2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Collision Risk Assessment via Awareness Estimation Toward Robotic Attendant

Kenji Koide[1] and Jun Miura[2]

***Abstract*— With the aim of contributing to the development of a robotic attendant system, this study proposes the concept of assessing the risk of collision using awareness estimation. The proposed approach enables an attendant robot to assess a person's risk of colliding with an obstacle by estimating whether he/she is aware of it based on behavior, and to take the requisite preventative action. To implement the proposed concept, we design a model that can simultaneously estimate a person's awareness of obstacles and predict his/her trajectory based on a convolutional neural network. When trained on a dataset of collision-related behaviors generated from people trajectory datasets, the model can detect objects of which the person is not aware and with which he/she at risk of colliding. The proposed method was evaluated in an empirical environment, and the results verified its effectiveness.**

(a) (b) (c)

Fig. 1: Example of a risky situation, in which an elderly person with dementia is not aware of a traffic cone and is alerted by an attending professional caregiver.

I. Introduction

Dementia is a category of diseases that affect the health and the quality of life (QOL) of a large number of elderly people worldwide. A telling symptoms of dementia is a lack of attention to objects in one's environment [1]. Even though elderly people with early-stage dementia often retain normal bodily functions, their lack of attention puts them at risk of injury by, for example, bumping into obstacles and falling off stairs.

Fig. 1 shows an example of the risk of accident to an elderly person suffering from dementia and the preventative action taken by a professional caregiver to avoid it. In this case, the elderly person was not aware of an obstacle (a traffic cone) because he was distracted by something else (Fig. 1 (a)). The caregiver noticed that the patient was going to bump into the obstacle and alerted him accordingly. Although the elderly person still slightly stumbled over the obstacle (Fig. 1 (b)), he was able to avoid injury because of the alert (Fig. 1 (c)). If he had been walking alone, he might have been injured. This example illustrates the need for caregivers to attend to elderly persons. Developed countries, however, suffer from a chronic shortage of caregivers that makes it difficult to provide adequate care to elderly persons, which in turn affects their QOL.

Our motivation here is to improve the QOL of elderly people by developing a robotic attendant system that imitates the behavior of a professional caregiver and enables elderly people to walk freely while protecting them from accidents. As mentioned above, patients with early-stage dementia often retain normal bodily functions. We thus think that if appropriately alerted of situations that put them at risk of injury, they can avoid accidents by themselves, as shown in the example above. If they are informed of the presence of every obstacle in the environment, however, this might be annoying or overwhelming for them. It is therefore desirable to assess the risk of an accident and alert them only when necessary.

To strike a balance between safety and comfort, we consider assessing the risk of collision via awareness estimation. Fig. 2 illustrates the proposed robotic attendant system. The robot estimates whether the person is aware of an obstacle in his/her path and takes preventative action (e.g., informing the person or physically intervening) only when he/she is not aware of it. Such an attendant robot would be suitable not only for the elderly, but also for caring for children in private and public environments.

In this paper, we propose the concept of simultaneous awareness estimation and trajectory prediction and implement it using a convolutional neural network (CNN). Fig. 3 shows the proposed network. The network takes as input a sequence of local environmental maps in which the target person is located at the center. It outputs a collision risk map that represents the positions of obstacles in the area of which the person is not aware as well as a trajectory map that predicts their movement. To train this network on datasets on the behavior of people that do not contain sequences where a person actually bumps into an obstacle, we propose a method to generate unawareness-related behavioral data. An empirical assessment showed that the proposed method can identify obstacles in a person's path of which he/she is not aware.

*This work was in part supported by JSPS Kakenhi No. 25280093 and the Leading Graduate School Program R03 of MEXT.

[1]Kenji Koide is with the Department of Information Technology and Human Factors, the National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba, 3050061, Ibaraki, Japan, k.koide@aist.go.jp

[2]Jun Miura is with the Department of Computer Science and Engineering, Toyohashi University of Technology, Hibarigaoka 1-1, Tempaku, Toyohashi, Aichi, 4418580, Japan, jun.miura@tut.jp

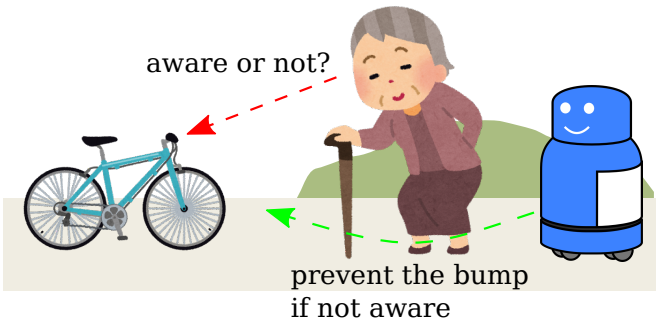


Fig. 2: Collision risk assessment via awareness estimation for a robotic attendant system.

II. Related Work

Awareness estimation has been widely studied in the field of driver assistance. When a driver is unaware of a pedestrian on the road, there is a high risk of an accident that can be prevented by informing the driver accordingly. A hidden Markov model proposed by Phan *et al*. estimates a driver's awareness of a pedestrian based on his/her driving operations and states, such as steering, accelerating, and braking [2]. Tateiwa *et al*. estimated the driver's awareness of pedestrians based on the motion of the car and developed a system to inform the driver of the risk of accident [3]. Bar *et al*. estimated the driver's awareness of pedestrians, other cars, and traffic cones by constructing a decision tree containing gaze and traffic obstacles as features [4]. Chutorian and Trivedi used a driver's head pose instead of gaze because gaze estimation is sometimes inapplicable to driving-related scenarios [5].

Awareness estimation has also been studied in human–computer interaction. Stiefelhagen and Zhu [6] estimated the focus of a person's attention based on the head pose to analyze discussion situations. Doshi and Trivedi [7] also estimated people's attention using head pose estimation to identify distracting objects in a meeting room. Dini *et al*. [8], [9] estimated people's awareness of objects of interest in a scene through 3D gaze analysis obtained from wearable eye-tracking glasses to improve human–robot collaboration in a manufacturing environment.

The above methods rely on features that directly reflect a person's awareness, such as gaze, head pose, and driving operations. In case of a mobile robot that follows a person, however, it is not feasible to observe these features without special devices, like eye-tracking glasses. In this study, we estimate a person's awareness of his/her surroundings based on his/her behavior because it is the most basic feature that can be observed by a mobile robot. A person's behavior contains enough information to estimate his/her awareness of the environment because a human observer can divine a person's intentions based on behavior without looking at his/her face (i.e., without gaze and head pose).

Koide and Miura proposed a method to estimate a person's awareness of an obstacle based on his/her trajectory [10]. They exploited the hidden conditional random field to classify pedestrian trajectories into cases of awareness and a lack of awareness. However, the target object and environment were limited in their work (comprising a box in a narrow corridor).

III. Methodology

We define the problem of awareness estimation as one of finding a function $\mathcal{F}$ that takes as input a person's sequence of behaviors $\mathcal{P}_0^t = \{p_0, \cdots, p_t\}$ and environmental information $\mathcal{E}$, including obstacle-related information, and outputs a list of obstacles $\mathcal{U} = \{u_0, \cdots, u_M\}$, of which the person is not aware, and a prediction of the person's behavior $\mathcal{P}_t^{t+N}$:

$$\mathcal{U}, \mathcal{P}_t^{t+N} = \mathcal{F}(\mathcal{P}_0^t, \mathcal{E}). \tag{1}$$

To implement this problem formulation, we propose a CNN-based model shown in Fig. 3 that takes a sequence of local environmental maps, and outputs the risk of collision to the person and maps of his/her predicted trajectory. Sec. III-A explains our design choices for the model, and Sec. III-B describes a method to generate data on people's behaviors when they lack awareness of their surroundings to train the proposed model.

A. Awareness Estimation and Trajectory Prediction Model

Input ($\mathcal{P}_0^t$ and $\mathcal{E}$): To estimate a person's awareness, the model needs to know his/her behavior, the positions of the obstacles, and the structure of the environment. Although [10] proposed hand-crafted features that describe a person's behavior with respect to an obstacle, this approach has two major drawbacks. First, features that can describe arbitrary behavior, objects, and environments are difficult to design. Second, the approach is difficult to extend to the multi-object case because we need to extract features for each

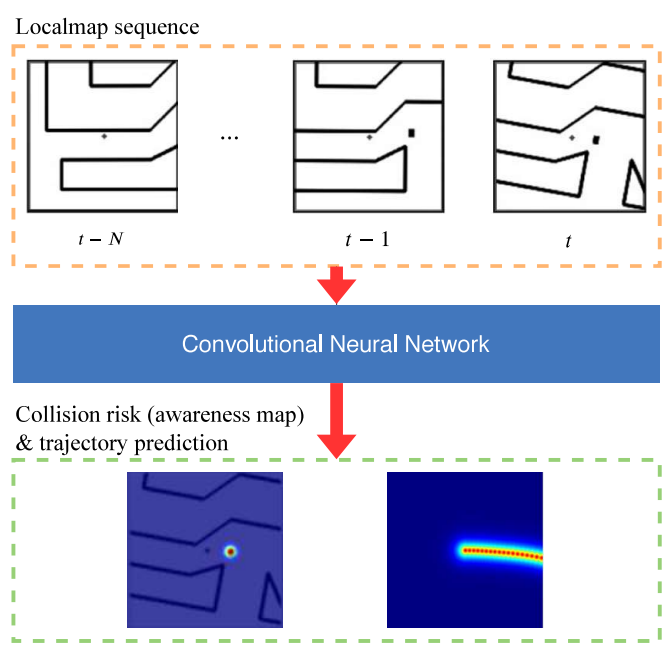


Fig. 3: Proposed awareness estimation model.

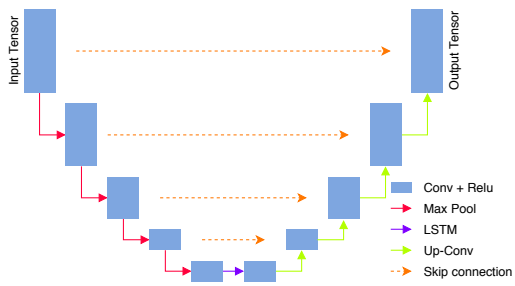


Fig. 4: *U-Net* architecture [11]

target object. We choose here as input a sequence of local environment maps, where the target person is located at the center, because this representation can represent any behavior and environmental structure.

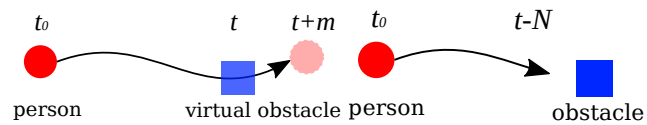
Output ($\mathcal{U}, \mathcal{P}_t^{t+N}$): Awareness estimation and trajectory prediction are a kind of chicken-and-egg problems. To estimate a person's awareness, we need to know how they will move, whereas we need to determine if a person is aware of objects in his/her surroundings to predict trajectory. We use an approach whereby the model simultaneously outputs the results of awareness estimation and trajectory prediction. We output the results of estimation in the image form because this can represent any distribution (e.g., multimodal distribution to represent multiple objects). Moreover, image-to-image translation has been successfully used in computer vision in recent years [12], [13].

Estimator ($\mathcal{F}$): As the backbone of the proposed model, we use the *U-Net* architecture [11] shown in Fig. 4. The convolutional layers were first applied to an input map to extract structured features, and deconvolution layers were then applied to obtain an output map with the same dimensions as the input. Skip connections transmit low-level data to the deconvolution layers to help them capture fine details of the input data. We extend *U-Net* to form a recurrent network by adding an LSTM layer at the bottom. We feed local environmental maps to the recurrent *U-Net* one by one, and the network outputs maps of the estimated awareness and the predicted trajectory at each time step.

The image-based input and representation of the output allows the model to capture arbitrary behavior, objects, environmental structures, and the relationship among them, in contrast with traditional behavioral models that are based simply on the distance between a person and an object, and their relative positions [14], [15].

B. Generating Unawareness-based Behavioral Data

An important issue when training the proposed network is that it is difficult to collect unawareness-related behavioral data on people for ethical and technical reasons. There is a risk of the subject being injured if he/she is not aware of an obstacle. Such experiments thus need to be very carefully controlled. Additionally, it is impossible to enable a person to intentionally become unaware of an object. It was thus not feasible to collect the large amount of unawareness-related behavioral data required to train the CNN model.



(a) Placing a virtual obstacle along a person's trajectory.

(b) Generated a virtual trajectory where the person bumps into the obstacle.

Fig. 5: Generating an unawareness-based behavioral data for people by introducing a virtual obstacle.

To avoid this problem, we propose generating data by introducing a simple assumption: "If a person is not aware of an object in his/her surroundings, he/she acts as if the object does not exist." For example, if a person is walking in a corridor and aware of an obstacle, he/she changes trajectory to avoid it. Conversely, if he/she is not aware of it, he/she moves as if there were no obstacle. The existence of the object does not influence the person's behavior if he/she is not aware of it.

This assumption allows us to generate the behaviors of people when unaware of obstacles from their usual behaviors, which are available in popular people tracking datasets [16], [17], [18], [19]. Because the above assumption implies that a person's behavior when unaware of an obstacle is independent of its properties, by placing a virtual obstacle along the person's position at time t, we can imitate an unawareness-based trajectory until t, when the person bumps into the virtual obstacle (see Fig. 5).

We first generated sequences of local environmental maps from the people tracking datasets and placed virtual obstacles along a person's path using random offsets. Maps of the estimated risk of collision and the predicted trajectory were then generated from positions of the virtual obstacles obs_j and the trajectory of the person p_t. The map of risk of collision $y^{\mathcal{U}}$ is defined as follows:

$$w_j = \exp\left(\frac{-\min_t \|obs_j - p_t\|}{C_\alpha}\right), \tag{2}$$

$$y^{\mathcal{U}} = \sum_j w_j \cdot \exp\left(\frac{-\|obs_j - x_i\|}{C_d}\right), \tag{3}$$

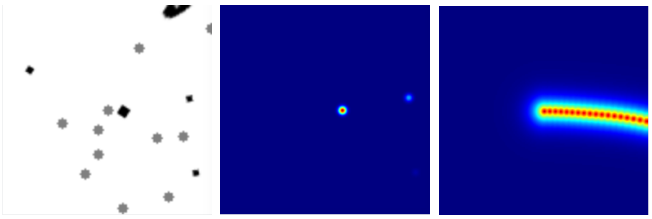
where x_i is the position of pixel i in the frame of the map, C_α and C_d are constants, and w_j is the weight of the obstacle j, given by the minimum distance between the trajectory p_t and the obstacle obs_j. We assign larger weights to obstacles that the person approaches because the risk of collision with them is higher.

The trajectory map $y^{\mathcal{P}}$ is defined as follows:

$$y^{\mathcal{P}} = \exp\left(\frac{-\min_t \|p_t - x_i\|}{C_t}\right), \tag{4}$$

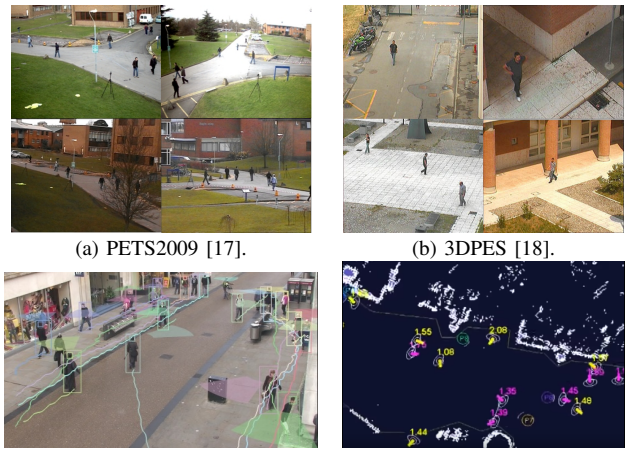
where C_t is a constant.

Fig. 6 shows examples of local input maps as well as the corresponding maps of the risk of collision and the predicted trajectory. In addition to virtual obstacles, we



(a) Input local map (b) Collision risk map (c) Trajectory map

Fig. 6: The generated local environmental map used as input, and the corresponding maps of the risk of collision and predicted trajectory. Black and gray pixels in the local map respectively represent surrounding obstacles and people.



(a) PETS2009 [17]. (b) 3DPES [18].

(c) Towncenter [16]. (d) ATC shopping center [19].

Fig. 7: Datasets used to generate the dataset of unawareness-based behaviors.

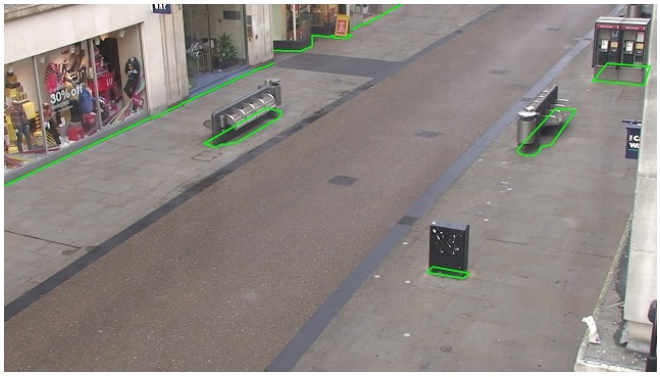
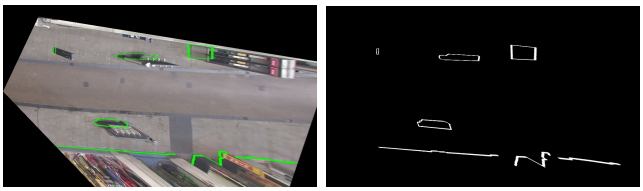


Fig. 8: Map of obstacle footprint created by hand.

rendered people on the local map so that the model could capture indirect interactions between people. The black and gray pixels in Fig. 6 (a) respectively represent the obstacles and the people. In the maps of the risk of collision (Fig. 6 (b)), a strong response appeared on an obstacle into which the person was going to bump, while weak responses were observed on the other obstacles farther from him/her.

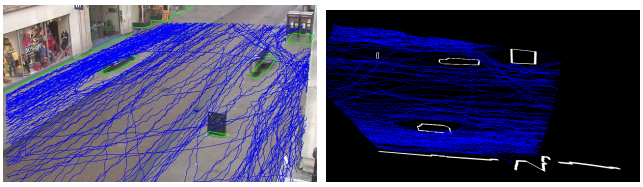
C. Training on Datasets of People's Behaviors

We used the *Towncenter* [16], *PETS2009* [17], *3DPES* [18], and *ATC shopping center* [19] datasets to generate a



(a) Projected map of obstacle footprint (b) Environmental map

Fig. 9: Generation of environmental map.



(a) Trajectories in the camera space (b) Trajectories projected in the bird's eye view

Fig. 10: People's trajectories in the *Towncenter* dataset.

dataset of unawareness-based behaviors (see Fig.7). The *ATC shopping center* dataset provided 3D trajectories of people in a shopping center. We created sequences of local maps by directly using the 3D trajectories. Because the *Towncenter*, *PETS2009*, and *3DPES* datasets provided people's trajectories as optical frames of a camera, we needed to transform the trajectories into a bird's eye view to generate sequences of local maps. We first created a map of obstacle footprints for each camera setting by hand (see Fig. 8) and projected it into the bird's eye view using the extrinsic parameters of the camera to obtain a 2D environmental map (see Fig. 9). We then projected people's trajectories on the 2D map to create local map sequences (see Fig. 10).

The number of sequences used for training was 2,000. We used the L2 loss function to calculate the residuals of the estimated risk of collision and the predicted trajectory maps. We assigned a small weight to the residual of the trajectory maps when calculating the L2 loss ($w = 1$ and 10^{-3} for the risk of collision and the trajectory, respectively) because maps of the risk of collision are more important for our application (i.e., finding dangerous obstacles), and people's trajectories are sometimes not predictable.

For verification, we tested whether the model correctly detected obstacles in which the person bumped using 100 randomly sampled sequences. We shifted the obstacles with random offsets such that their positions were different from those in the training set. We detected obstacles with high responses on the map of the risk of collision by thresholding. Fig. 11 shows a plot of the detection accuracy against the distance to the obstacle. Although the accuracy is low when the obstacle was far away from the subject (30% at 10 m), it increased as the person approached the obstacle. At 4 m from it, the detection accuracy was over 80 %. Assuming the person was walking at 1 m/s, the time to collision was about 4 s at this distance, which we think is sufficient duration for the robot to alert the person of the existence of the obstacle

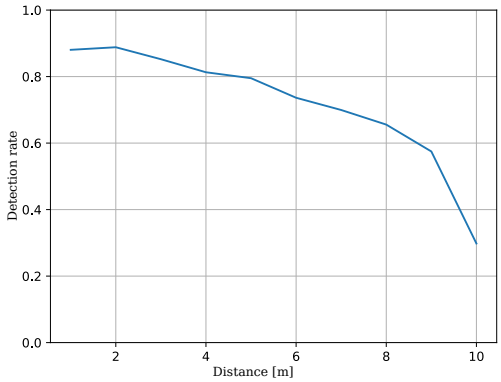


Fig. 11: Accuracy of collision prediction versus distance to the obstacle.

(a) Experimental environment (b) Half-blind glasses

Fig. 12: Experimental setting. The subjects wore half-blind glasses so that they could not see the obstacle on the ground. We predicted whether they bumped into the obstacles using the trained model.

TABLE I: Accuracy of collision prediction

	F1	precision	recall	TP	TN	FP	FN
Far (10 m)	0.909	0.833	1.000	10	3	2	0
Middle (5 m)	0.909	0.833	1.000	10	3	2	0
Near (1 m)	0.952	0.909	1.000	10	4	1	0
Total	0.923	0.857	1.000	30	10	5	0

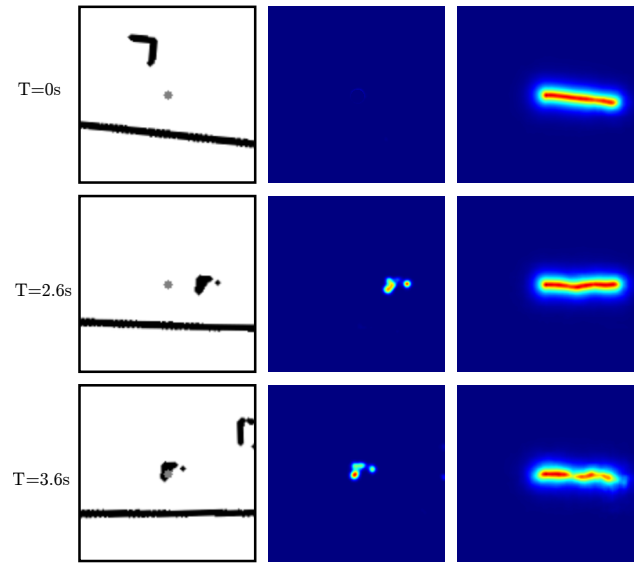
TP: True Positive, TN: True Negative,
FP: False Positive, FN: False Negative

and for them to alter their course.

IV. Experiments

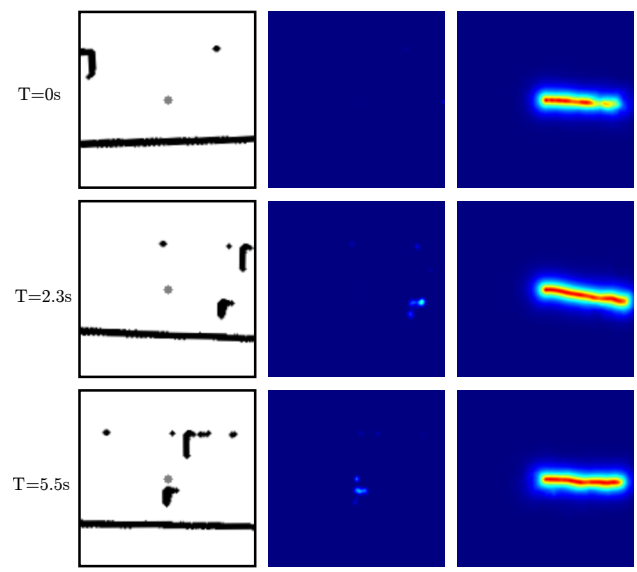
To validate the trained model in an empirical environment, we collected a dataset of people's behaviors (see Fig. 12 (a)). In this experiment, the subjects wore half-blind glasses (Fig. 12 (b)) so that they could not see an obstacle on the ground. The position of the obstacle was changed after every trial; therefore, the subjects did not know where the obstacle was, and bumped into it in some trials. We chose a light cardboard box as the obstacle and controlled the experiment to prevent serious accidents. We collected 15 sequences using four subjects (university students) in total. The subjects bumped into the obstacle in 10 out of the 15 sequences.

Fig. 13 shows a trial in which a person bumped into an obstacle. The figures on the left show the input local maps, and those at the center and on the right respectively show the estimated risk of collision and the trajectory maps. At $T = 0$ s, the model correctly predicted that the person would move along the corridor.



(a) Local map (b) Collision risk map (c) Trajectory map

Fig. 13: A trial in which the person bumped into an obstacle. The images to the left show the input local maps, and the maps in the center and to the right show the estimated risk of collision and the predicted trajectory, respectively.



(a) Local map (b) Collision risk map (c) Trajectory map

Fig. 14: A trial in which the person did not bump into an obstacle.

As the subject approached the obstacle, a strong response appeared at its position in the collision risk map ($T = 2.6$ s), and he eventually bumped into the obstacle at $T = 3.6$ s . Fig. 14 shows another trial in which the subject did not bump into the obstacle. In this trial, the collision risk map showed a much weaker response at the obstacle's position until the subject had passed it by ($T = 2.3$ and 5.5 s).

To quantitatively evaluate the proposed model, we sampled maps of collision risk at points where the distance between the obstacle and the subject was 10, 5, and 1 m,

and determined whether the model had correctly predicted whether the person would bump into the obstacle. We applied thresholding to extract high-response pixels from the map of collision risk. If the number of extracted pixels was larger than a threshold, the model was considered to have correctly predicted that the person would bump into the obstacle.

Table I shows the results of the evaluation. The positive cases represent those where the subject bumped into the obstacle while the negative cases represent the opposite. The model correctly estimated the subjects' awareness, and had a good recall rate in the positive cases. In two negative cases, where the person did not bump into the obstacle, the model incorrectly predicted otherwise. However, in one of the false positive cases, as the person approached the obstacle, the response of position of the obstacle in the collision risk map weakened, and the model correctly predicted that the subject would not bump into the obstacle when the distance to the obstacle was less than 5 m. In the other false positive case, the model could not properly estimate the subject's awareness until they had passed by the obstacle. In practical situations, however, some false positives are acceptable whereas false negatives are not. This is because they imply a failure to prevent accidents while a few false positives merely imply some unnecessary intervention.

V. Conclusion and Future Work

This paper proposed the concept of assessing the risk of collision via awareness estimation to strike a balance between safety and comfort, in the context of a robot attendant system to prevent elderly patients of dementia from getting into accidents in everyday activities. The proposed concept enables the attendant robot to estimate the risk of collision from a person's behavior and take preventative action only when required. As an implementation of the concept, we proposed a CNN-based simultaneous awareness estimation and trajectory prediction model that takes as input a sequence of local environmental maps, and outputs maps of estimated awareness and predicted trajectory. We also proposed a method to generate unawareness-based behavioral data from datasets of normal behaviors by people to train the proposed model. The model was evaluated in an empirical environment, and the results show that it can detect objects of which a person is not aware and with which he/she is going to collide. The proposed scheme can be applied to a more diverse population by using a larger and more varied dataset.

The proposed model was shown to be able to estimate a person's awareness; however, there is room for improvement to it. First, the input local environment maps can be extended to a more informative representation by including, for instance, distance and velocity maps such that the model can capture a person's behavior more easily. Second, a sophisticated object detection method, like the SSD detector, can be used instead of thresholding. That would allow us to more robustly distinguish between objects in the environment of which a person is aware and those of which he/she is not. Furthermore, the manner of preventative action is an important topic that needs to be investigated. It is not always possible to prevent an accident by merely informing the subject, such as the elderly suffering from dementia, of the presence of obstacles, and physical intervention may be required. In future work, we will develop hardware to appropriately intervene to correct the subject's movement to enable him/her to avoid obstacles.

References

[1] WHO Media centre, "Fact sheet: Dementia," 2017. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] M. T. Phan, V. Fremont, I. Thouvenin, M. Sallak, and V. Cherfaoui, "Recognizing driver awareness of pedestrian," in *IEEE Conference on Intelligent Transportation Systems*. IEEE, Nov. 2014, pp. 1027–1032.

[3] K. Tateiwa, A. Nakamura, and K. Yamada, "Study on estimating driver awareness of pedestrians while turning right at intersection based on vehicle behavior utilizing driving simulator," in *IEEE Intelligent Vehicles Symposium*. IEEE, June 2016, pp. 388–393.

[4] T. Bar, D. Linke, D. Nienhuser, and J. M. Zollner, "Seen and missed traffic objects: A traffic object-specific awareness estimation," in *IEEE Intelligent Vehicles Symposium*. IEEE, Oct. 2013, pp. 31–36.

[5] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 300–311, Apr. 2010.

[6] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *Extended Abstracts on Human Factors in Computing Systems*. ACM, Apr. 2002, pp. 858–859.

[7] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, June 2010, pp. 21–27.

[8] A. Dini, C. Murko, S. Yahyanejad, U. Augsdorfer, M. Hofbaur, and L. Paletta, "Measurement and prediction of situation awareness in human-robot interaction based on a framework of probabilistic attention," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Sept. 2017, pp. 4354–4361.

[9] L. Paletta, A. Dini, C. Murko, S. Yahyanejad, and U. Augsdrfer, "Estimation of situation awareness score and performance using eye and head gaze for human-robot collaboration," in *ACM Symposium on Eye Tracking Research & Applications*. ACM, June 2019, pp. 1–3.

[10] K. Koide and J. Miura, "Estimating person's awareness of an obstacle using HCRF for an attendant robot," in *International Conference on Human Agent Interaction*. ACM, Oct. 2016, pp. 393–397.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*. Springer, Nov. 2015, pp. 234–241.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, July 2017, pp. 1125–1134.

[13] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2019, pp. 2337–2346.

[14] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, pp. 4282–4286, May 1995.

[15] E. Hall, *The Hidden Dimension: Man's Use of Space in Public and Private*, ser. Doubleday anchor books. Bodley Head, 1969.

[16] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2011, pp. 3457–3464.

[17] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. IEEE, Dec. 2009, pp. 1–6. [Online]. Available: http://www.cvg.reading.ac.uk/PETS2009/a.html

[18] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *ACM Workshop on Multimedia access to 3D Human Objects*, Scottsdale, Arizona, USA, Nov. 2011, pp. 59–64.

[19] T. I. T. M. D. Brscic, T. Kanda, "Person position and body direction tracking in large public spaces using 3d range sensors," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 522–534, Oct. 2013.