Motion Prediction of Virtual Patterns, Human Hand Motions, and a simplified Hand Manipulation Task with Hierarchical Temporal Memory

Lukas Tenbrink*, Benedikt Feldotto*, Florian Röhrbein*, Alois Knoll*
Contact: lukas.tenbrink@tum.de
*Robotics, Artificial Intelligence and Real-Time Systems,
Faculty of Informatics, Technical University of Munich

Index Terms

Hierarchical Temporal Memory, Motion Prediction, Neural Networks, Hebbian Learning, Human Robot Interaction

Abstract

In this paper we utilize Numenta's Hierarchical Temporal Memory implementation NuPIC for online visual motion pattern prediction and test its performance on virtual animations as well as real world human motion data. For evaluation we run a series of progressively more complex experiments testing specific capabilities: Prediction of fixed-time noise-free motion animations, prediction of protocol-directed tasks with real-world camera captured human motion data, and lastly prediction of repetitive tasks performed without a strict protocol. Results show that the presented setup is able to predict time sequenced images as well as highly variable human motions increasingly well over several iterations. Limits are faced for non sequential variable hand motion execution: Here, predictions are made but do not improve in quality over time. The network runs online in real time and can be transferred to different tasks without expert knowledge. These characteristics qualify the setup for human robot interaction scenarios without the need for verified prediction accuracy.

I. Introduction

Human robot interaction (HRI) scenarios constitute an increasingly growing field of research as robots can facilitate many novel practical applications in daily tasks. In interactive task execution the strengths and advantages of both human and robotic or algorithmic workers can be exploited to increase overall task efficiency. In 2013, Huber et al. [1] empirically demonstrated that for workflows in HRI scenarios, a fully predictive and adaptive assistant can react in the exact timeframes a human counterpart has finished a task and needs robotic input. In fact, robots predicting assistance time frames and handing over objects at the exact time the human required it sped up the overall task execution more than robots assisting in fixed time intervals.

This means that an optimal assistant acts in a *predictive* manner. In highly vision-based tasks such as assembly of machine parts or mechanical reparation, a motion prediction system has to be implemented that makes predictions based on visual data. Predictive algorithms are required to extract the required knowledge such as the co-worker's state, the robot position and any additional manipulable objects as well as obstacles from the visual information to compute an accurate prediction based on the current situation. Specific challenges in the prediction of human motion can be found in highly variable task executions and complex environmental setups which can induce high noise ratios.

Several approaches, both model-based and trained ones utilizing neural networks, have been implemented in an attempt to predict visual motions. However, most of them lack two fundamental characteristics essential for a system dealing with physical world motion prediction: Real-time online learning and high noise robustness. The HTM implementation offers both and includes systems for inherent temporality. The predictive state relation can be learned and re-learned online in a fast manner and multifold given the current sensory input. Hereby, the HTM can handle complex situations, e.g. with changing speeds or repeating frames by fast adapation. Additionally, in the real world, situations change both short-term and long-term. The HTM learns relations online and therefore does not require manual re-adjustment or parametrization with changing environmental conditions.

II. State of the Art

In robotics, visual human detection, recognition and analysis, or "looking at people" [2], has been one of the classic and most popular research fields, since it is essential for dynamic HRI. Required skills are well performed by humans while computers are, historically, very bad at it. We here focus on one subset of this task, namely the prediction of motions, which is essential to understand and adapt to observed environmental motions.

Extremely simple approaches, such as the constant pose predictor [3], can currently often outperform other very complex algorithms in metrics such as short-term 0.4 and 1 second prediction error. This gives researchers some very good baselines to compare state-of-the-art motion prediction to. Recent approaches for visual motion prediction include LSTM-3LR (3 layers of Long Short-Term Memory cells) [4] and ERD (Encoder-Recurrent-Decoder) that is able to produce plausible long-term human motion predictions. In [5] a structural recurrent neural network (RNN) is presented

that introduces expert knowledge by utilizing semantic knowledge about the network as input, and dynamically assigns different RNNs to similar parts of the data. Martinez et al. [3] analyzed the previous [5, 4] methods by looking at the architectures, loss functions, and training procedures. Three adaptations were proposed to improve motion prediction to compete with other state-of-the-art approaches. Blütepage et al. [6] presented a full-body motion prediction system based on RGB + depth data that utilizes a deep generative model for online motion predictions of up to 1660ms.

Recent work involving prediction of hand motions specifically include a system [7] with a convolutional neural network and RNN based model in order to adjust robot motion trajectories to improve human-robot collaboration. Predictive neural networks for human robot collaboration have been demonstrated in [8] for human like hand reaching with Long Short-Term Memory. A HTM has been utilized for proprioceptive biomimetic arm motion prediction and prediction based Classical Conditioning learned from a human partner in [9]. To the best of our knowledge, the predictive HTM implementation has not yet been applied for human-oriented visual motion prediction, which we introduce in this paper.

A. Hierarchical Temporal Memory

The Hierarchical Temporal Memory (HTM) [10] neural network family provides both online learning and inherent temporality and may therefore be well-suited for online motion prediction. Our setup is based on Numenta's open source implementation NuPIC.

During execution, HTM cells make constant predictions about the next timestep, and their synapses to other cells are reinforced or weakened according to variations of the Hebbian Rule of Learning [11]. From these constant next-step predictions, further predictions such as 5-step or 10-step predictions can be extrapolated using a weight matrix. Furthermore, HTM systems use Sparse Distributed Representations (SDRs) for both input and output. Compared to dense bit array encodings, SDRs have a higher noise tolerance and can thus be considered more stable under varying input dynamics [12]. Using a threshold overlap for comparison, SDRs can tolerate extremely high noise values with low false positives. Many systems using digital sensors, such as computer vision, are subjected to high signal to noise ratios and can thus benefit from high noise tolerance.

While pre-computation of raw image data can promise increasing network performance, we here take the raw images as direct input. This enables us to maintain general applicability as well as test the network's capability to perform all logical tasks required for a prediction. Any sophisticated pre-processing would be highly task and environment dependent, which is time consuming in adaptation and not the goal of our current conducted study of a general prediction framework. We therefore only make use of general basic thresholding and filtering techniques. A system setup not containing task-specific solution steps can be considered very versatile, as it can be employed to solve diverse tasks without additional adjustments.

III. Network Setup

For predictions, experiment execution data is processed through the neural network in fixed frame rates. Each step, the current image is captured, the network processes the data (Figure 1), predictions are extracted, and finally values are presented to the user for evaluation.

A. Sensory Input

A binary image with $32 \cdot 32$ pixels ($n = 1024$) serves as the online input to the networks sensory regions. The method of image capture differs from experiment to experiment, and may be pre-computed to be represented in this format, as described in the experiments' sections.

B. Region data processing

In the network (Figure 1), each pixel has one corresponding sensor and one classifier region, which is used for 5-step predictions. When running a step, the camera input is captured and split up into individual pixels, each of which is assigned onto its corresponding sensor region's data source.

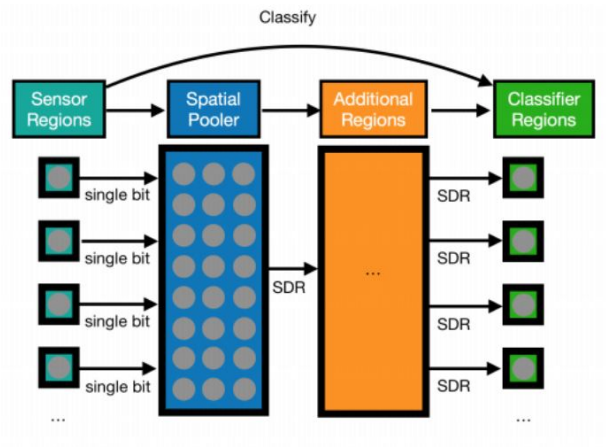

Fig. 1: Neural Network Setup: The network processes the image input in Sensor Regions pooled by a topological Spatial Pooler, additional regions and lastly Classifier regions.

More regions are linked, in sequence, whereas the first region receives input from the sensory regions, while the last region forward-feeds its outputs into the classifier regions. The regions are: A topological Spatial Pooler (SP) region ($n = 20 \cdot 20$), an untopological SP region ($n = 256$), and finally a Temporal Pooler region ($n = 256; cellsPerColumn = 5$).

Additionally, a Scalar Encoder ($w = 31$; $radius = 5$) is used to encode the grey value center x, y. It is inserted to feed-forward into the untopological SP region.

C. Prediction Outputs

For evaluation, we calculate the euclidian distance as a distance score between the binary image matrices of the current input and the predicted image extracted 5 steps earlier. Due to high variance in the accuracy of the motion prediction, a rolling average is calculated. An average distance score of 0 is considered as an optimal prediction.

IV. Virtual Motions

In the first series of experiments, visual motion patterns are generated virtually to allow for simple and fixed interval repetitive prediction tasks and therefore controlled evaluation. During this series, we successfully confirm the capability of the network to handle tasks requiring stateful machines.

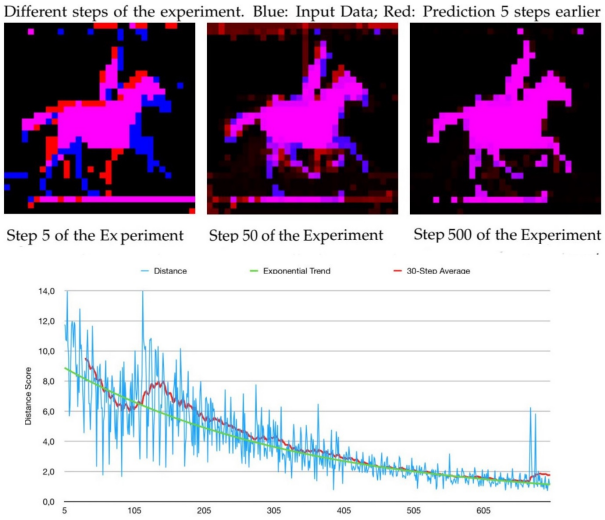

Fig. 2: Prediction of an animated running horse, image pixel input (blue), prediction output (purple) and accuracy over learning time.

In a first experiment motions of an animated running horse are predicted. Frames are extracted from Muybridges "The Horse in Motion" [13] and cycled forwards and backwards, serving as input for the system. The network's prediction accuracy increases exponentially over time and reaches a distance score of about 1 after 700 steps (Figure 2).

V. Uniform Human Hand Motions

In this series of experiments, a human performer is instructed to move according to a specific protocol. The protocol outlines the kind of motion and cycle speed of iterations. This ensures an easily predictable input, though it is not machine guided - no timer or mechanical help is used to ensure the exact action timing. The captured data is thus completely subjected to human motion.

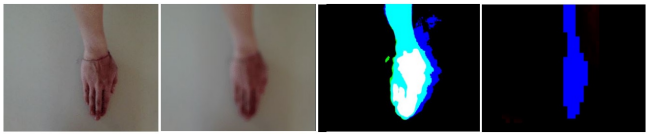

Fig. 3: Example of a captured hand image through the pre-computation steps.

To capture image data, an RGB camera mounted to a stand and pointed down towards a white table. A hand is then moved closely to the table, and the image pre-computed for the hand to finally be represented as white pixels in a black screen space.

To that end, the captured image is blurred with a $32 \cdot 32$ filter, a threshold difference to a background image is computed. The result is finally rescaled and color channels combined to encode the final network input (Figure 3). Images are captured at 5 frames per second.

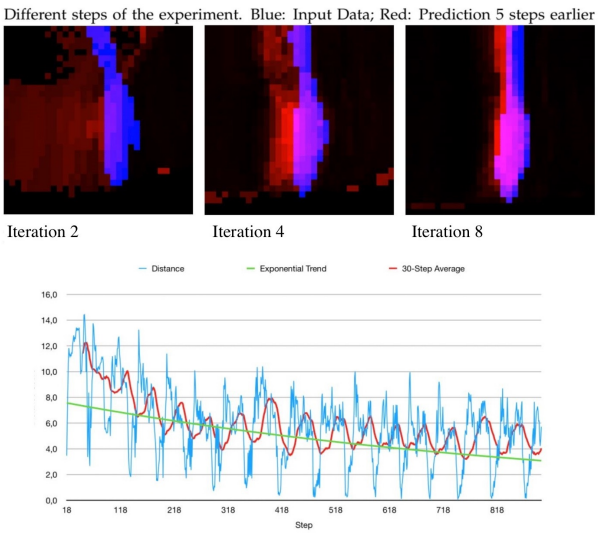

Fig. 4: HTM Prediction accuracy on a human hand in linear motion.

The first human motion experiment analyses linear hand motions. A hand is moved through the frame at a set speed. The network is able to predict the motion quite well and over time reaches an average distance score of about 6 (Figure 4).

In a next step circular hand motions are investigated: The participant is instructed to move his hand in circular motions inside the frame. The Network can predict these motions up to a distance score of about 7 after 300 steps only. High motion overlaps and variances in motion execution reduce prediction accuracy in contrast to the linear hand motions experiment (Figure 5).

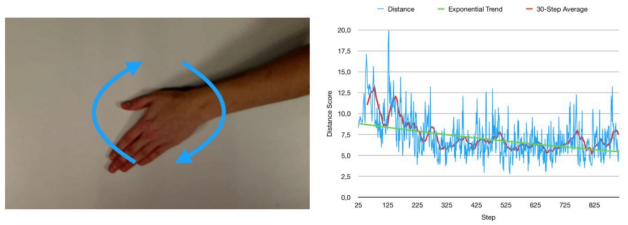

Fig. 5: Setup and prediction error over time of the circular hand motion experiment.

VI. Complex Hand Manipulation

In a next step the human participant is instructed to perform varying tasks without a protocol to follow. Compared to the previous series of experiments hereby the variability of motions increases, which leads to less predictability in the motion patterns. Here we find the capability limits of the presented system.

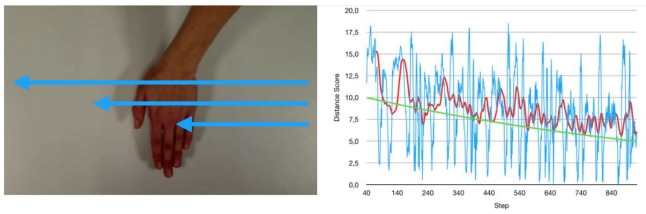

Fig. 6: Setup and prediction error over time of the varying hand speeds experiment.

As a first variation we introduce variation in speed: The participant executes linear hand motions through the frame with varying speeds (Figure 6. It is an adaption of the previous hand motion task (Figure 4) in which the network performed very well. While the network never quite reaches a stable prediction and prediction accuracy oscillates, we still find a a noticeable improvement in prediction accuracy over time.

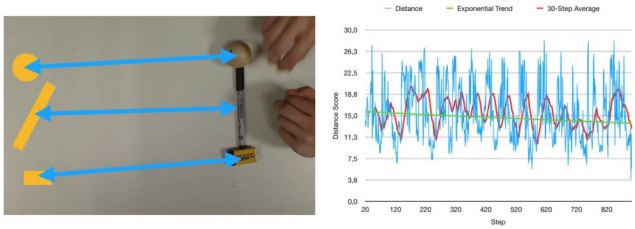

Fig. 7: Setup and prediction error over time of the simplified industrial assembly experiment.

Arguably the most challenging situation for motion prediction is posed by prediction of a simplified industrial assembly task (Figure 7) conducted as a final experiment. Here, three different object types need to be placed on certain spots and brought back afterwards. The task includes high variance in motion execution in terms of speed and location as well as multibody prediction. Over a learning period of about 900 steps including repeated manipulation iterations, the prediction is still oscillating in its accuracy and no strong improvements can be observed.

A. Network Setup Evaluation

Over the course of the experiments, we created several iterations of networks over the 3 experiment series. Network 1 consists of, in addition to the sensory and classifier regions, only a topological SP ($n = 32$), thus lacking a temporal component entirely. Network 2 includes a topological SP ($n = 20 \cdot 20$) and a Temporal Pooler ($n = 400; cellsPerColumn = 6$). Network 3 is the final iteration and thus the one presented in the network setup described above.

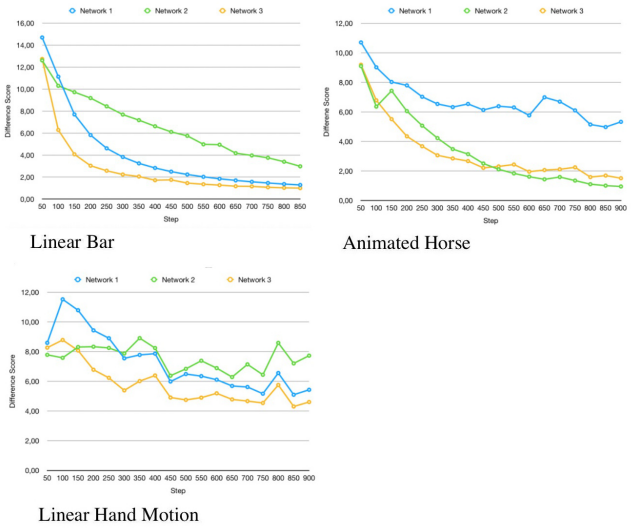

Fig. 8: Influence of Network Architecture on the prediction accuracy: We compare three architectures on the task of a virtual bar, horse animation and linear hand motion prediction. Although there is a trade-off between specialized problem solving and general applicability, we see a clear improvement of network 3 over the others.

To validate whether the network has indeed improved its general applicability in the field of visual motion pattern prediction, we test the networks on the tasks of linear bar, animated horse and linear hand motion prediction. Figure 8 demonstrates the result and indicates improvements in general applicability throughout the network setup iterations.

VII. Conclusion

We presented an experimental setup making use of a Hierarchical Temporal Memory network that demonstrates the capability for visual predicting of highly variable human motion tasks. Real world participant hand motions experiments show that the system presented in this paper is capable of predicting a number of different tasks within the problem space of real-world visual human motion prediction. These tasks include linear motion of a hand over the input space at semi-constant or varying speeds, and

rotating the hand inside the frame without a set pivot point. The system surpasses simple baselines such as the 1-second constant pose predictor [3] in these tasks and confirms Zhang et al.'s [14] and Ugolotti et al.'s [15] findings that systems using HTM are viable approaches to human action observation. We find limits of motion prediction for complex assembly tasks, that introduce high motion and execution variance. The system does not require specific adjustments such as expert knowledge or complex algorithms modifying the input data for reasonable predictions for each task. This proves HTM networks to be well-suited for use in the problem space of visual motion pattern prediction as the primary logical component of a system.

VIII. Acknowledgments

This research has received funding from the European Unions Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

References

[1] Markus Huber et al. "Predictive Mechanisms increase Efficiency in Robot-supported Assemblies: An Experimental Evaluation". In: *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*. Gyengju, Korea, 2013.

[2] D.M Gavrila. "The Visual Analysis of Human Movement: A Survey". In: *Computer Vision and Image Understanding* 73.1 (1999), pp. 82–98. ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1998.0716. URL: http://www.sciencedirect.com/science/article/pii/S1077314298907160.

[3] Julieta Martinez, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks". In: *CoRR* abs/1705.02445 (2017). arXiv: 1705.02445. URL: http://arxiv.org/abs/1705.02445.

[4] Katerina Fragkiadaki, Sergey Levine, and Jitendra Malik. "Recurrent Network Models for Kinematic Tracking". In: *CoRR* abs/1508.00271 (2015). arXiv: 1508.00271. URL: http://arxiv.org/abs/1508.00271.

[5] Ashesh Jain et al. "Structural-RNN: Deep Learning on Spatio-Temporal Graphs". In: *CoRR* abs/1511.05298 (2015). arXiv: 1511.05298. URL: http://arxiv.org/abs/1511.05298.

[6] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. "Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration". In: *CoRR* abs/1702.08212 (2017). arXiv: 1702.08212. URL: http://arxiv.org/abs/1702.08212.

[7] Yiwei Wang et al. "Hand Movement Prediction Based Collision-Free Human-Robot Interaction". English (US). In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*. Vol. 2017-July. United States: IEEE Computer Society, Aug. 2017, pp. 492–493. DOI: 10.1109/CVPRW.2017.72.

[8] Phongtharin Vinayavekhin et al. "Human-Like Hand Reaching by Motion Prediction Using Long Short-Term Memory". In: *Social Robotics*. Ed. by Abderrahmane Kheddar et al. Cham: Springer International Publishing, 2017, pp. 156–166. ISBN: 978-3-319-70022-9.

[9] Benedikt Feldotto et al. "Hebbian learning for online prediction, neural recall and classical conditioning of anthropomimetic robot arm motions". In: *Bioinspiration & Biomimetics* 13.6 (Oct. 2018), p. 066009. DOI: 10.1088/1748-3190/aae1c2. URL: https://doi.org/10.1088%2F1748-3190%2Faae1c2.

[10] Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. Times Books, 2004. ISBN: 0805074562.

[11] Donald Olding Hebb. *The Organizations of Behavior: a Neuropsychological Theory*. Chapman and Hall, 1957.

[12] Subutai Ahmad and Jeff Hawkins. "How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites". In: *CoRR* abs/1601.00720 (2016).

[13] Eadweard Muybridge. *The Horse in Motion*. 1878.

[14] Sen Zhang et al. "Detection of Activities by Wireless Sensors for Daily Life Surveillance: Eating and Drinking". In: *Sensors* 9.3 (2009), pp. 1499–1517. ISSN: 1424-8220. DOI: 10.3390/s90301499. URL: http://www.mdpi.com/1424-8220/9/3/1499.

[15] Roberto Ugolotti et al. "Multi-sensor system for detection and classification of human activities". In: *Journal of Ambient Intelligence and Humanized Computing* 4.1 (Feb. 2013), pp. 27–41. ISSN: 1868-5145. DOI: 10.1007/s12652-011-0065-z. URL: https://doi.org/10.1007/s12652-011-0065-z.