A Recurrent Encoder-Decoder Network Architecture for Task Recognition and Motion Prediction in Human-Robot Collaboration based on Skeletal Data

1st Dianhao Zhang
Centre For Intelligent Autonomous Manufacturing Systems
Queen's University Belfast
Belfast,UK
dzhang07@qub.ac.uk

2nd Ngo Anh Vien
The Institute of Electronics, Communications and Information Technology
Queen's University Belfast
Belfast,UK
v.ngo@qub.ac.uk

3rd Seán McLoone
Centre For Intelligent Autonomous Manufacturing Systems
Queen's University Belfast
Belfast,UK
s.mcloone@qub.ac.uk

***Abstract*—To achieve more accurate early prediction of human motion and enable robots to respond in a safe manner, a human-robot collaboration (HRC) architecture is proposed to predict future human activities and their trajectories using skeletal data collected from a Kinect. The architecture is the combination of two models. The first model is designed to predict motion trajectories and is based on a recurrent Encoder-Decoder long-short-term-memory (LSTM) network that takes a historical trajectory as input and predicts a future trajectory as its output. The second model predicts the next activity to be performed using a combined LSTM and conditional random field network (LSTM-CRF). Preliminary results are presented showing the efficacy of the approach with the LSTM-CRF able to achieve high-quality human activity classification, and the encoder-decoder LSTM able to accurately predict the coordinates of future human motion trajectories.**

***Index Terms*—human-robot collaboration, skeletal data, sequence-to-sequence, LSTM**

I. Introduction

Efficiency and safety are among the most important considerations in manufacturing automation. With collaboration between humans, aided by their ability to anticipate each other's actions, high-level tasks with multiple sub-activities can easily be handled to satisfy these criteria. In contrast, in human-robot collaboration (HRC) efficiency and productivity are frequently negatively impacted by safety related conservative robot operation, as well the computational overhead of dynamically computing actions sequences. It is therefore desirable to develop robots that can anticipate human actions in order to facilitate safe and effective collaboration. To achieve this target, an integrated HRC architecture is proposed consisting of real-time human dynamic motion tracking, human motion recognition and human trajectory prediction modules. The overall architecture of the proposed system is summarised in Fig 1. The input to the system is a sequence of subtasks expressed as trajectories of human skeletal joint coordinates, with each subtask corresponding to one action. These are processed by a long-short-term-memory Encoder-Decoder neural network (LSTM-ED) to give a prediction of the human joint coordinates for the next subtask. This information on human pose is then processed by a LSTM Conditional Random Fields (CRF) model to generate the label of the future subtask. The predicted human motion trajectory and subtask label are then passed to the robot motion planning and control algorithms. These algorithms are currently under development. The contribution of this paper is to introduce the architecture for human motion and activity prediction for HRC, and to present some preliminary results demonstrating its efficacy.

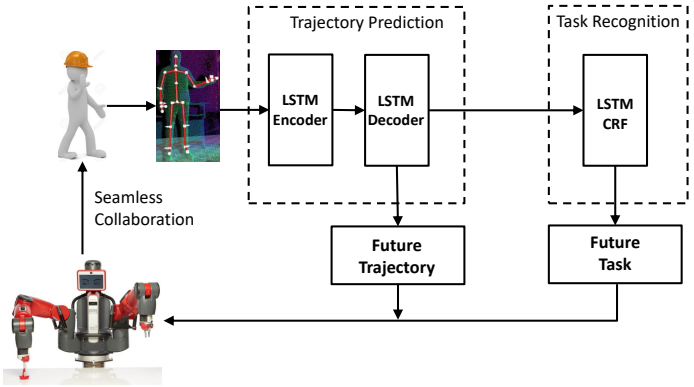

Fig. 1: Human robot collaboration architecture.

II. Network Typologies

A. LSTM-ED Network for motion prediction

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture that was designed by Hochreiter and Schmidhuber [2] to address the vanishing and exploding gradient problems of conventional RNNs, and are ideally suited to processing sequential (time-series) data. In our work we employ a LSTM-ED architecture to implement a motion prediction model. The architecture, which has previously been successfully applied to video prediction and intelligent translation applications [4], [6], is a many-to-many LSTM implementation consisting of a multi-layer encoder and a multi-layer

decoder (Fig 2). The encoder computes a representation s for each input sequence that represents a past motion sequence. Based on that input representation, the decoder generates an output sequence that represents a sequence of future motion. The input is a sequence of coordinates with 50 values (25 joints $\times$ 2D coordinates) from a set of 90 frames (3 seconds) of human skeleton reference points, as captured by a Kinect. The attention vector is the sum of hidden states of the encoder, weighted by attention scores. The input to the decoder is the concatenation of the previous hidden state and the attention vector. The first prediction is used as the input to the next LSTM cell.

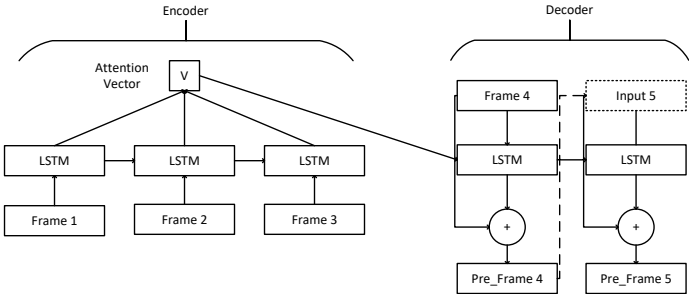

Fig. 2: LSTM sequence-to-sequence model architecture

B. *LSTM-CRF for activity classification*

Conditional random fields (CRF) are discriminative models for sequence labeling, which have been shown to be a powerful model for sequence tagging problems [3]. The motivation for introducing this model in combination with an LSTM, as first suggested by [1], is that it enables both previous and future context to be considered, and is therefore more suited to representing the diversity and time-varying nature of human activity. An LSTM-CRF network is obtained by employing the hidden states of the an LSTM as the inputs to a CRF layer, where the role of the CRF is to learn a mapping from the hidden state values to the subtask labels [5].

III. Results

We evaluate the performance of the proposed LSTM-ED and LSTM-CRF models for a basic activity involving the use of a screwdriver. The activity is partitioned into three sub-activities; sitting-down, bending over to pick up the screwdriver, and using the screwdriver to tighten a screw (simulated action). As the frame rate of the Kinect is 30 frames per second, one sample of the full screwdriver task consists of 180 frames (2 seconds per sub-activity). A dataset consisting of 120 repetitions of this task was recorded and used as training and test data for the models. Twenty percent of the data was retained as test data. Models were implemented in Python and trained using PyTorch.

First, we evaluate the LSTM-CRF by reporting its activity classification accuracy. Second, we evaluate the LSTM-ED by comparing the difference between the predicted future coordinates and the ground-truth in terms of the mean square error. In addition, we plot the predicted human pose for 90 future frames based on the previous 90 frames to check if the behaviour corresponding to the sub-activity can be predicted accurately. The results are shown in Fig 3.

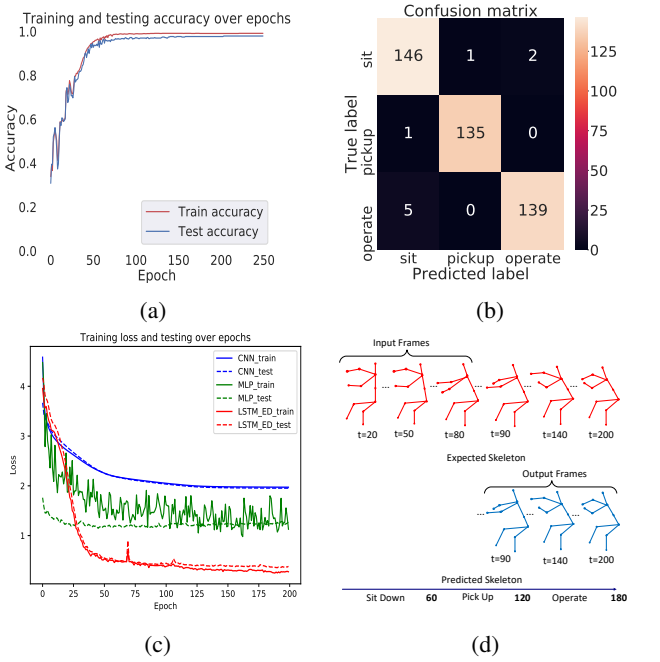

Fig. 3: Experimental results: (a) the LSTM-CRF model learning curves; (b) The test data confusion matrix for sub-activity recognition in the screwdriver usage task; (c) MSE comparison between MLP, CNN and LSTM-ED models for motion prediction; (d) Selected expected and predicted skeletal data frames for one instance of the screwdriver task using the LSTM-ED model.

For the screwdriver usage activity the LSTM-CRF model is able to achieve an accuracy of 99.24% and 98.76% on the training and test datasets, respectively, and for motion prediction the LSTM-ED substantially outperforms MLP and CNN based alternatives, which were also trained and tested for this activity.

IV. Conclusion

In this paper, we propose an architecture for providing robots with the capacity to anticipate human actions as an enabler for more effective human-robot collaboration. The architecture involves the use of skeleton-based human tracking data and LSTM-CRF and LSTM-ED based recurrent neural network models to perform human activity classification and human motion prediction, respectively. Preliminary results demonstrate the potential of the approach for predicting human activity subtasks ahead of time, offering the possibility of designing robot path planning and motion control algorithms that are more responsive and attuned to human interaction.

In future work, bespoke path planning and motion control algorithms will be developed and integrated with the proposed architecture and the overall system evaluated on more complex human-robot collaboration scenarios.

REFERENCES

[1] Tuan Do and James Pustejovsky. Fine-grained event learning of human-object interaction with LSTM-CRF. *arXiv preprint arXiv:1710.00262*, 2017.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[3] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv:1508.01991 [cs]*, August 2015.

[4] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025 [cs]*, September 2015.

[5] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Conditional Random Fields for Object Recognition. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1097–1104. MIT Press, 2005.

[6] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *International conference on machine learning*, pages 843–852, 2015.