Anticipating human presence for safer worker - robot shared workspaces

Emmanouilidis, Christos; Rica-Alarcón, Elena; Duqueroie, Bertrand

Published in:
Proceedings of the 2023 IFAC World Congress

Publication date:
2023

Anticipating human presence for safer worker - robot shared workspaces

Christos Emmanouilidis [*], **Elena Rica** [*], **and Bertrand Duqueroie** [**]

[*] *Department of Operations. University of Groningen, The Netherlands (e-mail: c.emmanouilidis@rug.nl)*
[**] *THALES SIX, France*

Abstract: The co-existence for human and mobile robots in modern industrial environments is increasingly common. Safety primitive behaviours are typically built-in mobile robots, to ensure safety. However, when fleets of multiple robots are operating in such environments, robot path planning becomes complicated and is often left sub-optimal to avoid compromising human, equipment, or process safety. Enhanced performance can be achieved if path planning takes into account not just current human presence, but projected human movement trajectories. While this problem has received extensive attention in outdoor environments in autonomous driving contexts, its indoors workspace equivalent has received less attention. This paper presents an approach for human movement prediction in industrial work environments, based on past and current heatmap occupancy grids and convolutional neural networks. The adopted heatmap format is appropriate for dealing with privacy concerns so as to avoid individual person identification. Obtained results from a range of simulation data are presented, following by a discussion on limitations, and challenges to be handled by further work.

Keywords: Human movement prediction, occupancy grid mapping, human-robot workspaces

1. INTRODUCTION

Although robots are capable of operating with efficiency in many processes, they present lack of flexibility to deal autonomously and with high operational efficiency in complex dynamic environments. Therefore, often industries seek to balance automation efficiency with the flexibility of involving the human in the loop (El Zaatari et al. (2019)). However, collaboration between robots and humans creates challenges for the safety of humans, equipment and processes (Müller et al. (2016)). As a consequence, human-robot collaboration spaces (HRCS) are being adapted to make possible the coexistence of humans and robots in a safe and secure manner (Robla-Gómez et al. (2017)). In such co-existence settings, mobile robots are typically equipped with systems that exhibit safety primitive behaviours, thus minimising relevant risks. Nonetheless, in scenarios when fleets of moving vehicles are needed, the operational performance impact of path planning hampered by robot movement stoppages triggered by human presence can be significant, leaving substantial room for further optimisations. Given the uncertainty inherent in a HRCS environment, it is desirable that a fleet of mobile robots is able to adapt in advance its individual and collective path planning, anticipating the movement trajectories of workers. Therefore, this paper deals with the problem of predicting human movement in industrial work environments. Aiming to act as precursor to a fleet of mobile robot vehicles paths planning, the developed predictive model receives as input outcomes from a video analytics module and produces as output projected heatmaps indicating future human presence. The remainder of the paper is structured as follows. Section 2 outlines related work. The problem formulation and proposed method are presented in section 3. Section 4 presents and analyses obtained results. Section 5 concludes the paper.

2. RELATED WORK

To approach the problem of human movement prediction, one can focus on the psychological perspective of the movement. With this strategy, it is assumed that people adjust their movement to the presence of other people and obstacles (Pellegrini et al. (2010)). Social forces such as desired destination and repulsion from other objects can be taken into account and the destination of one person could also affect the movement of other nearby people. Models to predict human trajectories in the street basing the models in social influences, position and velocity of each person have been presented (Luber et al. (2010)). A drawback of these approaches is that not all interactions can be captured and it is difficult to apply in crowded and complex scenarios. Regression models perform well on situations with missing and uncertain data to movement prediction of diverse targets and objects (Elnagar and Gupta (1998), Cai et al. (2006)). Prediction of pedestrians paths has been analysed through gaussian processes to model the uncertainty of unpredictability in human motion (Ellis et al. (2009)).

One of the most widely applied models in this research area is Hidden Markov model (HMM). HMM is based on discrete sequences and it is suitable for human motion prediction because the hidden state transition is useful in scenarios where the results of motion recognition are

uncertain (Liu and Wang (2017)). A downside of the HMMs is that they cannot learn the interactions between moving people in crowded environments and predict based on these interactions. Moreover, HMMs are only capable of providing short-term high-confidence predictions for motion planning. However, to make robot planning more efficient, longer term predictions are needed (Rudenko et al. (2020), Li and Fu (2014)).

Due to the spatio-temporal character of the data collected in these types of scenarios, deep-learning techniques are suitable (Wang et al. (2020)). Models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long-short term memory neural networks (LSTM) can be adapted to this problem taking into account the time and spatial dimension of the data. RNN and LSTM models can take into account sequences of time series data such as videos, which make them a popular modeling approach for predicting human movement. In (Alahi et al. (2016)) a trained LSTM allows sharing information from the hidden states of the LSTM with neighbouring moving individuals and adjusting the interactions between people in crowded situations. Bartoli et al. (2018), define a context-aware pooling layer that considers the fixed elements in the environment and models human-space interactions. LSTM models have also been applied to predict movement of pedestrians in crowded scenarios with adequate results (Su et al. (2017)). These characteristics make LSTM suitable for movement predictions in HRCS for a longer time horizons than other methods.

Although there is ample literature related to human movement prediction, research is more limited regarding such prediction in industrial workspaces. The characteristics or work spaces may have lower level of uncertainty compared to outdoor settings, as they may related to fairly well sturctured environments or specific work process sequences. Context-aware LSTM-based methods to predict workers' trajectories have also been proposed and used to calculate the optimal path of robots in exterior HRCSs. (Hu et al. (2020)). CNN models have been employed to predict human motion in work environments where tasks are developed with human-robot collaboration (Wang et al. (2018)). CNN models may be formulated to exploit spatio-temporal relations to detect interactions in dynamic environments, making them very suitable for movement predictions in HRCS for relatively long time horizons (Nikhil and Tran Morris (2019)).

Wang et al. (2015) considered mapping work environments into cell grids. The cell to which a subject will move is determined by the direction of cell entrance and the global starting point of the subject. In this grid, each cell represents a position in the workspace and is assigned a probability of being occupied. Another perspective is through a Bayesian approach of the problem in which the probability of each cell of being occupied is estimated by calculating a posteriori conditional probabilities, given a sequence of measurement vectors (Robbiano et al. (2022)). Nonetheless, complexity increases with higher grid dimensions. Furthermore, in real work environments, privacy concerns may limit the applicability of methods which can lead to the identification of individual persons. Therefore, methods which achieve adequate predictive performance, without compromising privacy are needed.

3. PROBLEM STATEMENT AND PROPOSED METHOD

Given a space that is partitioned into a grid of cells $P = \{P_i, i = 1, ..., N\}$ with N being the total number of cells in the grid, for each cell in the grid, and for each second t, a value $C^t(P_i)$ is assigned as follows:

$$C^t(P_i) = \begin{cases} 1, & \text{if the cell } P_i \text{ is occupied.} \\ 0, & \text{if the cell } P_i \text{ is not occupied.} \end{cases} \quad (1)$$

The problem is to determine if, given a known sequence in time $\{C^t(P_i)$, with $t = 0, ..., T\}$, it is possible to estimate the values of $\{C^{T+r}(P_i), r > 0\}$, where r indicates the future time window for the predicted human presence (grid occupancy). The proposed model is based on the CNN architecture, considering that CNNs can be applicable to spatio-temporal data. To explore this problem, a simplification assumption is made to consider a time window of three seconds in the time series of the occupancy grid. This assumption is based on the intuitive idea that workers have to go through the neighbouring cells to move from one cell to another one. The time window could also be set via estimating the average speed of a worker, during movements, but this can be seen as a further enhancement for the future. In a grid with N cells, the proposed strategy is based on the representation of each cell as a vector in $\mathbb{R}^n$ where n is the number of features obtained from the information of the cell. For each time instance t (one second in our experiments) and per each current cell P_i, $i = 1, ...N$ the features are defined as:

$$C^t(P_i) = \begin{cases} 1, & \text{if the } i^{th} \text{ cell is occupied in second } t. \\ 0, & \text{if the } i^{th} \text{ cell is not occupied in second } t. \end{cases} \quad (2)$$

The features C^t_s, D^t_s, are defined for $s = 1, 2, 3$ as:

$$C^t_s(P_i) = \begin{cases} 1, & \text{if the } i^{th} \text{ cell was occupied} \\ & s \text{ seconds ago.} \\ 0, & \text{if the } i^{th} \text{ cell was not occupied} \\ & s \text{ seconds ago.} \end{cases} \quad (3)$$

$D^s_t(P_i)$ = Sum of neighbouring cells occupied s seconds ago.

$$C^{t+1}(P_i) = \begin{cases} 1, & \text{if the } i^{th} \text{ cell is occupied} \\ & \text{for the next second in time.} \\ 0, & \text{if the } i^{th} \text{ cell is not occupied} \\ & \text{for the next second in time.} \end{cases} \quad (4)$$

$C^{t+1}(P_i)$ is the label that represents the occupancy of the cell P_i in the next second in time, $t+1$, and it is therefore the targeted output value for the predictive model. Thus, given a grid with N cells, at each time instance (each second in our experiments) and for every cell P_i, $i = 1, ...N$, there is a vector $v^t_i = (C^t, C^t_1, C^t_2, C^t_3, D^t_1, D^t_2, D^t_3)_i \in \mathbb{R}^7$ that represents the features of the cell and a value $C^{t+1}(P_i)$ that is the label that the method will predict. This feature set is employed for the training of the CNN-based predictive model, on the basis of past and current occupancy grid records.

3.1 Human trajectories data sets

In order to experiment with data sets which provide sufficiently persistent excitation for model training, our experiments are based on simulations which represent variations of workplace scenarios. Such simulations are currently employed for assessing crowding and human movement for critical infrastructure control centres, such as for airports and train stations, and therefore offer a rich testbed for creating realistic scenarios. This offers the possibility to model working spaces of varying complexity, structure, and with modifiable scenarios of human presence and trajectories.

The aforementioned simulations present four different industrial workspace layouts, comprising 8 types of working stations with 2 or 3 machines of each type, and supplementary attraction points. Such attraction points are used to emulate short breaks, for example as coffee machines. An example of a fictive factory can be seen in Fig. 1. In each of these fictive factories two simulations have been developed with 50 and 100 workers each. Thus, in total there are eight different sets of data corresponding to the eight simulations. Each simulation involves 4 types of persons, each with a role linked to a specific business process. These processes are mapped into sequences of worker visits to workstations, choosing among workstations of the same type the one with shortest queue. Furthermore, randomness is introduced using a uniform distribution over the number of workers of each type and a normal distribution over the worker movement speed. In addition, there is a 0.1 probability that workers deviate from their work process to go to the attraction point for such deviation, the coffee machine in this case. Each simulation has a duration of 2 hours and there is a text file representing the occupancy map at each time instance (each second in the reported experiments). Each text file represents an occupancy grid with 10000 cells of $1m^2$ each one. For each cell, there is a number indicating the number of workers currently present in the cell, otherwise the value is "0". These files also provide information about the immovable objects, such as well, workstations, or additional attraction points. Figure 2 represents a detail of the simulated work environment in which there is a work station in cells from one to eight, a wall in cells 12, 17 and 22. In cells 10, 11, 15 and 16 there are workers. It can be observed that one worker can occupy more than one cell at the same time. This implies that adding up the numbers in the neighbouring cells does not give the count of workers in the neighbouring cells as some of them may be counted more than once when their presence is mapped to more than one cell. Table 1 presents the average percentage of cells occupied with workers in all simulations. It can be observed that in all the cases the occupation is lower than 13%, which implies that the occupancy grid is sparsely populated.

3.2 Training and test sets

To build the training set of the model, the earlier defined features are extracted from each cell P_i in the data set. The first seven components $(C^t, C_1^t, C_2^t, C_3^t, D_1^t, D_2^t, D_3^t)_i$ are the input features of each sample in the data set and the value of $C^{t+1}(P_i)$ is the label that the model has to predict. The most crowded 15 minutes in each simulation have been selected for training set and the next 60 seconds of each one have been reserved as test set. An example of the occupation of cells in a workstation can be seen in Figure 3. Note that the number of occupied cells is not the same with the number of workers, as commented before.

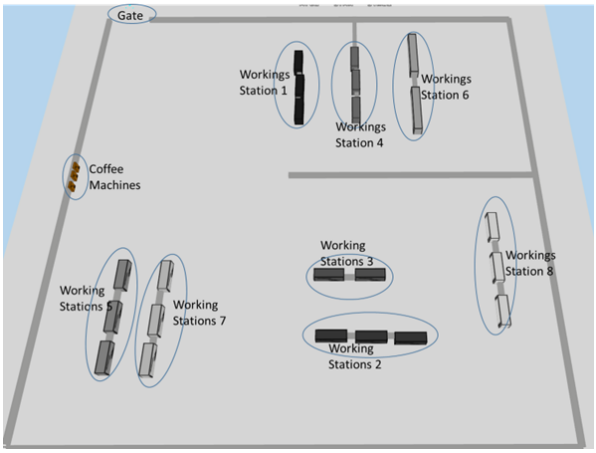

Fig. 1. Example of fictive factory.

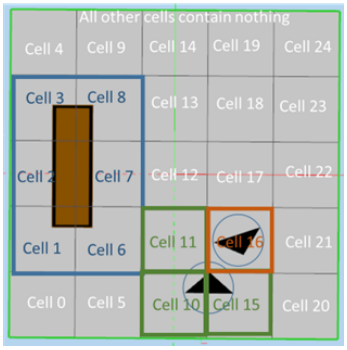

Fig. 2. Detail of the simulated work environment where in cells from one to eight there is a work station, and in cells 10, 11, 15 and 16 there are workers.

3.3 Structure of the CNN

To explore the predictive performance which can be obtained by training CNN models with such data, experiments involving different values of the hyperparameters of CNNs have been performed. The CNN structure has three convolutional layers with 128, 64 and 32 neurons, with correspondioning kernel sizes of 5, 2, and 1 each, and Rectified Linear Units (ReLU) as activation functions. A MaxPooling layer is applied after each convolution. The

Table 1. Average percentage of occupied cells in the different simulated environments.

Work environment	Nº of persons	% Cells with no presence	% Cells with presence
One	50	91.14	8.86
	100	87.79	12.21
Two	50	90.87	9.13
	100	87.63	12.37
Three	50	90.90	9.10
	100	87.67	12.34
Four	50	90.93	9.07
	100	87.60	12.40

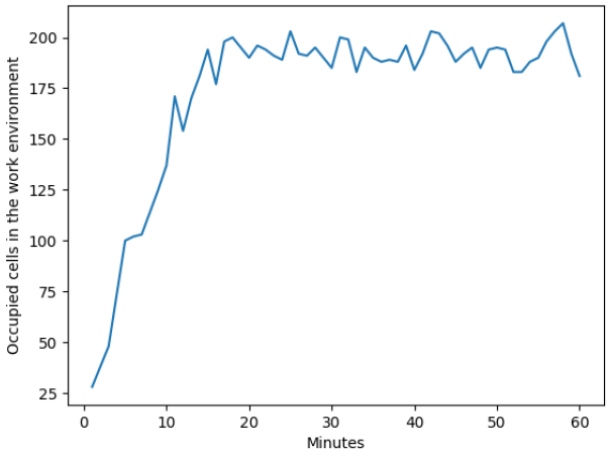

Fig. 3. Occupation of the cells during the simulation.

selected dropout rate to compensate for potential overfit was set at 0.2. The training has been performed for 100 epochs with a learning rate of 0.0001

3.4 Performance assessment

The performance of the methods is evaluated by the Root Mean Square Error (RMSE) for the regression outcomes of the model. When converting the model output to occupancy, Accuracy, Recall and Precision, where involved, as appropriate for classification:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Recall} = \frac{TP}{TP+FN}\text{ , Precision} = \frac{TP}{TP+FP}$$

With:
TP: Worker present in cell with model predicting presence.
FP: No worker in cell but model predicts presence.
TN: No worker present in cell, correctly predicted.
FN: Worker present in cell, correctly predicted.

4. RESULTS AND DISCUSSION

In this section the results obtained from the application of the CNN approach are presented and discussed. The CNN model directly produced outputs are real numbers, which can loosely be seen as belief or likelihood that a gird cell will be occupied, with higher values indicating higher such belief or likelihood. If this needs to be interpreted as a binary outcome (occupied or not), then network output can be passed through a threshold unit to decide whether the output of the model has to be interpreted as occupied or unoccupied cell. Specifically, for values higher than this threshold, the model will classify the cell as occupied. Different performance metrics, such as the values of root mean squared (RMSE), accuracy, recall and precision offer a different insight into the precictive performance. To analyse this influence, the CNN training has been performed with different thresholds.

For visual interpretability of the results, heatmaps representing the occupancy likelihood, as predicted by the model, have been created. These are illustrated in Fig. 4, Fig. 5, Fig. 6 and Fig. 7. Intuitively, for low thresholds more predictions are interpreted by the model as presence than with high thresholds. Therefore, low thresholds lead

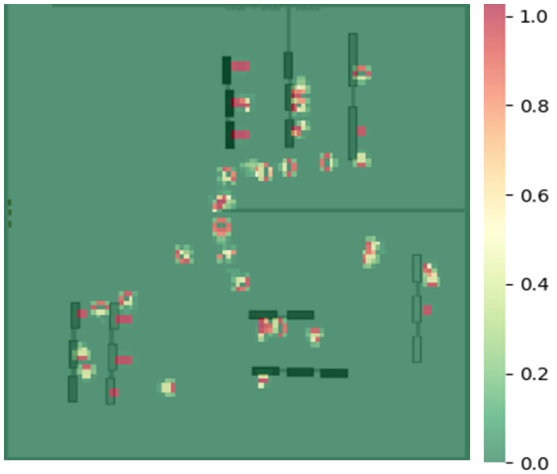

Fig. 4. Generated heatmap in workspace with 50 workers against the next second with threshold 0.3.

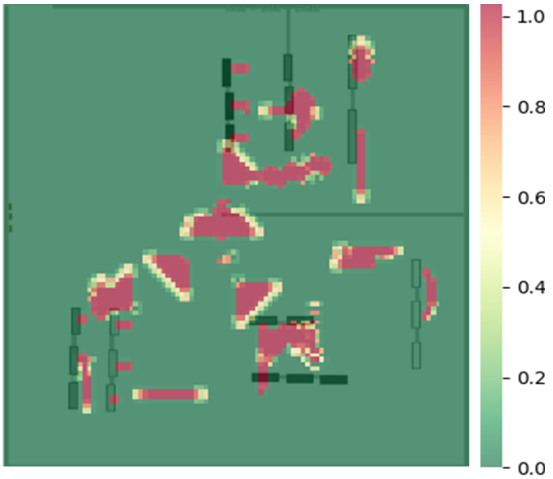

Fig. 5. Generated heatmap in workspace with 50 workers against the next 15 seconds with threshold 0.3.

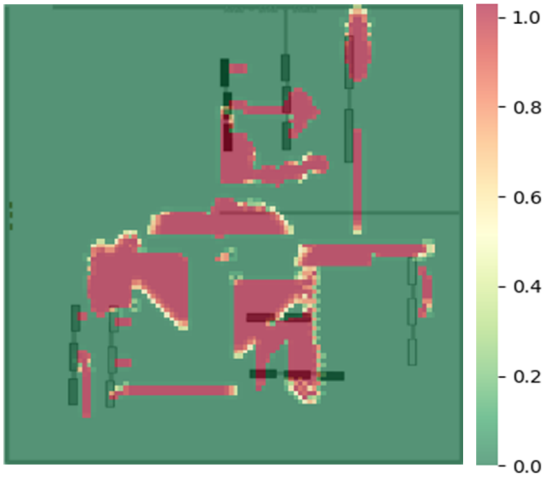

Fig. 6. Generated heatmap in workspace with 50 workers against the next 30 seconds with threshold 0.3.

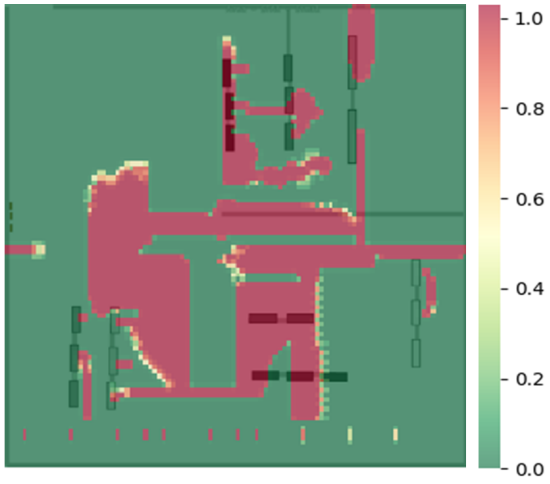

Fig. 7. Generated heatmap in workspace with 50 workers against the next 60 seconds with threshold 0.3.

to higher TPs, while high thresholds lead to higher FNs. Therefore, recall drops down significantly for thresholds 0.4, 0.5, 0.6 and 0.7 but increases for 0.2. However, lower thresholds of 0.2 and 0.3 also lead to a high RMSE and increased FPs. As it can be observed in Fig 8, for higher thresholds such as 0.6 and 0.7, the opposite happens. For these values, the probabilities of neighbouring cells are too low to be interpreted as presence and TP decrease over time. For instance, predictions obtained with threshold 0.6 one second ahead are very similar to the ones obtained with threshold 0.3 (Fig. 4). Nevertheless, the number of cells predicted as occupied after 15 seconds with threshold 0.6 represented in Fig. 9 is significantly lower than the number of cells predicted as occupied after 15 seconds with threshold 0.3 (Fig. 5). This is because the probabilities of being occupied are too low to be interpreted as presence and the number of cells predicted as occupied decreases over time leading to error propagation. Predictions beyond 15 seconds for the 0.6 threshold are not presented as not perceptible differences to the values obtained with 15 seconds have been found. It can be deduced that setting the threshold too high results in a lower rate of correct predictions of occupied cells (TP) for long time intervals.

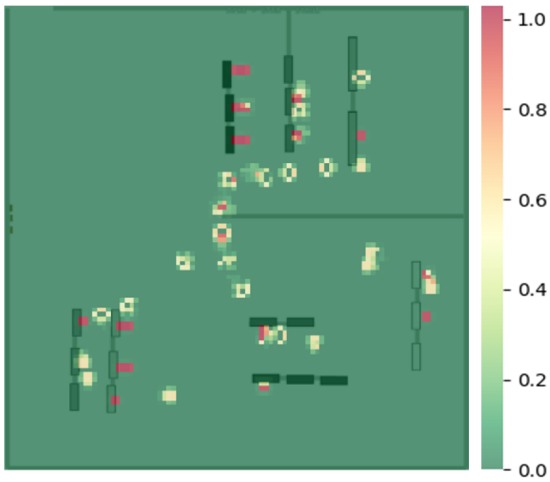

Fig. 8. Generated heatmap in workspace with 50 workers against the next second with threshold 0.6.

Table 2. Average RMSE, accuracy, recall and precision for predictions against a time horizon of 60 seconds with different thresholds.

No of workers	Thresh.	Average RMSE	Average accuracy	Average recall	Average precision
50	0.2	0.566	0.599	0.896	0.077
	0.3	0.333	0.867	0.650	0.125
	0.4	0.239	0.926	0.468	0.170
	0.5	0.184	0.957	0.338	0.221
	0.6	0.124	0.984	0.261	0.742
	0.7	0.123	0.984	0.246	0.744
100	0.2	0.576	0.583	0.919	0.130
	0.3	0.367	0.839	0.774	0.206
	0.4	0.245	0.924	0.654	0.314
	0.5	0.198	0.952	0.525	0.418
	0.6	0.147	0.978	0.476	0.855
	0.7	0.153	0.976	0.414	0.856

A summary of the results against a time horizon of 60 seconds with different thresholds is presented in Table 2. The recall values in Table 2 are higher for the experiments with 100 person than with 50 persons because the number of TPs is greater for experiments with 100 workers. However, the number of FP predicted for the experiments with 100 workers is also high, which results in a decrease in overall accuracy and precision. In more detail, Table 3 shows the results with time horizons 1, 15 and 30 in work space one with different thresholds for the cases of 50 and 100 workers respectively. It can be observed that accuracy of 0.99 and RMSE of 0.08 have been obtained with one step ahead predictions, but both values get worse when projecting further ahead into the future, due to prediction error accumulation. When the occupancy detection sensitivity is increased, by lowering the classification threshold, the higher is the number of cells that are predicted as occupied. Safety of workers should be a high priority in HRCS and therefore this justifies a higher rate of false positives. However, this also results in more cautious and therefore potentially sub-optimal path planning for the robot fleet. FN measures the number of cells that are occupied, but they are predicted as not occupied. If FN increases, it could lead to less secure work spaces than increasing FP. Thus, a low precision can be interpreted as a lower risk than a low recall.

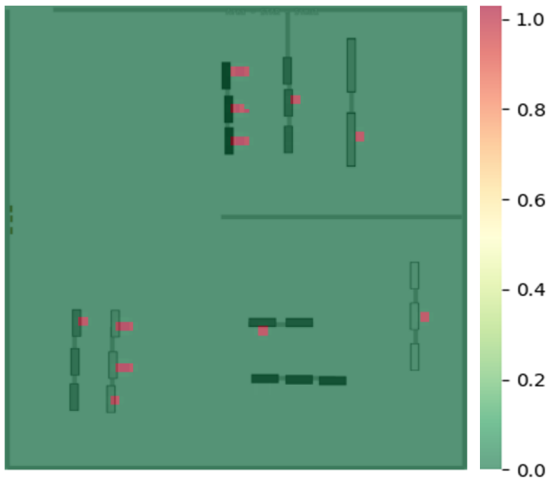

Fig. 9. Generated heatmap in workspace with 50 workers against the next 15 seconds with threshold 0.6.

Table 3. RMSE and classification performance; thresholds 0.2, 0.5; 50/100 workers.

50 workers

Threshold	Second	RMSE	Accuracy	Recall	Precision
0.2	1	0.080	0.989	0.856	0.648
	15	0.385	0.809	0.830	0.077
	30	0.613	0.584	0.957	0.041
0.5	1	0.080	0.991	0.802	0.743
	15	0.170	0.969	0.340	0.257
	30	0.192	0.955	0.310	0.154

100 workers

Threshold	Second	RMSE	Accuracy	Recall	Precision
0.2	1	0.123	0.974	0.803	0.747
	15	0.385	0.809	0.837	0.147
	30	0.614	0.581	0.919	0.073
0.5	1	0.090	0.988	0.876	0.801
	15	0.170	0.964	0.557	0.499
	30	0.213	0.947	0.528	0.345

5. CONCLUSION

This paper presented an exploratory analysis and a proposed model formulation for predicting human movement in indoor workspaces where humans and robots collaborate. Among the advantages of the approach is the ability to work without compromising individual person identification, thus respecting privacy. Current limitations relate to simplification assumptions, and the limited representation power of how input features are fed into the network. Future work can also consider the application of a Bayesian approach to align movement patterns to work process sequences, thereby biasing the prediction to the work context and working with additional realistic scenarios. Current work on mobile robot fleet planning is taking into account the human movement prediction to adjust and optimise the performance of the fleet, while ensuring human safety.

ACKNOWLEDGEMENTS

Research was supported H2020 grant ID 956573. Credits are due to the University of Groningen student Laurens Kuiper for his exploratory MSc thesis on the problem.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–971.

Bartoli, F., Lisanti, G., Ballan, L., and Del Bimbo, A. (2018). Context-aware trajectory prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 1941–1946. IEEE.

Cai, Y., Freitas, N.d., and Little, J.J. (2006). Robust visual tracking for multiple targets. In *European conference on computer vision*, 107–118. Springer.

El Zaatari, S., Marei, M., Li, W., and Usman, Z. (2019). Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems*, 116, 162–180.

Ellis, D., Sommerlade, E., and Reid, I. (2009). Modelling pedestrian trajectory patterns with gaussian processes. In *2009 IEEE ICCVC, 12th international conference on computer vision workshops*, 1229–1234. IEEE.

Elnagar, A. and Gupta, K. (1998). Motion prediction of moving objects based on autoregressive model. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 28(6), 803–810.

Hu, D., Li, S., Cai, J., and Hu, Y. (2020). Toward intelligent workplace: prediction-enabled proactive planning for human-robot coexistence on unstructured construction sites. In *2020 Winter Simulation Conference (WSC)*, 2412–2423. IEEE.

Li, K. and Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. *IEEE transactions on pattern analysis and machine intelligence*, 36(8), 1644–1657.

Liu, H. and Wang, L. (2017). Human motion prediction for human-robot collaboration. *Journal of Manufacturing Systems*, 44, 287–294.

Luber, M., Stork, J.A., Tipaldi, G.D., and Arras, K.O. (2010). People tracking with human motion predictions from social forces. In *2010 IEEE international conference on robotics and automation*, 464–469. IEEE.

Müller, R., Vette, M., and Mailahn, O. (2016). Process-oriented task assignment for assembly processes with human-robot interaction. *Procedia CIRP*, 44, 210–215.

Nikhil, N. and Tran Morris, B. (2019). Convolutional neural network for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 186–196. Springer International Publishing.

Pellegrini, S., Ess, A., and Gool, L.V. (2010). Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, 452–465. Springer.

Robbiano, C., Chong, E.K., Azimi-Sadjadi, M.R., Scharf, L.L., and Pezeshki, A. (2022). Bayesian learning of occupancy grids. *IEEE Transactions on Intelligent Transportation Systems*, 23(2), 1073–1084.

Robla-Gómez, S., Becerra, V.M., Llata, J.R., Gonzalez-Sarabia, E., Torre-Ferrero, C., and Perez-Oria, J. (2017). Working together: A review on safe human-robot collaboration in industrial environments. *IEEE Access*, 5, 26754–26773.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K.M., Gavrila, D.M., and Arras, K.O. (2020). Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8), 895–935.

Su, H., Zhu, J., Dong, Y., and Zhang, B. (2017). Forecast the plausible paths in crowd scenes. In *IJCAI*, volume 1, 2772–2778.

Wang, P., Liu, H., Wang, L., and Gao, R.X. (2018). Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. *CIRP annals*, 67(1), 17–20.

Wang, S., Cao, J., and Yu, P.S. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(8), 3681–3700.

Wang, Z., Jensfelt, P., and Folkesson, J. (2015). Modeling spatial-temporal dynamics of human movements for predicting future trajectories. In *Knowledge, Skill, and Behavior Transfer in Autonomous Robots", workshop in AAAI Conference on Artificial Intelligence, Austin, USA, January 25, 2015*, volume WS-15-09, 42–48. AAAI.