HUMAN ROBOT INTERACTIVE INTENTION PREDICTION USING DEEP LEARNING TECHNIQUES

Do Nam Thang[1*], Nguyen Viet Tiep[2], Pham Trung Dung[2], Truong Xuan Tung[2]

Abstract: *In this research, we propose a method of human robot interactive intention prediction. The proposed algorithm makes use of a OpenPose library and a Long-short term memory deep learning neural network. The neural network observes the human posture in a time series, then predicts the human interactive intention. We train the deep neural network using dataset generated by us. The experimental results show that, our proposed method is able to predict the human robot interactive intention, providing 92% the accuracy on the testing set.*

Keywords: OpenPose; LSTM; Interactive Intention Prediction.

1. INTRODUCTION

In recent years, autonomous robots are increasingly researched, developed and applied in social life and in the military field. The strong development of the fourth scientific and technological revolution together with the trend of globalization has been a strong driving force in manufacturing technology and the application of autonomous robots in all areas of life.

Although current modern robot navigation systems are capable of driving the mobile robot to avoid and approach humans in a socially acceptable manner, and providing respectful and polite behaviors akin to the humans [1-3], they still surfer the following drawbacks if we wish to deploy the robots into our daily life settings: (1) a robot should react according to social cues and signals of humans (facial expression, voice pitch and tone, body language, human gestures), and (2) a robot should predict future action of the human [4]. Predicting human interaction intent is an important part of the analysis of human movement because it allows devices to automatically predict situations to actively set up respective action scenarios.

Human-robot interactive intention has been studied and incorporated into robotic systems. Human intention essentially means the goal of his/her current and/or upcoming action as well as motion towards the goal. The human intention was successfully applied to trajectory planning of robot manipulation [5-6], mobile robot navigation [7] and autonomous driving [8]. However, these motion planning systems only predict and incorporate the human motion intention for human avoidance, not human approaching which is essential for applications of mobile service robots.

There have been many authors using OpenPose and LSTM/RNN in human activity recognition (HAR) approach [9-11], but no author has used human posture to predict human-robot interactive intentions. And there are not any published data sets of the human-robot interactive intention by posture. We propose a new approach of human-robot interactive intention prediction using OpenPose and LSTM network.

2. BACKGROUND INFORMATION

2.1. Overview of OpenPose Model

OpenPose is a real-time multi-person keypoint detection library for body, face,

hands and foot estimation.

OpenPose was created by CMU Perceptual Computing Lab, the first version was released in July 2017, so far the latest version is 1.7.0. It supports Ubuntu (20, 18, 16, 14), Windows (10, 8), MacOS and NVIDIA TX2 embedded computers. The algorithm of OpenPose is detailed in the article [12] and [13].

The original OpenPose architecture consisted of a CNN network with two branches, in which the first branch was the reliability map and the second branch was the PAFs set.

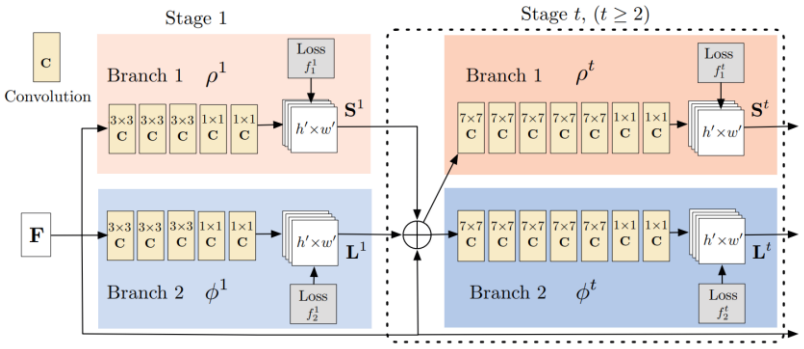

Fig. 1. *The original OpenPose architecture.*

The inputs of OpenPose are images, videos, webcams, Flir/Point Grey and IP Cameras. The outputs of OpenPose are basic images and keypoint display/saving in popular images formats such as PNG, JPG, AVI,... or keypoint saving as JSON, XML,... The number of body keypoints that can be exported is 15, 18 or 25-keypoint.

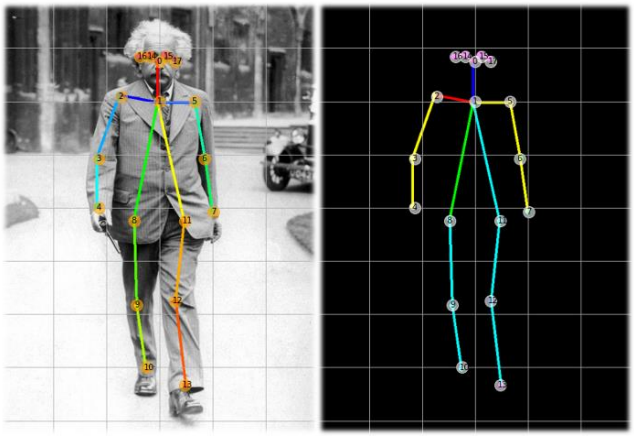

Fig. 2. *The output of OpenPose.*

In particular, the authors also provide API (Application Programming Interface) for two popular languages, Python and C++, which allow users to easily use OpenPose in their applications.

2.2. Long-Short Term Memory Technique

Long short-term memory (LSTM) is an artificial recurrent neural network

(RNN) architecture [14] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video).

A common LSTM unit is composed of a cell state, an input gate, an output gate and a forget gate. The cell state is used to remember values over arbitrary time intervals and the three gates are used to regulate the flow of information into and out of the cell.

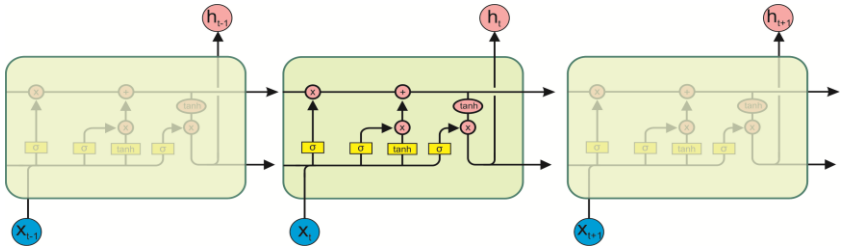

Fig. 3. *The architecture of LSTM network.*

LSTM networks are well-suited to classify, process and make predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

3. PROPOSED METHOD

We divide the human-robot interaction scenarios into nine cases shown in fig. 4.

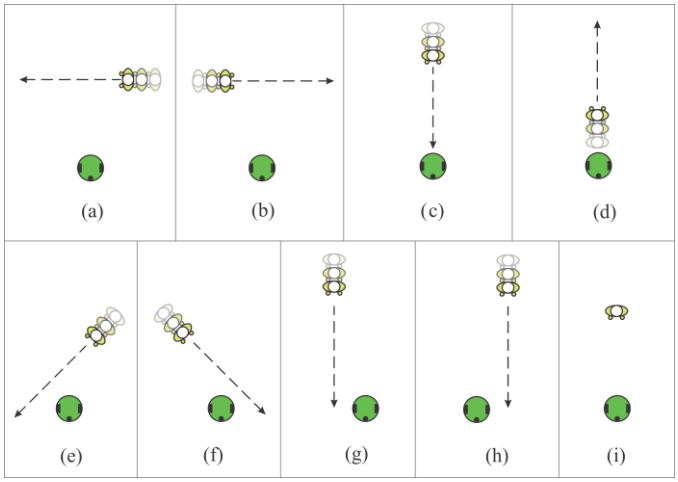

Fig. 4. *Human-robot interaction scenarios. (a) Human is crossing the robot to the left; (b) Human is crossing the robot to the right; (c) Human is meeting the robot; (d) Human is leaving the robot; (e) Human is avoiding on the left side of the robot; (f) Human is avoiding on the right side of the robot; (g) Human is moving towards the left side of the robot; (h) Human is moving towards the right side of the robot; (i) Human is standing.*

There have been many studies proving that a person's posture carries a lot of information, including emotions, health conditions [15]. Does person's posture contain information about intent to interact? We use the LSTM network, observe the person's posture in *n_steps* time steps, and then predict the person's intention to interact with the robot (fig. 5)

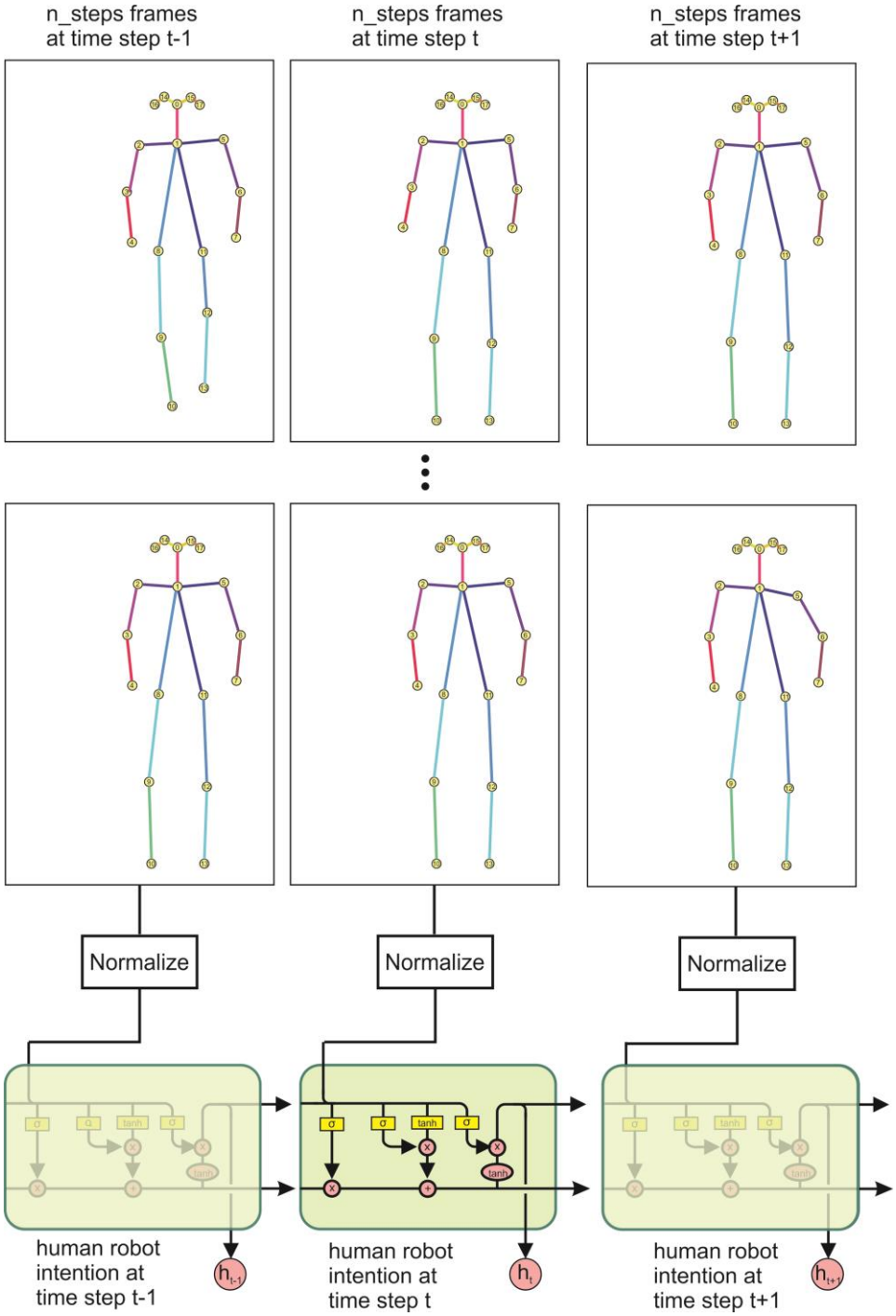

Fig. 5. *Our proposed method.*

Figure 6 gives an overview of the general flowchart of our process.

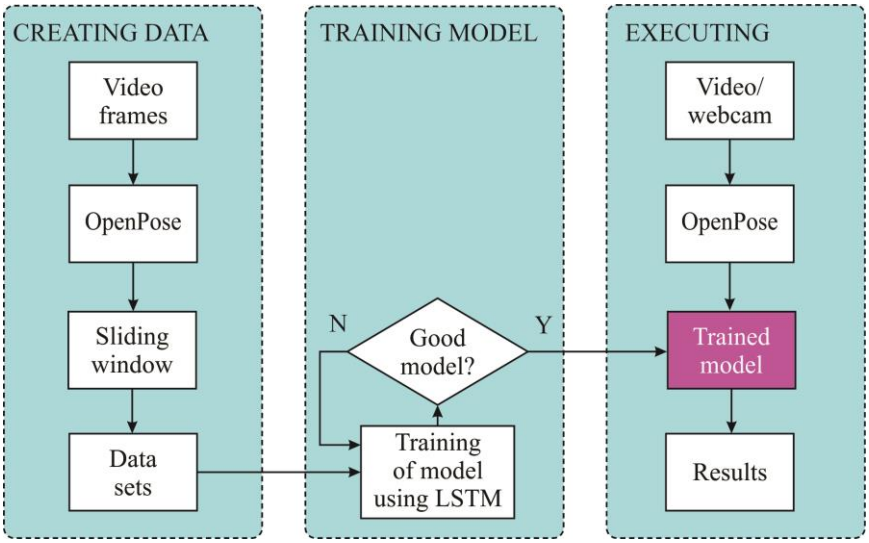

Fig. 6. *The flowchart of the propose system.*

Person's pose is extracted from OpenPose. A pose consists of 2D coordinates of *j* keypoints on the body, including 15, 18 or 25 keypoints. Each keypoint has two coordinates. Therefore at each time step we have a coordinate vector.

$$\mathbf{x} = [x_1, x_2, \ldots, x_k] \text{ with } x_i \in R,\ i = \overline{1,k},\ k = 2 * j.$$

The input of LSTM network is an *n_steps* x *k* matrix X, while the output are cases of human-robot interactive intention *n_case* (as shown above, *n_case* = 9), which is represented as an one-hot vector.

3.1. Data preparation

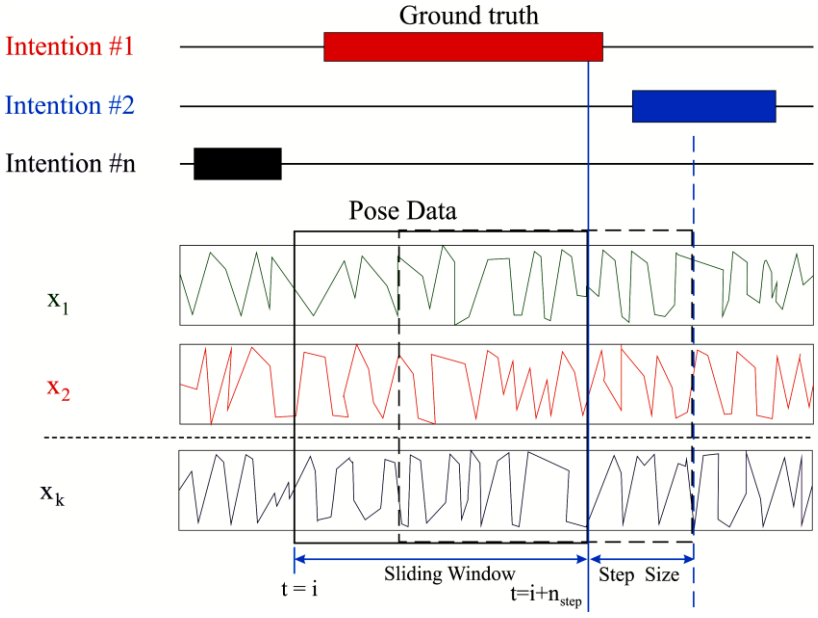

Fig. 7. *Sliding window.*

Dataset preparation is one of the most important step of deep learning model training process. It is crucial and can significantly affect the overall performance, accuracy and usability of trained model. The human-robot interactive intention dataset is not popular or not free in publishing so we created our own dataset by recording multiple videos in different environmental conditions.

To create the time step dataset, we used the sliding window technique, as shown in fig. 7. A window has *n_steps* width, which slides across the data series, for each step we have a data point and a corresponding label. The label is the intention of human-robot interaction.

The values of the keypoints in each window are written to input set X, while the ground truth, represented by a classification label, is written to output set Y. We do the same for training and testing set.

3.2. Training process

The dataset is split into two sets included 80% for training and 20% for testing. It is extremely important that the training set and testing set are independent of each other and do not overlap.

The batch size and epochs number were set in different values for training process. The training process was running automatically and finished when the pre-setting epochs number is reached. The model was saved after every certain number of epochs. At the end of the training process, we exported the prediction results on the training set and testing set to evaluate the newly trained model.

We filmed a variety of models with different heights, weights, and BMI (Body Mass Index) to create our datasets. The keypoints *j* is chosen 18, the time-step *n_steps* is chosen 32, so the input of the network is a 32x36 matrix.

We tested with different paramenters of the LSTM network to find the good ones. Several results were shown in table 1. The good network training results are shown in fig. 8. In this case, we tested with 30 hidden layers and 800 epochs.

Table 1. *Several training results.*

Training set	**Number Hidden layer**	**Batch size**	**Normalize data**	**Epoch**	**Training time** (minute)	**Accuracy**
15600	32	512	No	1100	40	74.3%
27736	34	512	No	1000	35	79.6%
27736	36	512	No	1200	45	79.4%
27736	36	512	Yes	1000	30	92.1%
27736	36	256	Yes	1000	75	89.5%
27736	36	8	Yes	800	200	88.1%
27736	30	512	Yes	800	30	92.7%

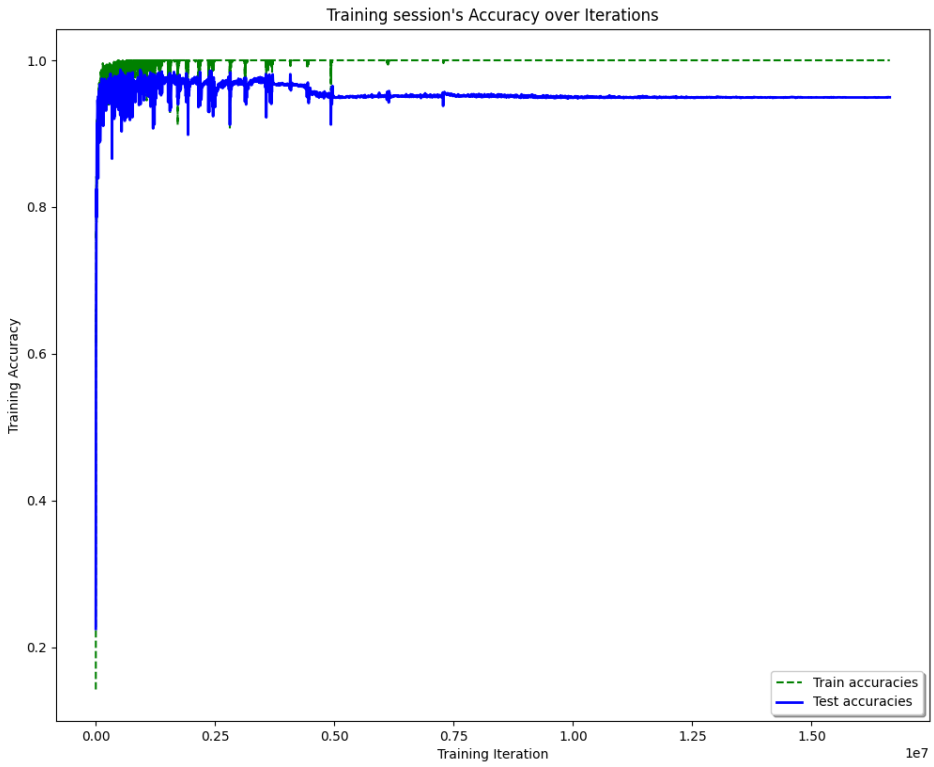

Fig. 8. *The prediction results after training.*

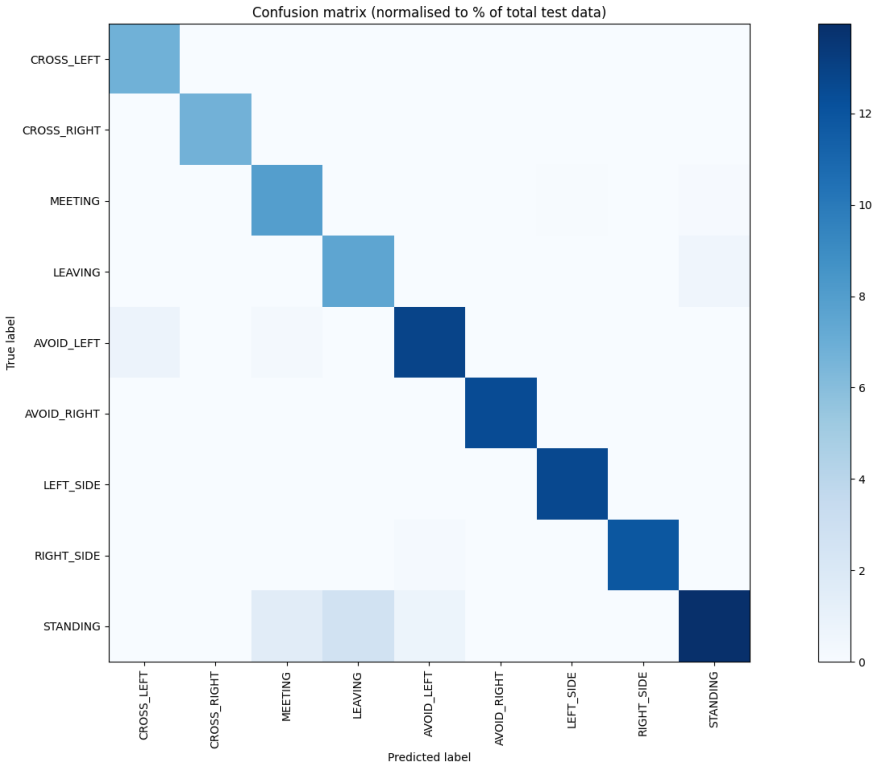

Fig. 9. *The results of evaluating the testing set.*

The results of evaluating the testing set on the confusion matrix are shown in fig. 9.

The evaluation results with the test set are very good. The model gets over 92% accuracy.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

We use smartphones to represent robots, which have a full HD (1920x1080) resolusion camera. Videos are scaled to 640x480 resolution before a being fed to the network model. The model stands 8-10m away from the robot and moves according to the scenarios, as shown in fig. 4. In the end, the model moves in a combination of several scenarios.

The testing process was run on a laptop with Intel core i7-8850 CPU, 8GB RAM and NVIDIA P1000 graphical card, which was installed Ubuntu 18.04 operating system.

4.2. Results

4.2.1. Single case results

The network model predicts very well with clear movements such as acrossing to the left, right, meeting and leaving, as seen in fig. 10 and fig. 11. In more difficult cases, for example a human avoids the robot to the left, as showwn in Fig. 12a or right (fig. 12b), at several early scenes, the network model may mistake for a human walking toward to the right, as illustrated in fig. 13b or left fig. 13a of the robot.

(a) (b)

Fig. 10. *The results of testing the network model (1).*

(a) Human is crossing the robot to the left;

(b) Human is crossing the robot to the right.

(a) (b)

Fig. 11. *The results of testing the network model (2).*
(c) Human is meeting the robot;
(d) Human is leaving the robot.

(a) (b)

Fig. 12. *The results of testing the network model (3).*
(a) Human is avoiding on the left side of the robot;
(b) Human is avoiding on the right side of the robot.

(a) (b)

Fig. 13. *The results of testing the network model (4).*
(a) Human is moving towards the left side of the robot;
(b) Human is moving towards the right side of the robot.

Fig. 14. *The results of testing the proposed model when the human is standing.*

4.2.2. Combination case results

In this case, the person moves combining 7 single cases, as shown in fig. 15. Although the movement is complicated, the proposed model is well predicted.

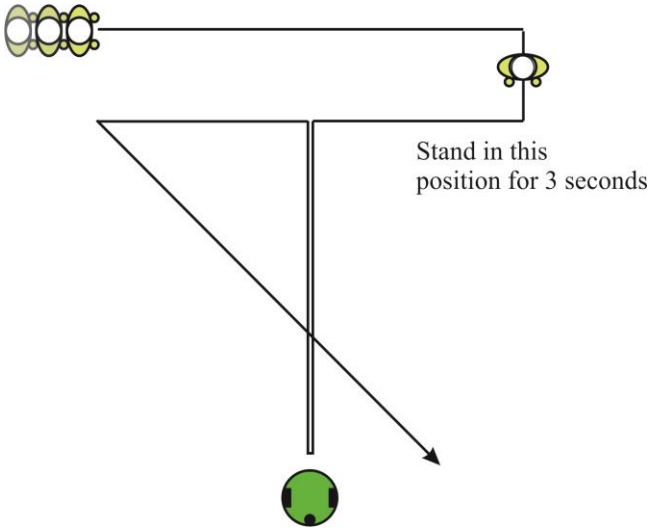

Fig. 15. *The combination case studies.*

5. CONCLUSIONS

In this article, we have presented an approach that predicts human-robot interactive intention using deep learning techniques. We utilized the OpenPose to extract human posture and LSTM network to observe a person over a certain period of time, then we predicted the human intent to interact with the robot. This approach initially gave some very positive results. In the future, we will continue to develop the algorithm by pre-processing information from the image and combine with information about euclidean distance from humans to robot to increase the prediction accuracy.

REFERENCES

[1]. M. Shiomi, F. Zanlungo, K. Hayashi, and T. Kanda, *"Towards a socially acceptable collision avoidance for a mobile robot navigating among*

pedestrians using a pedestrian model," International Journal of Social Robotics, vol. 6, no. 3, pp. 443-455, 2014.

[2]. X.-T. Truong and T. D. Ngo, *"Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model,"* IEEE Transactions on Automation Science and Engineering, vol. 14, no. 4, pp. 1743-1760, 2017.

[3]. Y. F. Chen, M. Everett, M. Liu, and J. P. How, *"Socially aware motion planning with deep reinforcement learning,"* in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017: IEEE, pp. 1343-1350.

[4]. X. T. Truong and T. D. Ngo, *"Social interactive intention prediction and categorization,"* in ICRA 2019 Workshop on MoRobAE-Mobile Robot Assistants for the Elderly, Montreal Canada, May 20-24, 2019.

[5]. Y. Li and S. S. Ge, *"Human–robot collaboration based on motion intention estimation,"* IEEE/ASME Transactions on Mechatronics, vol. 19, no. 3, pp. 1007-1014, 2013.

[6]. J. S. Park, C. Park, and D. Manocha, *"I-planner: Intention-aware motion planning using learning-based human motion prediction,"* The International Journal of Robotics Research, vol. 38, no. 1, pp. 23-39, 2019.

[7]. R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, *"Understanding human intentions via hidden markov models in autonomous mobile robots,"* in Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, 2008, pp. 367-374.

[8]. T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus, *"Intention-aware motion planning,"* in Algorithmic foundations of robotics X: Springer, 2013, pp. 475-491.

[9]. F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen, *"A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network,"* in Scandinavian conference on image analysis, 2019: Springer, pp. 299-310.

[10]. C. Sawant, *"Human activity recognition with openpose and Long Short-Term Memory on real time images,"* EasyChair, 2516-2314, 2020.

[11]. M. Z. Uddin and J. Torresen, *"A deep learning-based human activity recognition in darkness,"* in 2018 Colour and Visual Computing Symposium (CVCS), 2018: IEEE, pp. 1-5.

[12]. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, *"Realtime multi-person 2d pose estimation using part affinity fields,"* in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291-7299.

[13]. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. J. I. t. o. p. a. Sheikh, and m. intelligence, *"OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields,"* vol. 43, no. 1, pp. 172-186, 2019.

[14]. S. Hochreiter and J. J. N. c. Schmidhuber, *"Long short-term memory,"* vol. 9, no. 8, pp. 1735-1780, 1997.

[15]. 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,"* arXiv preprint arXiv:2003.01062, 2020.

TÓM TẮT

DỰ ĐOÁN Ý ĐỊNH TƯƠNG TÁC CỦA NGƯỜI ĐỐI VỚI ROBOT SỬ DỤNG KỸ THUẬT HỌC SÂU

Trong bài nghiên cứu này, chúng tôi đề xuất một phương pháp tiếp cận dự đoán ý định tương tác của người đối với robot. Phương pháp mà chúng tôi đề xuất sử dụng thư viện OpenPose cùng với mạng nơ ron học sâu Long – Short Term Memory quan sát tư thế chuyển động của người trong một vài bước thời gian, sau đó đưa ra dự đoán về ý định tương tác của người đối với robot. Chúng tôi đào tạo mạng nơ ron học sâu trên chính tập dữ liệu chúng tôi tổng hợp. Kết quả thí nghiệm chỉ ra rằng, phương pháp chúng tôi đề xuất đã dự đoán tốt ý định tương tác của người đối với robot với độ chính xác trên tập kiểm tra lên đến trên 92%.

Từ khóa: OpenPose; LSTM; Interactive Intention Prediction.

Received March 29th 2021
Revised May 06th 2021
Published May 10th 2021

Author affiliations:
[1]Academy of Military Science and Technology;
[2]Faculty of Control Engineering, Le Quy Don Technical University.
*Corresponding author: thangdonam@gmail.com.