and interact with robots [5][6][7]. Previous research focused on understanding factors that could impact how human-robot proximity is shaped. However, little work has been devoted to allow robots to predict comfortable proximity towards human users, for example via machine learning algorithms [8]. This has motivated us to further investigate the use of deep learning techniques in predicting human-robot proxemics from experimental interaction data. We believe that the multivariate normal cumulative distribution function we learned around different people can be used to build a realistic reward function. This can then help robot to learn realistic behaviours using deep reinforcement learning algorithms.

In this paper we contribute to the HRI domain with the following: (1) The development of a machine learning system from user data to predict suitable human-robot proximity positions. This includes introducing a deep-learning model that is trainable end-to-end and proposing a pre-processing method to allow for the smoothing of the user data representing discomfort. (2) A comparative analysis of three model configurations investigating their suitability and performance when predicting human-robot proxemics.

II. Related work

Human-human interpersonal distances (proxemics) is one of the non-verbal behaviours that shapes the communication spaces between individuals. Vast research has been done to understand how humans position themselves within a given communication context such as [9][10]. Hall [11] presented one of the predominant theories defining proxemics into four zones: ”intimate”, ”personal”, ”social” and ”public” distances. His definition has served many researchers within the human-robot interaction community to investigate how human users position themselves when interacting with a robot in a given context and environment. For example, Walters et al. [12][7][13] presented studies addressing the impact of different social behaviours on human-robot proxemics. Satake et al. [14] contributed a study on understanding a robot’s approach behaviour and its influence on initiating interaction. Mumm and Mutlu [6] presented an investigation based on the theory presented by Argyle and Dean [9], where they report that one of the main factors to influence distance is disliking a robot.

Moreover, Takayama and Pantofaru [15] presented a study using a mechanical-like robot to investigate human-robot approaching distances in three scenarios: the robot autonomously approaching the user, the robot being teleoperated to approach the user, and the user approaching the robot. One of the findings indicated that pet owners as well as users that had previous experience with robots had a smaller interpersonal distance with the robot. In addition, they found gender to have an effect on the interpersonal distance when the robot gazed at the user. Złotowski et al. [16] provided empirical evidence on the influence of approach angle on proximity while the user is either walking or standing. Their results showed that when users were in

All authors are members of the Social Robotics Lab at the Visual Information and Interaction Division, Department of Information Technology, Uppsala University, Sweden alex.yuan.gao@it.uu.se

motion, they preferred the robot to approach them from the right and left frontal directions, but when standing it was acceptable for the robot to approach them from all directions (front, left and right).

In addition, researchers also looked at techniques for robots to learn the rules of such social interactions, for example via the adaptive strategies presented by Rossi et al. [17]. Some techniques focused on creating computational models that can enhance a robot's capability in path planning while considering social factors, such as the work presented by [18], [19], [20]. Other techniques include the work by Mitsunaga et al. [21] that used reinforcement learning for adaptive distance selection based on the user's body signals, or Mead and Mataric [8] proposing the use of Bayesian belief networks to model the probability of success in a human-robot interaction scenario. Kosinksi et al. [22], on the other hand, looked at using fuzzy logic modeling to directly learn and represent human-robot proxemics.

To our knowledge, little research has investigated the applicability of deep learning on the human-robot proxemics scenario so far. Hence we choose to investigate an end-to-end deep learning approach as we believe that it provides a more automated way of modeling influential factors. In this paper, we aim to learn such models while leveraging previous work to influence our choice of input features such as age, gender, height, pet ownership and previous robotic experiences.

III. Data Collection

The data used in our research is a subset of data used in an earlier work by one of this paper's co-authors, which is presented in Konsinski et. al. [22].

The collected data is gathered from 27 Swedish participants (13 female, 14 male). The participants average age is 30.5 years (SD = 10.7) and the majority had little to no experience with robots.

Data collection was setup in a motion capturing studio of 10×12.25 m equipped with a Qualisys system[1] that was used to track reflective markers.

Within the tracked space, a telepresence Double robot[2], with four reflective markers on the corners of its screen, has been used. In addition, reflective markers have been attached to a glove and a hat given to each participant to enable tracking of their positions. Generally, the hat has been used to locate the position of the participant, while the tracking of the glove has been used to measure discomfort. The user was asked to start raising their hand when they started to feel uncomfortable and they could stop the robot by putting their hand all the way up (stop gesture) - this is explained in further detail in [22]. For the remainder of this paper we will refer to this hand height as "discomfort level".

The procedure for each participant started by giving an introduction to the study and getting their informed, signed consent. Before commencing to the study sessions, the participant was asked to fill out a short demographic questionnaire. The study consisted of five short sessions, where Double would approached the user from five different angles and stopped when the user raised their hand and gestured a stop. The setup is illustrated in Figure 1 and the order of approach was randomized for each participant. After all sessions, the participant was asked to fill a post-study questionnaire followed by a short interview lead by the experimenter.

A. Extracted Data

Data has been extracted from the tracked recordings of the motion capturing system and serves as the bases for this work. We use a subset of this data, in particular, we only include data tracked from three angles (0, 45, 90) named A_3, A_4 and A_5 in Figure 1. In addition, we utilized the data collected during the initial demographic questionnaire.

The following inputs are available to each algorithm:

- distance between participant and robot (in mm)
- angle of approach (in degree, available for angles: 0, 45, 90)
- gender (male, female)
- age (in years)
- previous experience with robots (5-point likert scale)
- preferred writing hand (left, right)
- pet ownership (yes, no)

Using these values, the model predicts discomfort during it's approach towards the user.

IV. Machine Learning Setup

We consider different deep neural network architectures to predict a region where the robot should stop. We use the topology described in Figure 2 and vary the layer type of the fourth and fifth layers to create different networks.

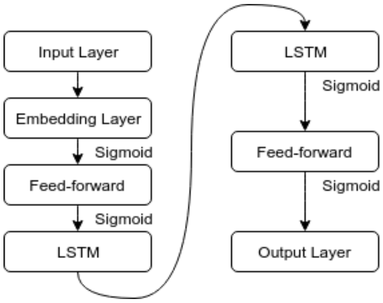

Fig. 2. The proposed topology of our networks, in which the fourth and fifth layer can be configured to other layer types.

In total we test three different layers that are frequently used in deep learning: (1) Feed-forward layers (FF) [23], (2) Gated Recurrent Units (GRU) [24] and (3)

[1] http://www.qualisys.com

[2] http://www.doublerobotics.com/

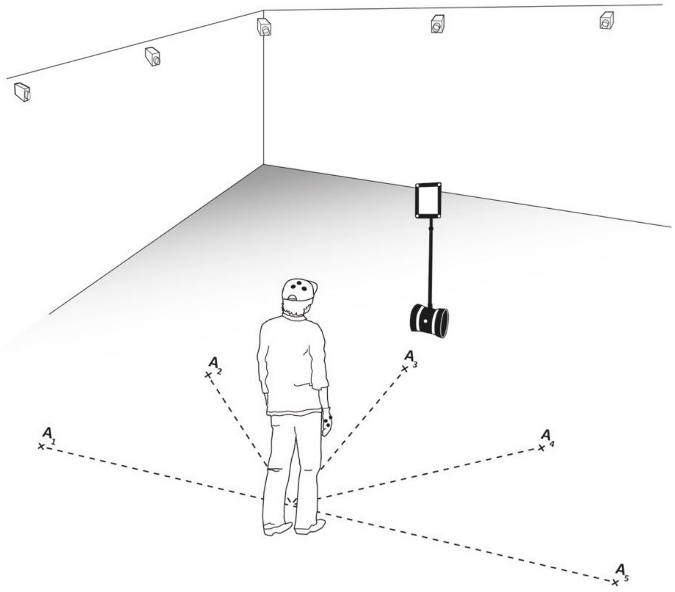

Fig. 1. The experimental setup of the motion capturing space with an illustration of the approach angels, the Double robot and a user in action. Only data from the three angles A_3, A_4 and A_5 is used in this paper. The image is adopted from [22]

Long Short-Term Memory (LSTM) [25]. Each model is trained in a supervised fashion; that is the training set, consisting of example inputs and desired outputs, is fed into the network to learn the relationship between them. The exact list of inputs is given in Section III-A and the predicted output for all networks is the perceived user discomfort.

We use the same general topology (Figure 2) for each model, starting with an embedding layer that maps all the features into a high dimensional space so that the model can represent interaction between factors more efficiently. This is followed by a FF layer with 128 hidden units as preparation for the core layers of each model: (1) two LSTM layers, (2) two GRU layers or (3) two FF layers. For each model, each core layer has 25 hidden units. We follow the core layers with another FF layer (10 hidden units) and then map the captured information to the output. The network uses sigmoid activation functions after each dense layer and the two core layers. The weights of the network are initialized using glorot uniform initialization [26] and are optimized using the Adam [27] method.

1) Feed-forward Layer: The FF Layer is one of the oldest layer types for neural networks and is conceptually quite simple. Mathematically the FF layer is defined as:

$$FF(\vec{x}) = \sigma(W\vec{x}) + \vec{b} \tag{1}$$

Where $\vec{x}$ is the input vector, W is a weight matrix and σ is a non-linear activation function.

The FF layer is used in the constant part of the architecture (see Figure 2) and, for one model, as an option of the core layers. Using them instead of LSTM or GRU layers creates a sequential, feed-forward model which, in our case, acts as a baseline to estimate the performance of the other models.

2) Long Short-Term Memory: Considering that our data has temporal dependencies, we use recurrent layers to learn from the input features of the system. At first, LSTM layers were introduced to address the vanishing gradient problem in recurrent architectures [28] and since then the model has proven to be very successful in learning long term temporal dependencies in a variety of tasks [29]. It works by using gated units to control how the information propagates through the network, i.e. three gates: input gate, forget gate and output gate, are used for controlling the cell. In our study, we would like to use LSTM for learning temporal dependencies to build a prior model for robot.

3) Gated Recurrent Unit: A Gated Recurrent Unit (GRU) is a simplification of aforementioned LSTM layer. It has less gates but achieves similar performance as a LSTM [30]. The first successful application of a GRU is in the field of Neural Machine Translation [31] and has since become popular in many tasks as in [32]. We hope the GRU model can serve as another model that can learn the temporal dependences in our dataset.Direct: r = 0.1()

V. Analysis and Results

In this section we introduce the modeling process of the topology presented in Figure 2 using LSTM layers as an example. Then we show the results of comparing the three models (FF, GRU and LSTM based) using human-robot proxemics data.

As a first step we convert the data into a machine learning friendly format by applying a smoothing technique and generating a statistical representation of the data. We normalize all discomfort levels from the three angles of each participant by fitting a Gaussian kernel based Cumulative Distribution Function (CDF) to them via kernel density estimation. The CDF is represented mathematically by formula $F(x) = \int_{-\infty}^{x} f(t)dt$, where $f(x)$ is any Gaussian function.

Figure 3 gives an example of the normalized discomfort levels for the robot approaching from Angle 0 in relation to the probability of the robot making a stop.

After this initial smoothing of the output data, we trained our system models (Figure 2) to learn the relationship between the user data and the target CDF, solving a regression task; this is illustrated in the example shown in Figure 4. To produce a cleaner result, we apply the same CDF smoothing technique to each of the model outputs that we used for the target denoising; this is illustrated in the exmaple shown in Figure 5.

Since we model the angle as a continuous variable rather than a discrete one, we are able to generate CDFs of the estimated discomfort from different unseen angles. Figure 6 shows an example of our model predicting several unseen angles (generated by doing

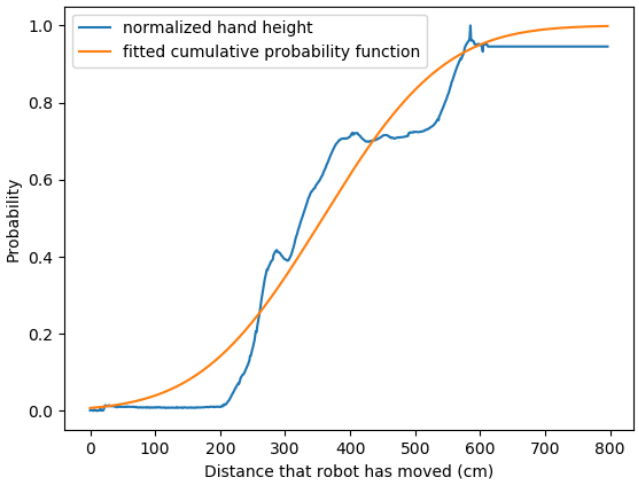

Fig. 3. The relationship between the estimated Cumulative Distribution Function (CDF) and the normalized discomfort level.

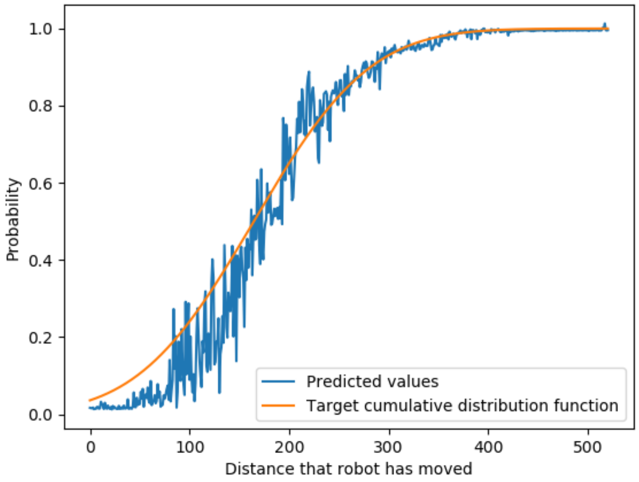

Fig. 4. The predicted probabilities values and its corresponded target values. The blue line indicates the predicted probabilities whilst the orange line shows the target CDF. We can observe that the generated values oscillates up and down around the target CDF.

an interpolation between the known angles (0, 45, 90)). We can see that the the model is able to generate similar CDFs based on the three angles provided.

To better understand how our trained models relate to findings in human-robot proxemics, we visualized the discomfort level in relation to the robot's space and its position. In this case, we first applied a Gaussian filter and selected discomfort probabilities from 0.2 (low discomfort level) to 0.8 (high discomfort level) as the thresholds. Figure 7 illustrates this relationship and highlights the two boundaries with labels 0.2 and 0.8, thus indicating the region where the robot should stop.

Next, we conducted further analysis to investigate the performances of the three deep learning layers by using the average euclidean distances between the model's outputs and the target CDFs. Thus, we compared the output of the three models using a test set (10% of the dataset), in which we train the network 50 times for each model and compare their average accuracy. The comparative results are shown in Figure 8.

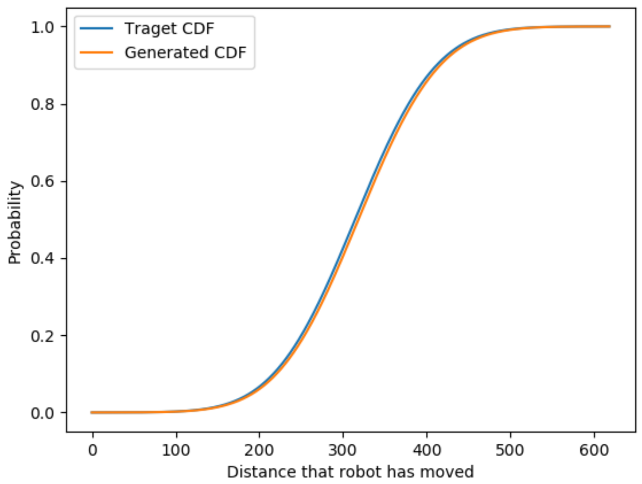

Fig. 5. The generated CDF by the model made with LSTM layers and the target CDF. We can see that for this particular angle of the participant, the two CDFs are almost indistinguishable.

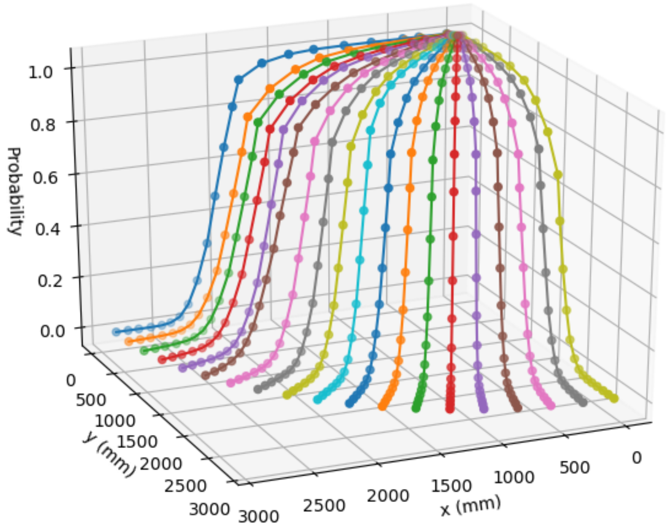

Fig. 6. The interpolated angles between 0 and 90 degrees. In this example, the generated CDFs are similar to the learned CDFs from 0, 45 and 90 degrees.

VI. Discussion

In the analysis of the results, our approach reveals several interesting discussion points. First, using the given data inputs, our deep learning approach was able to model and output proximity distances that are suitable in the HRI context, as illustrated in Figure 7. Considering the variability of human factors involved in determining appropriate proximity, we can foresee that deep learning approaches can be more suitable to such variability in contrast to other machine learning approaches. This is mainly due to the fact that classical approaches need a feature engineering phase before modeling the data inputs, while deep learning auto-

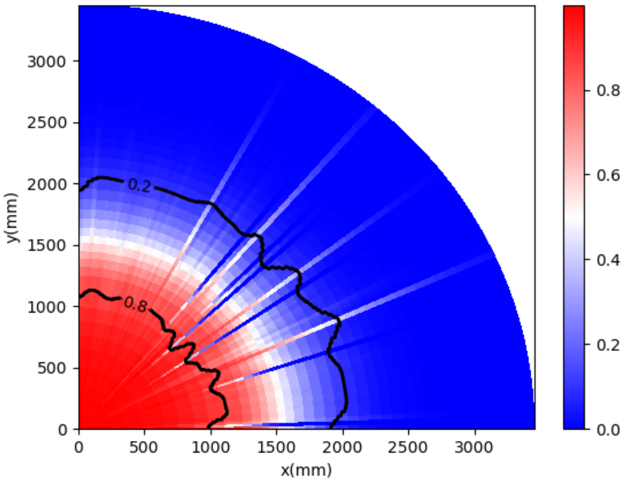

Fig. 7. The map of estimated probabilities of stopping on different positions. The participant stands on the origin, i.e. (0,0). The sector on the graph shows the area we consider. On this graph, only 0, 45 and 90 degrees are learned probabilities from the dataset. The other angles are generated from interpolating the angles and then sent to the model. The probability map is listed on the right side of the graph. Blue color indicates probabilities less than 0.5 and red color shows probabilities greater than 0.5.

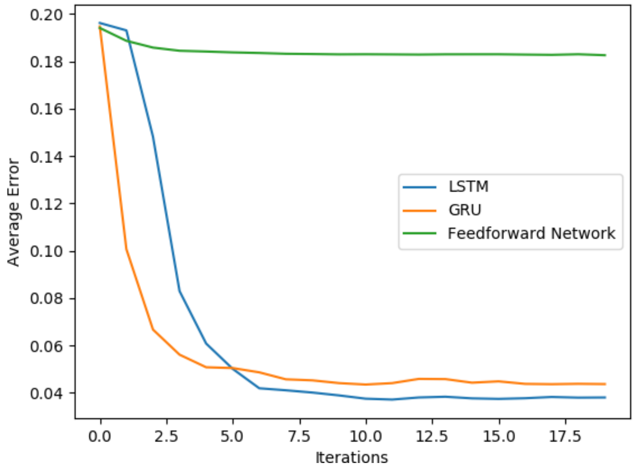

Fig. 8. The comparison of the three different configurations. GRU and LSTM have relatively good performance of fitting the target CDF. However, a simple FF neural network does not perform very well.

matically develops such abstractions; this can be seen in different research articles such as [1].

It is important to note that the introduction of a data smoothing phase is crucial in creating a suitable model to represent the input data. In our modeling process, we proposed an automatic pre-processing step that fits the CDFs to the data, thus allowing for better predictive performance. In addition, the pre-processing step allows the data to be defined statistically for further analysis in the system's pipeline.

Second, our approach illustrated that, using input data from three angles, we are able to model and predict a larger unseen region between the angles. This feature is due to the fact that deep learning models can accept continuous inputs and produce continuous outputs, thus making it possible to generate CDFs from unseen angles. One limitation we have noticed is that when we use a trained model to generate the CDFs over different angles for a single participant, the result can sometimes have a sharp transition. From Figure 7 we can observe that the probability transition among different angles are not smooth enough. On one hand, this is probably caused by the CDFs of the interpolated angles being influenced by the trend of other participants, and on the other hand, it also shows that we may need more data to have better CDFs.

Finally, when comparing the three different deep learning layers (FF, LSTM and GRU), we can see that the LSTM outperforms the other two layers in modeling the human-robot proxemics data, as shown in Figure 8. It is due to the fact that the LSTM based model has the lowest error in modelling target CDFs.

VII. Conclusion

We developed a system that can adapt and predict comfortable human-robot proxemics. The system first estimates CDFs of discomfort based on the users' hand height. It then uses a neural network architecture to learn the correspondence between users' discomfort and the distances that the robot has travelled towards the user from three angles. Thereafter, the system generates probabilities for unseen angles by interpolation and forms an area where the robot should stop.

Our experiment shows that among the three possible core layers of the neural network architecture, i.e. LSTM, GRU and FF layers, the configuration with a LSTM layer is the best at modelling HRI proxemics data. The final result in Figure 7 shows that we are able to produce a distribution estimation from only three angles. We also argue that because this model is neural network based and end-to-end trainable, it can lean from more data without any further modification.

Future work is directed towards improving the quality of the generated CDFs i.e. generating a probabilistic distribution with smoother transitions. In addition, we aim to investigate the performance of the system with more data and other configurations of the model.

Acknowledgment

The work was partly supported by the Swedish Research Council (grant n. 2015-04378) and the COIN project (RIT15-0133) funded by the Swedish Foundation for Strategic Research.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," *arXiv preprint*, 2016.

[4] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 873–883.

[5] M. Obaid, E. B. Sandoval, J. Zotowski, E. Moltchanova, C. A. Basedow, and C. Bartneck, "Stop! that is close enough. how body postures influence human-robot proximity," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Aug 2016, pp. 354–361.

[6] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, ser. HRI '11. New York, NY, USA: ACM, 2011, pp. 331–338.

[7] M. Walters, D. Syrdal, K. Koay, K. Dautenhahn, and R. te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *The 17th IEEE International Symposium on Robot and Human Interactive Communication, 2008. RO-MAN 2008.*, 2008, pp. 707–712.

[8] R. Mead and M. J. Matarić, "Autonomous human–robot proxemics: socially aware navigation based on interaction potential," *Autonomous Robots*, vol. 41, no. 5, pp. 1189–1201, 2017.

[9] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, vol. 28, no. 3, pp. 289–304, 1965.

[10] J. J. Hartnett, K. G. Bailey, and C. S. Hartley, "Body height, position, and sex as determinants of personal space," *The Journal of Psychology*, vol. 87, no. 1, pp. 129–136, 1974.

[11] E. T. Hall, *The Hidden Dimension*. Doubleday, 1966.

[12] M. L. Walters, K. Dautenhahn, S. N. Woods, K. L. Koay, R. Te Boekhorst, and D. Lee, "Exploratory studies on social spaces between humans and a mechanical-looking robot," *Connection Science*, vol. 18, no. 4, pp. 429–439, 2006.

[13] M. L. Walters, K. Dautenhahn, R. te Boekhorst, K. L. Koay, D. S. Syrdal, and C. L. Nehaniv, "An empirical framework for human-robot proxemics," in *Proceedings of the New Frontiers in Human-Robot Interaction : symposium at the AISB09 convention*, 2009, pp. 144–149.

[14] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, "How to approach humans?: Strategies for social robots to initiate interaction," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, ser. HRI '09. New York, NY, USA: ACM, 2009, pp. 109–116. [Online]. Available: http://doi.acm.org/10.1145/1514095.1514117

[15] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009.*, 2009, pp. 5495–5502.

[16] J. A. Złotowski, A. Weiss, and M. Tscheligi, "Navigating in public space: Participants' evaluation of a robot's approach behavior," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, NY, USA: ACM, 2012, pp. 283–284. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157795

[17] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for hri: a survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, 2017.

[18] O. A. I. Ramírez, H. Khambhaita, R. Chatila, M. Chetouani, and R. Alami, "Robots learning how and where to approach people," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),*. IEEE, 2016, pp. 347–353.

[19] D. Lee, C. Liu, Y.-W. Liao, and J. K. Hedrick, "Parallel interacting multiple model-based human motion prediction for motion planning of companion robots," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 52–61, 2017.

[20] R. Ramón-Vigo, N. Pérez-Higueras, F. Caballero, and L. Merino, "A framework for modelling local human-robot interactions based on unsupervised learning," in *International Conference on Social Robotics*. Springer, 2016, pp. 32–41.

[21] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Adapting robot behavior for human–robot interaction," *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 911–916, 2008.

[22] T. Kosiński, M. Obaid, P. W. Woźniak, M. Fjeld, and J. Kucharski, "A fuzzy data-based model for human-robot proxemics," in *proceedsings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 335–340.

[23] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[24] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

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

[28] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *arXiv preprint arXiv:1211.5063*, 2012.

[29] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[30] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 694–707, 2016.

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR*, 2015.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *International Conference on Machine Learning*, 2015, pp. 2067–2075.