 DOI: 10.3217/978-3-85125-663-5-12

RNN-based Human Pose Prediction for Human-Robot Interaction

Chris Torkar[1,2], Saeed Yahyanejad[1], Horst Pichler[1], Michael Hofbaur[1], Bernhard Rinner[2]

Abstract— **In human-robot collaborative scenarios human workers operate alongside and with robots to jointly perform allocated tasks within a shared work environment. One of the basic requirements in these scenarios is to ensure safety. This can be significantly improved when the robot is able to predict and prevent potential hazards, like imminent collisions. In this paper, we apply a recurrent neural network (RNN) to model and learn human joint positions and movements in order to predict their future trajectories. Existing human motion prediction techniques have been explored in a pseudo scenario to predict human motions during task execution. Building upon previous work, we examined their applicability to our own recorded dataset, representing a more industrial-oriented scenario. We used one second of motion data to predict one second ahead. For better performance we modified the existing architecture by introducing a different output-layer, as opposed to common structures in recurrent neuronal networks. Finally, we evaluated the artificial neuronal network performance by providing absolute positional errors. Using our method we were able to predict joint motion over a one second period with less than 10 cm mean error.**

I. INTRODUCTION

Robots are widely used in different fields due to their precision, reliability, strength, and speed. Initially, in industrial applications, robots operated separated from humans in isolated areas. With advances in technology and the necessity for coexistence of robots and humans (e.g., medical application, service robots, collaborative production lines), a new era of human-robot interaction has emerged. Human-robot collaboration is one of the key aspects of future industrial manufacturing. When humans work closely with robots, the safety of the human becomes an important issue [19].

No matter how accurate and safe a system is designed, continuous monitoring is still required to ensure safety in collaborative scenarios. Naturally, perception plays an important role, identifying hazardous scenarios by using various types of advanced sensors. For instance, robot-integral force and torque sensors are used to detect collisions in order to immediately stop the robot. In addition, tactile and touch sensors can be used to detect collision areas, to further improve a robot's reactive capability. Quite recent research showed how proximity sensors even can detect an approaching object at closer ranges and prevent a collision in advance [16], [18]. Similarly, many other sensors such as conventional cameras, RGB-D cameras, time-of-flight cameras, and laser scanners can be used to monitor the environment allowing for prevention of hazardous situations. However, performing sensory data acquisition, analysis, and fusion is a computationally expensive task. Therefore, in presence of fast robot and human motions, the robot may still need extra time for a reliable perception of the environment. By predicting human and robot motions in advance, the robot can foresee hazardous situations and adjust the robot movement trajectory or speed according to safety standards [8].

Advances in machine learning and deep neural networks brought new forms of scene understanding. Miseikis et al. [14], [15] showed how to find a robot and localize its 3D joint positions using a single image applying a convolutional neural network. With an increasing number of available datasets containing a variety of human motions [7], the momentum of applying machine learning to human motion related tasks increases. For instance it is possible to extract the 2D joint positions of multiple humans from an image or video in a timely manner [1]. By using recurrent neural network topology, [13] showed that a short term human motion prediction is possible based on observed human motion patterns. Even more recent findings [3] show smoothly synthesized long-term motion and short-term prediction of human motion. As shown by [4], accurate motion predictions have been accomplished by using action-specific motion patterns to train a time-series-aware neural network to predict human motion.

These methods build the foundation to incorporate motion prediction in robotic systems and enable them to plan their own movement ahead of time and avoid risky situations such as collisions with other humans or robots in collaborative scenarios. In this paper, we investigate if existing human motion prediction, using a RNN, is in principal applicable to collaborative human-robot scenarios. The goal is to evaluate the prediction accuracy of human joint motion for up to one second. To achieve this we adapted an existing RNN topology and evaluated the performance on our dataset representing a simple industrial assembly task.

II. RELATED WORK

Here we provide an overview of the most relevant literature. Our work builds upon prior research on human-robot collaboration and human-motion modelling using artificial neural networks.

Human-robot collaboration - Previous works in the field of safe robot operation in conjunction with human interaction e.g. [5], [10], have led the consideration and characterization of safety requirements. Safety is considered as the most important design criterion in drafting a new robotic system or implementing an industrial robot in a manufacturing process. Furthermore, it is desired to account for more convenient robot interaction with humans and more efficient operation.

[1] Joanneum Research - Robotics first.lastname@joanneum.at
[2] Alpen-Adria-Universität Klagenfurt first.lastname@aau.at

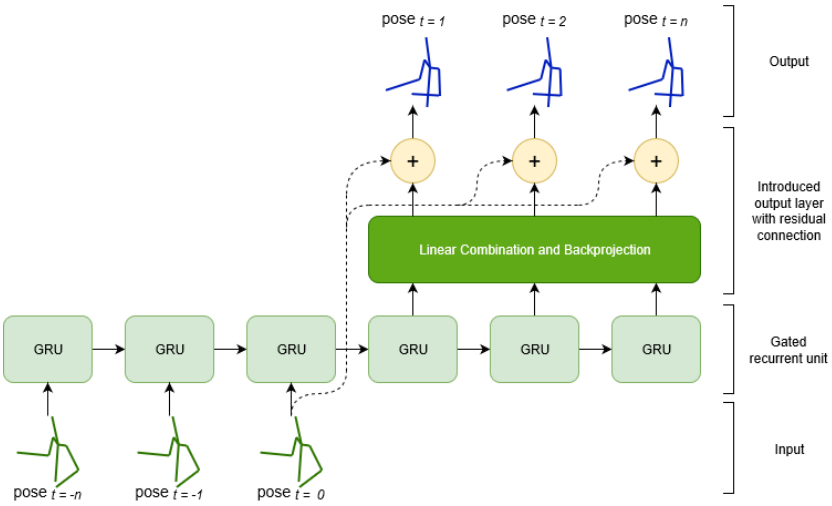

Fig. 1. Proposed architecture with extended back projection layer (green box). Green skeletons represent the input sequence and blue skeletons represent the predicted output. The dotted line show the residual connection added to the final pose output.

Similar related work has been published by [11], [12] who propose human motion prediction methods for human-robot collaborative tasks. Being able to foresee human motion in a short time span, prior information can be used to avoid hazardous situations up front.

Liu and Wang [11] intended to model a product assembly operation as a sequence of tasks. Decomposing an assembly operation in different tasks with specific motion patterns allowed them to map their assembly problem to a Hidden Markov Model. Hidden Markov Models are well suited for discrete sequence models. They showed the feasibility of applying a Markov Chain to human motion on a task level view to predict the next likely task. Lacking detailed positional joint information in their approach, [12] proposed a framework that allows robots and humans to operate safely in close proximity. Using the Gaussian Mixture Model representation of human motion they where able to predict the upcoming work space occupancy for a certain time span. Applying the existing trajectory optimization methods, they showed the practicability using a work space occupancy map to compute collision free trajectories. Providing a framework to offer real time motion recognition and prediction they assessed the frameworks capabilities by conducting experiments to measure the human interference during task execution.

Human motion modeling using artificial neural networks - Application of artificial neural network in human motion modeling and estimation has grown drastically with the introduction of datasets containing annotated human motion and pose samples [7]. Commonly used by [2], [3], [4], [13], they all focused on the task of human motion prediction and long-term synthesis. By pursuing the achievements in human motion prediction using Deep RNN, [13] introduced simple and scalable RNN architecture. Using a single Gated Recurrent Unit (GRU) rather than concatenated Long Term Short Memory cells [3] they showed that the much simpler architecture is capable of achieving similar performance. Simplifying the architecture increases performance and allows faster training. Applying ideas from [6] they where able to improve first frame discontinuities significantly. Rather than modeling human poses and motion, they added a residual connection to the architecture which forced it to model joint velocities instead. Gui et al. [4] showed, most recent improvements in the area of human motion modelling and prediction. They used a similar approach as [13], but introduced a global discriminator which examine the quality of the prediction. Inspired by Generative Adversarial Networks (GAN) they adapted this concept to the motion prediction domain and resulted in a motion GAN. GANs have shown great progress in sequence generation problems since the predicted sequence is judged from a global perspective. Jointly optimizing the discriminator together with the predictor, the predictor's performance is not only measured by the loss function, but further the predicted sequence is rated quality-wise from a global perspective.

Although state of the art showed promising results by applying RNNs in the field of human motion prediction, RNNs have not been extensively exploited in the field of human-robot collaboration. Due to the lack of available datasets containing actions similar to assembly tasks in an industrial environment, we conducted experiments using our own data samples.

III. METHOD

Recent work in human pose prediction mainly focuses on casual tasks, e.g walking, discussions, eating and etc. It showed promising results for motion prediction in a short-term manner and qualitative smooth motion synthesis in a long-term scenario. However, for the distinct scenario of motion prediction in an industrial environment, we did not find publicly available datasets which can be used for human-robot collaboration. Further more, we saw the potential to improve predictions on our own dataset by

using a more complex back projection layer in the network architecture.

Our approach - As Figure 1 shows, we apply a similar structure as [13], who based their proposed method on the advantage of modelling velocities in the RNN [6]. We focused on the importance of high accuracy in the first frames to accommodate for continuous prediction. Applying a RNN in a sequence prediction task requires the output to have the same dimensionality as the input. This means in our case the input is given by a sequence of observed body poses in the joint angle domain (green skeletons) and the output is a continued sequence of the input (blue skeletons). The input size represents the number of joint angles we consider for reconstruction of the body skeleton, see Figure 3. The output size is dependent on the number of units in the GRU. Using a higher number of internal units allow to extract more features from the input, although it mismatches the desired output-size, e.g. using 1024 outputs instead of 15 output angles. To apply RNN to sequence prediction problems, the output size is projected from the GRUs output units to the original dimensionality. Commonly this is done by a linear layer which back-projects frame by frame. Temporal information is only contained in the RNN and the linear back-projection is only used for size adaptation. We saw potential in expanding the back-projection layer to incorporate temporal information. This was done by feeding all GRU output information to a fully connected layer. Dependent on the sequence, the fully connected layer is able to combine all available information and fuse this to the output sequence. All information fusing in one layer (Linear Combination and Backprojection) allows this layer to weight the extracted information from the GRU in a second instance. Applying a residual connection forces the overall structure to model joint velocities rather than poses. Due to the changes in the network structure, the same residual architecture as used by [13] was no longer possible and is changed to use the last pose of the input sequence as input for the residual connection (dotted line).

Setup - We conducted our experiments with our own recorded dataset to evaluate the applicability of the proposed method in a human-robot collaboration scenario. A simple rectangular object with one screw in each corner was used to simulate the assembly procedure. Each screw was tightened by reaching for the screwdriver, tightening and placing the screwdriver again in the rest position, see Figure 2. The recorded dataset contains joint positions in $[x, y, z]$ format for 9 joints on the upper human body. The recorded joint positions resemble a simplified human torso skeleton, see Figure 3. A sequence of 120 seconds of motion capture footage, using OptiTrack motion capture system [17], was recorded at 120 Hz.
The recorded data was pre-processed applying error removal and smoothing. After this, the pre-processed raw data was fed to a skeleton generator to convert joint positions to the joint angle space. Our proposed architecture was trained on the converted dataset.

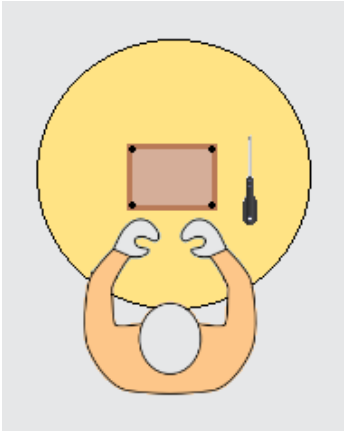

Fig. 2. Schematic representation of the setup.

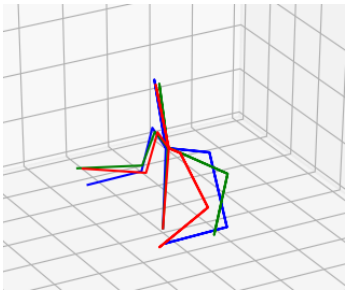

Fig. 3. Computed torso skeleton from joint angle space.

Implementation Details The results presented in this paper where obtained using the proposed method for an input sequence of 120 frames followed by 120 predicted frames. Experiments with the structure showed that 512 units in the GRU are sufficient for our requirements. The recurrent neural network is implemented using Tensorflow and was trained for 12h on a Nvidia GTX 1050Ti GPU.

IV. RESULTS

Finally we present the achieved results of our proposed method for motion prediction. Due to the small dataset we used, it was not feasible to split the dataset into test and train data. However, the results should be representative enough to evaluate the applicability of the proposed method. We considered one second as input sequence to predict one second ahead. We measured the error as mean angle error as it has been done by [4], [13] to be able to compare it to their results. Direct comparison to their results is not possible since they consider full body motion compared to torso motion in our tests. Figure 5 shows the development of the mean angle error over a time span of one second. It indicates low error for the first frames of prediction and increases as time evolves. Showing a faster error development in the first 200 milliseconds, error develops slower and linear after 200

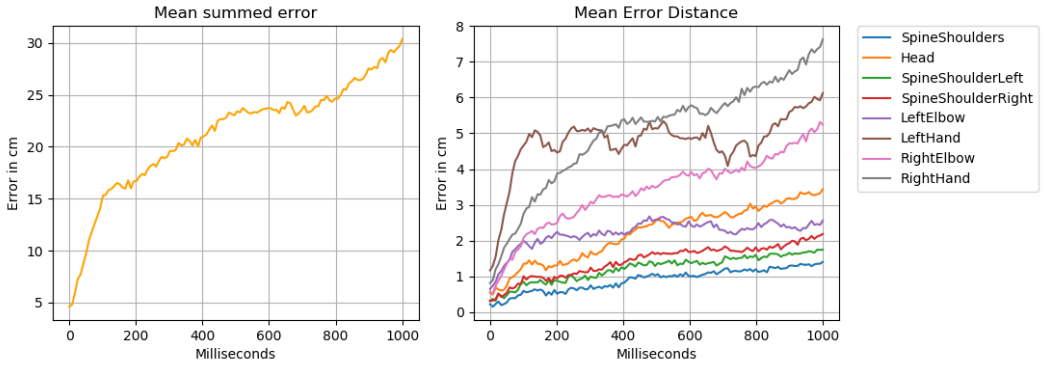

Fig. 4. Absolute errors considering a average human person with 180 cm body height. Left: Shows the summed error of all joints. Right: Shows the error of each joint individually. The results represent the average performance of 60 randomly selected samples.

milliseconds.

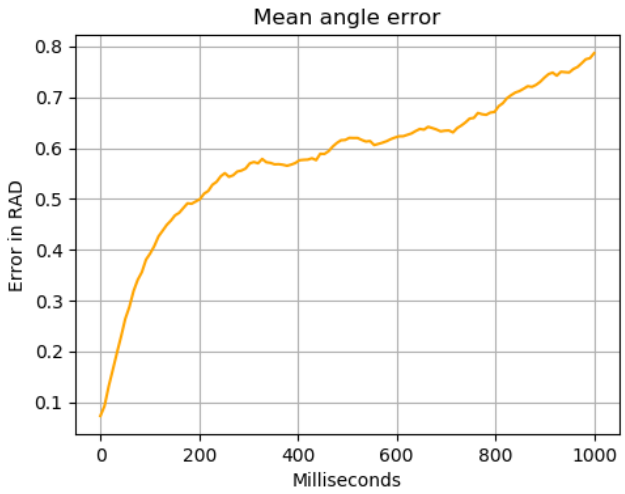

Fig. 5. Mean angle error plot of the proposed method as commonly used by previous work.

Figure 4 was used to asses the achieved performance of the proposed method. The left plots validate the results from Figure 5 by showing a fast progressing error up to 200 milliseconds and slower development after. The right plot reveals highest error for hand joint prediction. Still the maximum error is less than 8 cm at its peak. The higher error for the hand joints can be explained by assuming the highest velocity for hand joints. [3], [4], [13] commonly provide their results in mean angle error, for more easy comparison Figure 5 is provided.

V. CONCLUSIONS

We demonstrated the potential of sequence to sequence artificial neural networks for human robot collaboration scenarios. The obtained results are still in a early stage, but support our vision to use the proposed architecture as scalable human motion prediction solution in human-robot collaborative tasks. However, it allows to assess this technology for usability and shows promising performance. The introduced architecture offers sufficient accuracy with low first frame errors, allowing for motion prediction on continuous data streams. Continuous prediction results can be used to dynamically adjust motion trajectories of robots for collision avoidance. Due to the scarcity of industrial datasets, the next planned steps include to improve the dataset and represent a more general motion model. Additionally, we have an ongoing work to expand our current dataset and split it into test and train sequences. This would allow to verify the presented performance and asses the relevance for industrial integration. Furthermore, for safe industrial integration, it would be valuable to provide a certainty measure [9]. The certainty of a predicted sequence could give a rough estimate of how probable the predicted sequence is. We plan to cover these points in our future work.

ACKNOWLEDGMENT

This work has been supported by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) programme ICT of the Future, managed by the Austrian Research Promotion Agency (FFG), under grant no. 861264.

REFERENCES

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1310, 2017.

[2] K. Fragkiadaki, S. Levine, and J. Malik, "Recurrent network models for kinematic tracking," *CoRR*, vol. abs/1508.00271, 2015. [Online]. Available: http://arxiv.org/abs/1508.00271

[3] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, "Learning human motion models for long-term predictions," *CoRR*, vol. abs/1704.02827, 2017. [Online]. Available: http://arxiv.org/abs/1704.02827

[4] L.-Y. Gui, K. Zhang, Y. Wang, X. Liang, J. M. F. Moura, and M. Veloso, "Teaching robots to predict human motion," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2018.

[5] S. Haddadin, A. Albu-Scheffer, and G. Hirzinger, "Requirements for safe robots: measurements, analysis and new insights," *The International Journal of Robotics Research The International Journal of Robotics Research*, vol. 28, pp. 11–12, 01 2009.

[6] J. N. Ingram, K. P. Körding, I. S. Howard, and D. M. Wolpert, "The statistics of natural hand movements," *Exp Brain Res*, vol. 188, no. 2, Jun 2008, 18369608[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18369608

[7] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2014. [Online]. Available: https://dblp.uni-trier.de/db/journals/pami/pami36.html

[8] ISO, *ISO/TS 15066:2016: Robots and robotic devices – Collaborative robots*. Geneva, Switzerland: International Organization for Standardization, Feb. 2016.

[9] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," *CoRR*, vol. abs/1807.00263, 2018. [Online]. Available: http://arxiv.org/abs/1807.00263

[10] D. Kulic and E. Croft, "Pre-collision safety strategies for human-robot interaction," *Auton. Robots*, vol. 22, pp. 149–164, 01 2007.

[11] H. Liu and L. Wang, "Human motion prediction for human-robot collaboration," *Journal of Manufacturing Systems*, vol. 44, pp. 287–294, 08 2017.

[12] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 299–306.

[13] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," *CoRR*, vol. abs/1705.02445, 2017. [Online]. Available: http://arxiv.org/abs/1705.02445

[14] J. Miseikis, I. Brijacak, S. Yahyanejad, K. Glette, O. J. Elle, and J. Tørresen, "Transfer learning for unseen robot detection and joint estimation on a multi-objective convolutional neural network," *CoRR*, vol. abs/1805.11849, 2018. [Online]. Available: http://arxiv.org/abs/1805.11849

[15] J. Miseikis, P. Knöbelreiter, I. Brijacak, S. Yahyanejad, K. Glette, O. J. Elle, and J. Tørresen, "Robot localisation and 3d position estimation using a free-moving camera and cascaded convolutional neural networks," *CoRR*, vol. abs/1801.02025, 2018. [Online]. Available: http://arxiv.org/abs/1801.02025

[16] S. Mhlbacher-Karrer, M. Brandsttter, D. Schett, and H. Zangl, "Contactless Control of a Kinematically Redundant Serial Manipulator Using Tomographic Sensors," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 562–569, Apr. 2017.

[17] NaturalPoint, Inc., "Optitrack," *[Online]. Available: http://www.naturalpoint.com/optitrack. [Accessed: Feb. 2019].*

[18] S. E. Navarro, S. Koch, and B. Hein, "3d contour following for a cylindrical end-effector using capacitive proximity sensors," in *IROS*. IEEE, 2016, pp. 82–89.

[19] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human-srobot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, 03 2018.