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

Dianhao Zhang, Vien Ngo, Seán McLoone, Queen's University Belfast

Introduction

- 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.
- Desirable to develop robots that can anticipate human actions in order to facilitate safe and effective collaboration.
- An integrated HRC architecture is proposed consisting of real-time human dynamic motion tracking, human motion recognition and human trajectory prediction modules (Fig .1).

How do the architecture works?

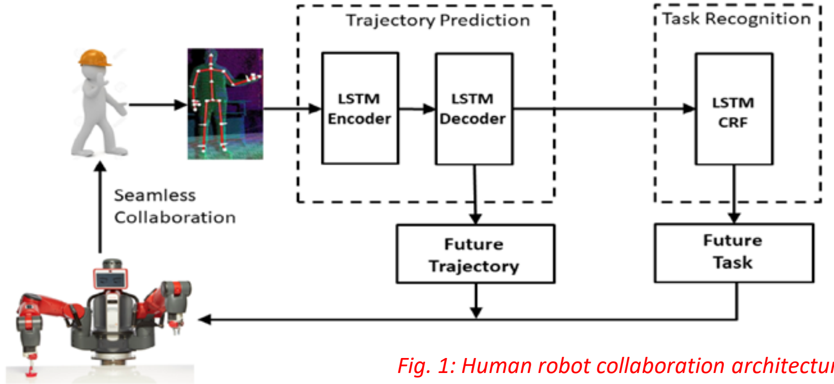

Fig. 1: Human robot collaboration architecture

- Input to the system: a sequence of subtasks expressed as trajectories of human skeletal joint coordinates, with each subtask corresponding to one action.
- A long-short-term-memory Encoder-Decoder neural network (LSTM-ED) predicts the human joint coordinates for the next subtask.
- Human pose information is processed by a LSTM Conditional Random Fields (CRF) model to generate the label of the future subtask.
- Predicted human motion trajectory and subtask label then used to augment the robot motion planning and control algorithms.

Network Typologies

LSTM-ED Network for motion prediction

- LSTM-ED architecture: a many-to-many LSTM implementation consisting of a multi-layer encoder and a multi-layer decoder.
- The encoder computes a representation for each input sequence that represents a past motion sequence
- The decoder generates an output sequence that represents a future motion sequence.
- 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.

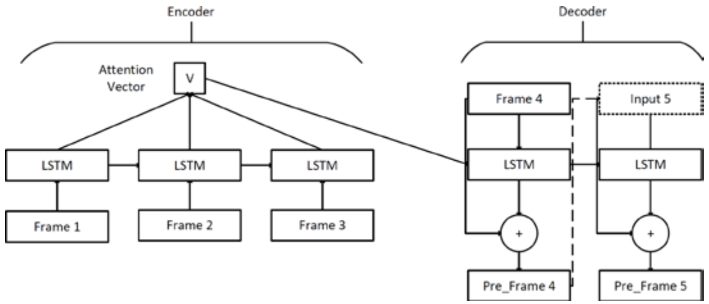

Fig. 2: LSTM-ED architecture

LSTM-CRF for activity classification

- A Conditional random fields (CRF) network is a discriminative models for sequence labeling
- In combination with an LSTM it enables both previous and future context to be considered.
- An LSTM-CRF network is obtained by employing the hidden states of an LSTM as the inputs to a CRF layer.
- The role of the CRF is to learn a mapping from the hidden state values to the subtask labels.

Results

Experimental setup

- Preliminary results obtained for a basic activity involving a screwdriver partitioned into three sub-activities; sitting-down, bending over to pick up the screw-driver, and using the screwdriver to tighten a screw.
- One sample of the full screwdriver task consists of 180 frames of Kinect skeletal data (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.

Preliminary results

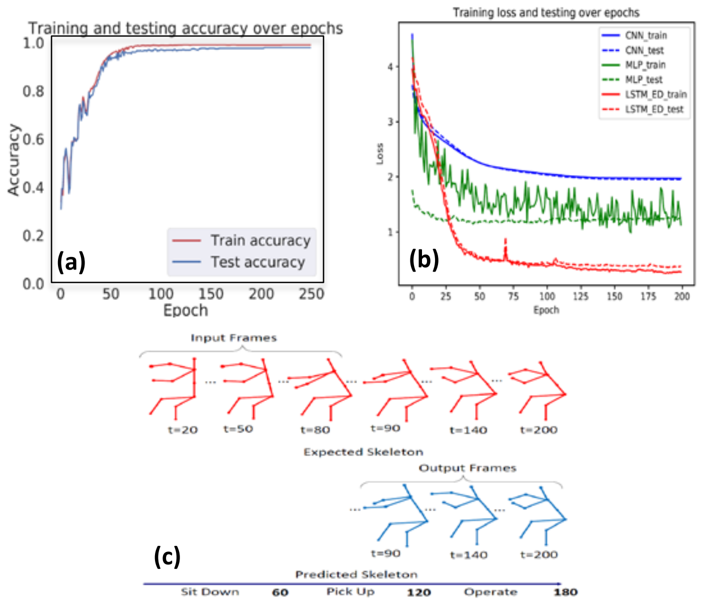

Fig. 3: (a) LSTM-CRF model learning curves; (b) MSE comparison between MLP, CNN and LSTM-ED models; (c) Expected and predicted skeletal data frames using the LSTM-ED model.

Queen's University Belfast
School of Electronics, Electrical Engineering and Computer Science

Email: {dzhang07, s.mcloone}@qub.ac.uk

iAMS Centre for Intelligent Autonomous Manufacturing Systems