Surgical Gesture and Error Recognition with Time Delay Neural Network on Kinematic Data

Giovanni Menegozzo, Diego Dall'Alba, Chiara Zandona and Paolo Fiorini

University of Verona, Italy
paolo.fiorini@univr.it

INTRODUCTION

Surgical Robotic System (SRS) supports the surgeon in executing complex surgical actions with more confidence and firmness, which ultimately translates into better outcomes for the patients. Automatic surgical gesture recognition and classification is an enabling technology for improving surgeon's support in multiple applications. On-line recognition of surgical gestures will improve the training process and the overall performance thanks to user-specific feedbacks and the prompt detection of errors due to excessive fatigue or cognitive overloading [1]. Thus, surgical gesture recognition recently gained more attention from robotic and computer vision researchers, thank also by the availability of syncronized video and kinematic data acquired by reseach SRSs and surgical simulators. The introduction of new datasets and methods support generalization of results to several procedure. Zappella et al. was among the first to propose gesture recognition from video data using linear dynamical systems and bag of words [2]. Kinematic data have been firstly modeled with Hidden Markov Model (HMM) [3], then other approaches were proposed with Condition Random Field (CRF) and skip-chain CRF (SKCRF) [4]. Recent research works combine kinematics and video data to outperform gesture recognition results previously obtained [5]. However, integration of kinematic and video data is challenging due the etherogeneous nature of these data types and further research efforts is required. Actually, many research focuses on modeling the temporal information that is essential for gestures recognition. New types of neural networks have been proposed to handle videos and kinematics information, exploiting intrinsic temporal constraint as a new feature and considering multiple time steps simultaneously for a more precise context evaluation, e.g. Temporal Convolutional Networks (TCN) or Long Short Term Memory (LSTM) [6,2]. Time Delay Neural Networks (TDNN) are becoming successful due to their ability to model time relationships over long sequences while maintaining computational efficiency [7]. TDNN has a pyramidal structure with a progressively wider temporal context, i.e. the initial transforms are learnt on narrow contexts and the deeper layers process the hidden activations from a wider temporal context due to the dilatation on nodes, as shown in Figure 1.

MATERIALS AND METHODS

In Figure 1 we represent the architecture of the proposed TDNN that is composed of two blocks (layer 1 and 2 in figure) and a time distributed activation level for classification (layer 3). Each block consists of a mono-dimensional convolutional layer (with kernel filter size of 3) and a Rectified Linear Unit (ReLU) activation function. Thanks to the proposed architecture, the convolutional layer in the second block will have a dilated kernel for increasing the reception field.

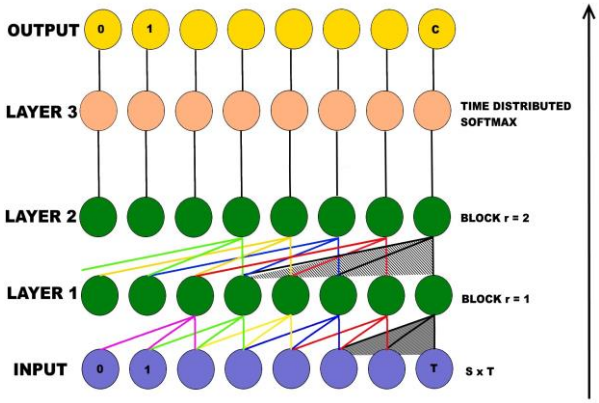

Figure. 1. *Architecture of the proposed Time Delay Neural Network (TDNN). The dark shadow shows the context pyramidal structure of the TDNN*

To evaluate the performance of the proposed method in error detection we have created the V-RASTED (Virtual Robotic Assisted Surgical Training Evaluation Dataset) dataset by collecting training exercises performed by a group of 18 medical students taking an elective course in robotic surgery. We set up an experiment to acquired synchronized kinematic data and endoscopic stereo images from the simulator developed by BBZ s.r.l. The exercise consists in lifting a ring with one robotic arm, passing the ring on the other arm and positioning it in the corresponding pole (Figure 2). This exercise replicates the one offered in standard MIS training curriculums and available in all SRS virtual training simulators.

Each student had a time slot of one hour, the first half is dedicated to practicing with the simulator interface and the second half is dedicated to the recorded trials. Each student performed from a minimum of ten to a maximum of twenty trials resulting in a total 276 sequences. The dataset included synchronized stereo images, kinematic variables for each slave manipulators simulated and some status variable for recording foot-pedals activities. To maintain consistency with JIGSAWS [5], for each slave manipulator 13 variables are provided: Cartesian position (3), rotation matrix (9), and instrument gripping angle (1). The classification classes are six and are defined as ring collection, ring transfer between surgical tools, ring

positioning in matching pole and their corresponding errors.

The dataset and more detailed documentation are available at *gitlab.com/altairLab/v-rasted.git.* The proposed TDNN implementation is based on Keras, with GPU accelerated Tensorflow framework, and it has been tested on mobile workstation equipped with Intel i7 CPU, 8GB RAM and a Nvidia Geforce 960m video card with 2GB video memory. We used ADAM stochastic optimization algorithm with categorical cross entropy loss function during training with 200 epochs and batch size of 16.

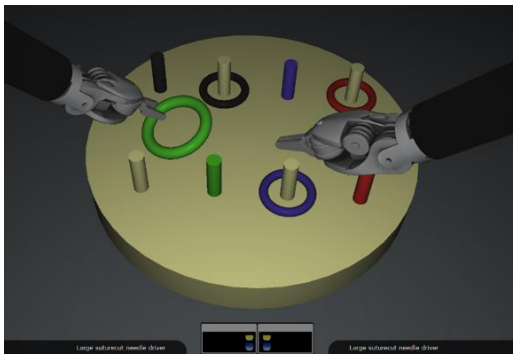

Figure. 2. *Example of image extracted from V-RASTED dataset showing the training exercise considered.*

RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed method we followed the protocol described in [5] using only micro accuracy and precision. For detailed description of the experimental evaluation please refer to [5]. We obtained 68.4±6.95% on micro accuracy and 69.75±12.07 on precision evaluation. The values presented for each measure correspond to the average values (± standard deviation) of all users tested with Leave One User Out (LOUO) methodology. This means that for each user, the training and the testing set are recalculated. In Figure 3 we report the normalized confusion matrix resulted from our evaluation, it shows which types of classes are correctly recognized and which classes are the most complex to predict. In general, the results obtained on V-RASTED correspond to a good level of gesture recognition, aligned with results obtained on other datasets [2,6]. We tested our network with benchmark dataset [8]. We Most of the wrong recognitions occur between gesture and corresponding errors classes. This fact is clearly represented by the normalized confusion matrix showed in Figure 3. These results demonstrate the limits of error detection based only on kinematic variables, which can accurately represent robotic system movements, but they are unsuitable for recognizing errors deriving from interaction with the environment, such as the fall of a ring during grabbing or the incorrect positioning of the ring in the corresponding pole. To improve this result, in future works we will introduce environment sensing, based on video data as spatiotemporal features or on external sensors as marker. Moreover, the proposed V-RASTED dataset is more challenging than previously proposed ones (e.g. JIGSAWS) since its present strong class imbalance that is typical of fault/error detection. This public dataset would be an important component in the development and benchmarking of future surgical error detection methods.

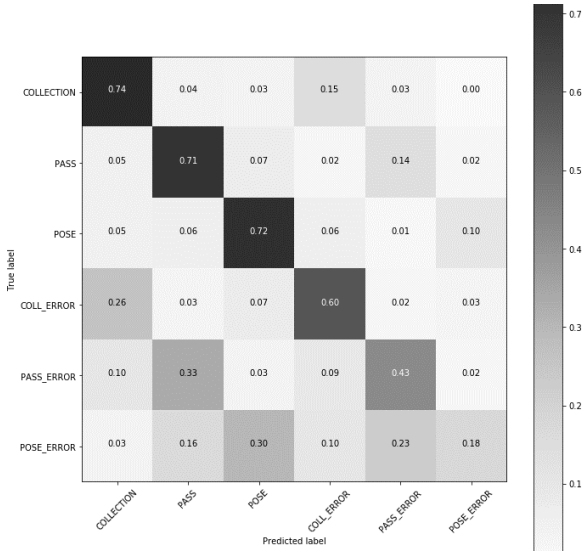

Figure. 3. *Normalized Confusion matrix obtained for the proposed method during LOUO evaluation on V- RASTED*

CONCLUSIONS

In this work we have introduced TDNN for surgical gestures and errors recognition based on kinematic data. We have evaluated the proposed method on novel V-RASTED public dataset. The obtained results demonstrate state that TDNN applied to kinematic data can be very effective in modeling instrument movements, thus they are suitable for gesture recognition, but they need to be fused with other sensing modalities for obtaining error detection performance suitable for on-line situation awareness systems.

ACKNOWLEDGEMENT

This project has received funding from ERC under the H2020 R&I programme (grant agreement No 742671).

REFERENCES

[1] Lee, G. I., et al. "Surgeons' physical discomfort and symptoms during robotic surgery: a comprehensive ergonomic survey study." Surgical endoscopy 31.4 (2017)

[2] Zappella, Luca, et al. "*Surgical gesture classification from video and kinematic data*." Med. image analysis 17.7 (2013)

[3] Rosen, Jacob, et al. "*Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills*."

[4] Tao, Lingling, et al. "*Surgical gesture segmentation and recognition*."

[5] Ahmidi, Narges, et al. "*A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery*."

[6] DiPietro, Robert, et al. "*Analyzing and Exploiting NARX Recurrent Neural Networks for Long-Term Dependencies*."

[7] Peddinti, Vijayaditya et al. "*A time delay neural network architecture for efficient modeling of long temporal contexts*."

[8] Menegozzo, Giovanni et al. "*Surgical Gesture Recognition with Time Delay Neural Network based on kinematic data*."