Surgical gesture recognition with time delay neural network based on kinematic data

Giovanni Menegozzo*, Diego Dall'Alba*, Chiara Zandonà*, Paolo Fiorini**Life Fellow, IEEE*

**Department of Computer Science*, University of Verona, Verona, Italy

Corresponding Author: giovanni.menegozzo@univr.it

***Abstract*—Automatic gesture recognition during surgical procedures is an enabling technology for improving advanced assistance features in surgical robotic systems (SRSs). Examples of such advanced features are user-specific feedback during execution of complex actions, prompt detection of safety-critical situations and autonomous execution of procedure sub-steps. Video data are available for all minimally invasive surgical procedures, but SRS could also provide accurate movements measurements based on kinematic data. Kinematic data provide low dimensional features for gesture recognition that would enable on-line processing during data acquisition. Therefore, we propose a Time Delay Neural Network (TDNN) applied to kinematic data for introducing temporal modelling in gesture recognition. We evaluate accuracy and precision of the proposed method on public benchmark dataset for surgical gesture recognition (JIGSAWS). To evaluate the generalization capability of the proposed method, we acquired a new dataset introducing a different training exercise executed in virtual environment. The dataset is publicly available to enable other methods to be tested on it. The obtained results are comparable with other methods available in literature keeping also computational performance compatible with on-line processing during surgical procedure. The proposed method and the novel dataset are key-components in the development of future autonomous SRSs with advanced situation awareness capabilities.**

***Index Terms*—Time Delay Neural Network, TDNN, surgical gesture segmentation, surgical action/gesture recognition, kinematic modelling**

I. Introduction

The introduction of robotic systems in minimally invasive surgery (MIS) has provided many advantages to the surgeon compared to other surgical approaches, especially magnified tri-dimensional visualization and improved instruments dexterity [1]. 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. In the near future, these SRSs will provide even more advanced features, such as multi-modal intra-operative sensing modalities fully integrated in the vision system, novel instrumentations and improved cognitive and technical support in the surgical procedure execution [2]. This substantial impacts in the healthcare system and on the society wellbeing [3].

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 742671).

978-1-5386-7825-1/19/$31.00 ©2019 IEEE

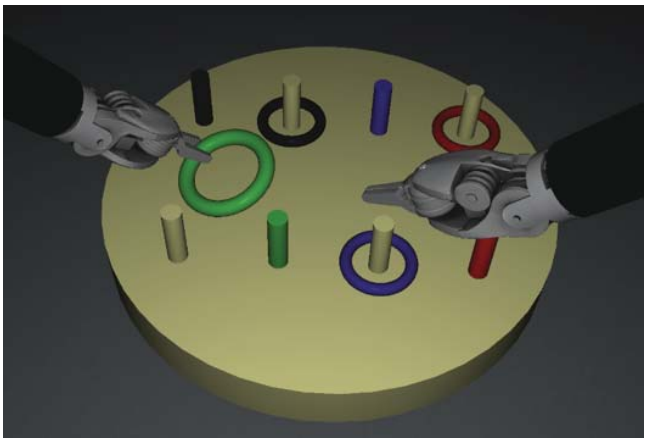

Fig. 1. Example of the virtual task considered in the V-RASTED dataset: 4 colored rings need to be placed in the corresponding pegs. See text for more details.

Automatic surgical gesture recognition and classification is of paramount importance to support this tendency since it could be applied for multiple purposes. Automatic surgical workflow analysis would enable forecasting of possible dangerous situations and to suggest corrective actions before their critical consequences [4]. Evaluating operator performance on-line during the procedure will improve the training process and the overall performance thanks to surgeon-specific feedbacks and the possibility of promptly detecting excessive fatigue and cognitive overloading [5]. Many works related to automatic subdivision of surgical operations into sub-tasks (i.e. gesture, maneuvers, phases) are based on video data, since this type of information is available for every type of MIS procedure. Kinematic data acquired from robotic manipulators could provide accurate measurements for gesture recognition since these data directly correlate to surgeon commands and therefore instruments movements. In the past decade, gaining access to kinematic data was complex due to concerns about patients' privacy and SRS manufacturer's trade secrets. The availability of SRS research platforms, such as da Vinci Research Kit (DVRK) or Raven II, virtual training simulators and the intro-

duction of novel SRSs on the market is changing this situation and enabling an easier acquisition of dataset including video and kinematic recordings [6]. The possibility of acquiring data from virtual simulator during training exercises (or simulated surgical procedures) will further facilitate data collection of larger dataset.

Therefore, the first novel contribution in this work is a gesture classification method based on the analysis of kinematic data with time delay neural networks (TDNN). We selected TDNN because they are suited for contextual analysis and recognition of temporal patterns over long sequences while maintaining computational efficiency. We have experimentally evaluated the proposed method on standard benchmarking JIGSAWS dataset and on a new annotated dataset acquired from a virtual training simulator to evaluate its generalization performance. This public dataset, detailed described in section IV.B, is the second novel contribution described in this work. The results confirm that the proposed method obtains gesture recognition performance aligned with methods previously proposed on JIGSAWS dataset. Thanks to the novel dataset introduced we have also evaluated the generalization performance of the proposed method, demonstrating its capability of processing different datasets.

The paper is organized as follows: in the next section we contextualize the proposed methods with other work available in literature, then we describe the proposed method in detail. In section IV we describe the experimental setup including the novel dataset acquired then the obtained results and related discussions. In the last part of the paper we draw some conclusions and possible future works.

II. Background and Related works

Surgical gesture recognition recently gained more attention from researchers, becoming a meeting point between robotic and computer vision communities. Increased research activities with SRSs and surgical simulators eased this providing concurrently video and kinematic data allowing to integrate techniques related to different areas. Depending on the SRS used, different sensor information can be provided, such as instrument Cartesian positions, gripping angle and other commands status (e.g. clutching usage). Regardless of the SRSs considered, the dimensionality of kinematic feature is inferior to the ones extracted from video, and they do not require any pre-processing, as demonstrated in [7].

Zappella et al. was among the first to propose gesture recognition from video data using linear dynamical systems (LDSs) and bag of words (BoW) [8]. Nowadays, neural networks are predominant processing methods for image analysis and have also been widely accepted in the field of automatic gesture recognition. Kinematic data have been firstly modeled with Hidden Markov Model (HMM) [9], then other approaches were proposed with Condition Random Field (CRF) and skip-chain CRF (SKCRF) [10], [11]. SKCRF classification results are influenced by kinematic values selected or on their representation (i.e. relative distances and velocities between objects and instruments), as demonstrated in [12]. Kinematic data also provide high level information suitable for the recognition of gestures based on unsupervised methods [13]. Recent research works combine kinematic and video data to outperform gesture recognition results previously obtained [12].

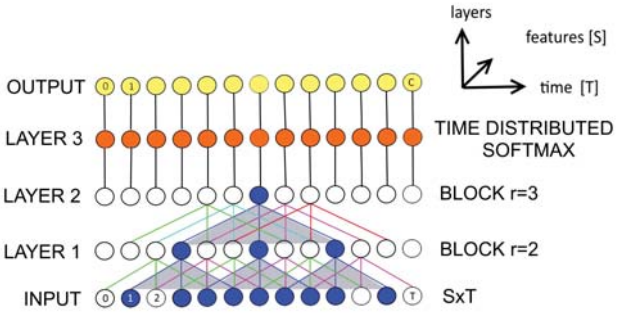

Fig. 2. Architecture of the Time Delay Neural Network (TDNN). The gray shading shows the context pyramidal structure of the TDNN.

New types of neural networks have been proposed to handle video and kinematic information exploiting intrinsic temporal constraint as a new feature and considering multiple video frames simultaneously for a more precise context evaluation, e.g. Temporal Convolutional Networks (TCN) or Long Short Term Memory (LSTM) [7], [8]. These methods provide accurate and precise classification results, but their processing performance decreases due to increased configuration parameters and training times. In general, TCNs applied to spatio-temporal features are widely used in action detection, statistical language model and speech recognition [14], [15]. TDNNs are becoming successful due to their ability to model time over long sequences while maintaining computational lightness, both during training and prediction [16]. Several time-series models, including TDNN and TCN, have been applied and compared for action segmentation and detection in video classification in [17] demonstrating their superior performance on this task. TDNN consists in enlarging the receptive fields of network neurons, i.e. increasing the width of the time window considered for classification. TDNN has been introduced in [18] to create a network for speech recognition that has the ability to represent time relationship between events, that is invariant under time-translation with minimal number of weights compared to the dimensionality of training data. These requirements are very similar to those required in the field of recognition and surgical actions. In fact, to recognize surgical gestures based on kinematic data we need to consider their context in term of temporal sequence of kinematic recordings. Therefore, we decided to use a TDNN model that instead of working on previous words in a sentence, it works on kinematic data sequences for gesture recognition.

III. Method

TDNN has a pyramidal structure due to a wider temporal context, 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 2. The higher layers can learn wider temporal relationships, thus providing a higher feature abstraction. This enables the recognition of longer time features thanks to wider receptive field, thus modelling gestures and movements considering their temporal context. During back-propagation, due to tying, lower layers of the network are updated by a gradient accumulated over all the time steps of the input temporal context. Thus, the lower layers of the network are forced to learn translation invariant feature transforms [16].

The inputs of our neural network will be a set of kinematic data described by Cartesian tri-dimensional position $\mathbf{C}$, orientation represented as quaternions $\mathbf{Q}$ and instrument gripping angle $\mathbf{G}$. We selected this reduced set of kinematic variables since they are available for all robotic systems, in particular for all datasets considered in this work. We grouped all the kinematic features considered in the $\mathbf{S}$ vector, i.e. $\mathbf{S} =< \mathbf{C}, \mathbf{Q}, \mathbf{G} >$. Let $\mathbf{S_t}$ be the input for time steps t for $1 \leq t \leq T$ where T depend on each trial length. The action label for each time step is given by vector $\mathbf{Y_t} \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{Class}}$ such that the true class is 1 and all others are 0. We define a block B as a tuple $B =< A, Re >$ where A is an atrous convolutional [19] and Re is an activation function (using the notation introduced in [20]) with:

$$A = \sum_k x[r \cdot k]w[k]$$

$$Re = ReLU(x)$$

Where the atrous rate r corresponds to the stride used for sampling the input signal $S \times T$ defined as $x =< S_1, S_2, S_3, \ldots, S_T >$, k is the length of the convolutional kernel and w represented its weights. We could observe that standard discrete convolution with kernel dimension k is a special case of atrous convolution with rate $r = 1$. Therefore, our network is composed by the input layer $I^S \times T$, two blocks $\sum_{r=2}^{3} B$ with increasing atrous rate followed by a fully-connected layer with SoftMax activation function to obtain classification results, i.e. $\mathbf{Y_t}$ vector.

The proposed methods 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 2 GB video memory. We used ADAM stochastic optimization algorithm during training with 200 epochs and batch size of 16.

IV. Experiments

We tested our method on two different datasets, public available JIGSAWS benchmarking dataset and a novel annotated dataset called V-RASTED (Virtual Robotic Assisted Surgery Training Evaluation Dataset). The objective of both datasets is to provide ground truth annotation for evaluation of automatic actions recognition. Although they show differences on the classification objective (i.e. different actions labelling) they present comparable synchronized kinematic variables and video data. A method able to obtain optimal results on both datasets implies the ability to generalize actions recognition abstracting the specific dataset and actions considered.

TABLE I
DATASETS GESTURES DESCRIPTIONS

JIGSAWS DATASET

ID	DESCRIPTION
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G7	Pulling suture with right hand
G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

V-RASTED

ID	DESCRIPTION
1	Colleting the ring
2	Passing the ring from the right arm to the left arm
3	Posing the ring in the correct pole
4	Failing grabbing the ring
5	Failing passing the ring from right arm to left arm
6	Failing posing the ring in the correct pole

A. JIGSAWS

JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) is a public dataset that contains synchronized video and kinematic data from three tasks, Suturing (SU), Needle Passing (NP) and Knot Tying (KT), which are common exercises used for skills assessment in MIS training. For each task, eight users performed five trials resulting in 39 sequences. The motion of each manipulator was described by a local frame attached at the far end of the manipulator using 19 kinematic variables, therefore there are 76-dimensional data considering the 4 manipulators involved: left and right for master and slave side. The 19 kinematic variables for each manipulator include Cartesian position, rotation matrix, linear velocities, angular velocities, and instrument gripper angle. The classification classes are fifteen and are defined in Table I. For more details about this dataset please refer to [21], [22].

B. V-RASTED

The JIGSAWS dataset contains a handful of samples of few surgical tasks and multi-class skill labels of beginner, intermediate, and expert. However, this dataset does neither establish a correlation to training in a virtual environment, which is a more cost-effective alternative to using real surgical robotic system, nor it allows following the learning progress of a trainee, since a limited number of trials are considered for each user and exercise.

To address these limitations, we have created the V-RASTED dataset by collecting training exercises performed by a group of 18 medical students taking an elective course in robotic surgery. All the students have no previous experience with SRSs, therefore to introduce a slight variability in the

Fig. 3. The hardware training console used by one of the student during data acquisition. The subject is looking inside the stereo viewer and using the two master manipulator for controlling the virtual surgical robotic instruments.

initial skill level we have divided them into two groups of 9 students each. The first group performed a training session on a da Vinci surgical robot, where each student had one hour time-slot to perform the same task that will be performed on the virtual simulator. The second group skipped the training session with the real robot and used only the virtual simulator.

We set up an experiment to acquire synchronized kinematic data and endoscopic stereo images from a research version of Xron virtual training simulator (BBZ srl, Verona, Italy). Xron is a realistic robotic surgery simulator based on Bullet Physics Library and developed following the training guidelines of different training centers and surgeons in Europe and USA. Xron accurately simulate SRS experience in the execution of a wide range of training exercises of increasing difficult level. The physic simulation includes patient-side manipulator kinematic together with realistic stereo video rendering. The simulator is running on a Leo master console (BBZ srl, Verona, Italy) visible in Figure 3, a compact hardware device integrating two maters manipulators, high-definition stereo viewer and foot-pedals tray. The console guarantees an immersive user-experience, replicating the most significant features offered by clinical SRSs.

Each student repeated the same basic-skill exercise consisting in a pick-and-place task. The task consisted of positioning a set of colored rings in their correct position on a peg board, as shown in Figure 1. The exercise consists in lifting a ring with one of the robotic instrument, passing the ring on the other robotic arm and positioning it in the corresponding pole. This exercise replicates the one offered in standard MIS training curriculum and available in all SRS virtual training simulators.

Each student had a time slot of one hour on the simulator, divided in two half-hour sessions: the first is dedicated to practicing with the specific platform and the second is dedicated to the recorded trials. During the first session the ring positioning sequence was randomly selected while in the second one a fixed sequence was used. Each student performed from a minimum of ten to a maximum of twenty recorded trials resulting in a total 276 sequences. The dataset included synchronized stereo images, kinematic variables for each slave manipulators simulated (one camera arm and two instrument arms) and some status variable for recording foot-pedals activities. For each slave manipulator 13 variables are recorded describing instrument status: Cartesian tri-dimensional position, orientation represented as rotation matrix, and gripper opening angle. All the trials have been manually annotated and checked by two experts. The six annotated classes are defined in Table 1, they include errors occurring during task execution to better describe the training process.

The dataset and related detailed documentation is available at *gitlab.com/altairLab/v-rasted.git*.

C. Evaluation Methodology

For both datasets we consider the following kinematic variables for each slave manipulator: (3) Cartesian position in meters, (4) Cartesian orientation represented with quaternion, (1) grasping angle expressed in radiant. Therefore, we consider a total of sixteen input features for each dataset. In both dataset the camera is not moved during the trials, therefore not considering camera information is not a limitation. Using quaternion for describing the orientation instead of other representations (e.g. rotation matrix, axis-angle, Euler angles) reduce the redundant information and lower the number of features by keeping at the same time a unique representation.

We performed the evaluation using Leave One User Out (LOUO) methodology as described in [22]. We use the LOUO methodology because both evaluation datasets are natively structured to support it and LOUO is usually more challenging than other evaluation strategies. To try to maintain the evaluation methodology consistent with those previously adopted we used the macro/micro accuracy, precision and their standard deviation. Micro-average is preferable for classification on unbalanced classes since it aggregates the contributions of all classes to compute the average metric in terms of precision also considering false positive. On the other hand, macro accuracy will compute the metric independently for each class. All these metrics have been described in detail in [12]. We add a new metric defined as F_1 score, described below:

$$F_1 = 2\frac{(precision \cdot recall)}{(precision + recall)}$$

Where precision is defined as $tp/(tp + fp)$ and recall define $tp/(tp + fn)$ where tp, fp, fn stand for true positive, false positive and false negative respectively. F_1 is the harmonic average of precision and recall, and it ranges from 1 in case of perfect precision and recall to the minimum of 0. We have selected this metric because it allows a better generalization of the results without making assumptions about which metrics is most relevant for the evaluation of the specific study. F_1 represent a balance between precision and recall. For all the metrics we calculate the standard deviation on all subject tasks.

TABLE II
BEST PERFORMANCE OF TDNN VALIDATED ON THE JIGSAWS AND V-RASTED DATASETS, FOR LOUO CROSS-VALIDATION

	JIGSAWS			V-RASTED
Evaluation	Suturing	Knot Tying	Needle Passing	Peg & Ring
Micro	74.4	73.44	64.36	68.4
±Std	7.41	9.58	11.65	6.95
Macro	53.89	70.27	48.26	55.81
±Std	8.39	12.17	7.83	6.08
Precision	67.35	62.19	55.55	69.75
±Std	9.55	21.68	16.36	12.07
F_1 Score	0.7574	0.7239	0.6383	0.7689
±Std	0.0885	0.1785	0.1556	0.0829

TABLE III
COMPARISON OF MEAN LOUO PERFORMANCE BETWEEN LSTM AND TDNN FOR JIGSAWS SUTURING TASK

	Evaluation			
Model	Micro [%]	Macro [%]	F_1	Time [s]
TDNN	80.4	60.43	0.7362	135
LSTM	61.99	40.43	0.66	10093

V. Results

Table II shows the results obtained by applying the proposed TDNN to the two datasets. The data reported were obtained trying to maximize the micro accuracy. The values presented for each measure correspond to the average value of the users tested with LOUO methodology. This means that for each user the training and the testing set are recalculated, and the value shown in the table is the average of these results for all subjects. The standard deviation also allows us to evaluate how much the result is influenced by the quality execution of the tasks by different user considered. All results are presented in percentage value except for the F_1 score which instead ranges from 1 to 0. The different exercises included in the JIGSAWS dataset have been trained separately to follow the same procedure described in [12].

In Figure 3 we report the V-RASTED normalized confusion matrix resulted from LOUO evaluation. This matrix is particularly explanatory to show which types of classes are correctly recognized and which classes are the most complex to predict, as discussed in the following section.

VI. Discussions

The micro accuracy shows uniform results regardless of the exercise considered, see Table II. This leads to the successful recognition of surgical gestures by TDNN from kinematic data with not significant influence from the dataset considered and regardless of the specific type of actions. The proposed method can give a semantic interpretation to the kinematic sequence by defining the temporal correlation as a fundamental characteristic for the recognition of surgical gestures. Considering results reported in Table II, the proposed method obtains accuracy performance comparable to state of art methods, which provide micro accuracy between 74.77 and 81.74 on JIGSAWS tasks [12].

The introduction of F_1 metric proposed in this work, confirms that this metric is suitable for describing gesture classification performance independently of the specific dataset considered giving a score that is robust to different type of statistical hypothesis testing error. The results presented in Table II confirm this, in particular considering that the Suturing task of JIGSAWS and the peg and ring task in V-RASTED are the two most similar exercises. To further support this observation, we have kept the same network models for both datasets avoiding operations of fine tuning for a specific dataset. However, the results presented in Table III demonstrate the outstanding result on JIGSAWS Suturing dataset by tuning the input features selection (including also Cartesian velocities) and network dimension adding one more block level as defined above. The gesture classification results obtained on V-RASTED are mainly limited by incorrect recognition of gesture classes and their corresponding error classes. This fact is clearly represented by the normalized confusion matrix showed in Figure 3. This result was expected since based on the kinematics it is not possible to observe errors deriving from the environment and not from the system such as the fall of a ring during grabbing or passing gesture or the incorrect positioning of the ring in the corresponding pole. To improve this result, it would be necessary to introduce environment sensing, based on video data or external sensors.

Another significant characteristic of the TDNN networks is the training speed. TDNN and TCN are lighter and faster than other model used for modeling temporal context as recurrent network. In the last column of Table III, we have compared the proposed method with LSTM network considering the training time of a single user LOUO evaluation, showing the much faster performance of the proposed method (more than 8x speed-up). This can become an essential aspect for the development of real time methods for the use of causal data given the need to continuously train the network and avoid delays during the evaluation. The results obtained using kinematic reproduced by the BBZ simulator enable the development of future datasets in simulated environments that would allow a greater collection of data and a possible expansion of the gestures to be recognized thanks to automatic annotation of procedure actions. Although there are numerous positive aspects on TDNN we must admit that results on macro accuracy are not particularly high. This is due to the poor distribution of classes within datasets that make it difficult to model classes represented by few observations. The class frequencies for both datasets are reported in Figure 4, and the represented histogram further supports this observation by showing not homogeneous classes distribution. Although the TDNN networks are designed to robustly process different datasets, the structure of the training datasets are still influencing the accuracy of movements detection. One aspect that should be investigated through the analysis of the features and the corresponding results consists in how single movements are modeled to understand in depth what are the limitations of neural networks compared to the classical models and what type of gestures are more easily recognized with parity of

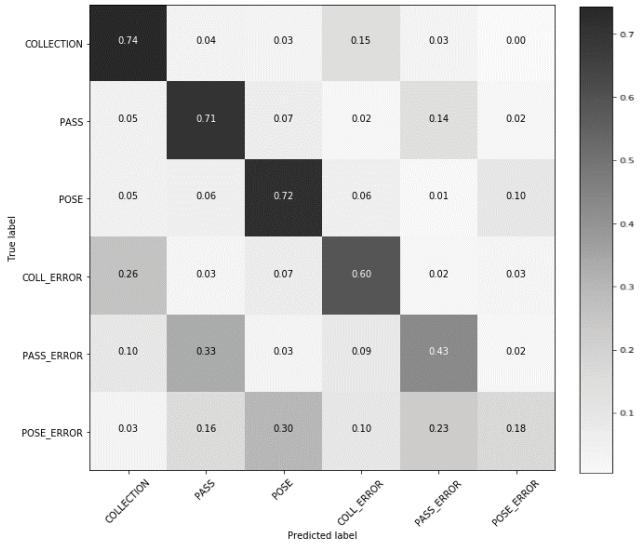

Fig. 4. Normalized Confusion matrix obtained for the proposed method during LOUO evaluation on V-RASTED

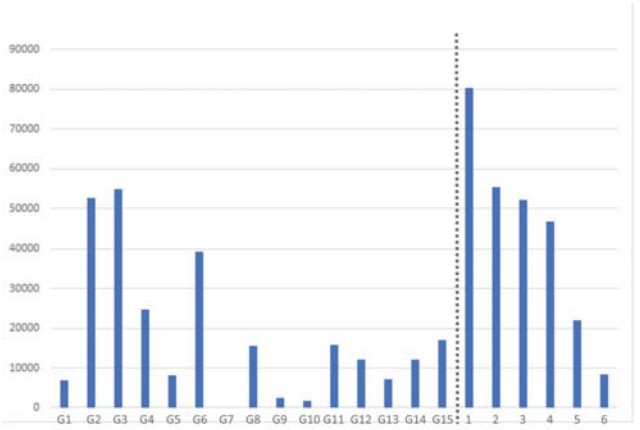

Fig. 5. Class frequencies (expressed as number of occurences) distribution for JIGSAWS and V-RASTED datesets. See Table I for class labels description.

competence on the surgeon. This would bring a greater link between the physical meaning of the features and its semantic meaning and would allow the search for new measures able as features to further distinguish between surgical gestures

VII. Conclusion

In this work we have introduced a Time Delay Neural Network able to achieve gesture recognition results in line with state of the art methods and we have tested the proposed method on an additional dataset to prove its robustness. We have reported the results obtained using the same types of input. We have shown that our network also has improvements in efficiency.

As soon as possible we would like to extend the V-RASTED dataset by adding other exercises including different movements, executed both in virtual environment and with a real SRS. Another future work would be to create a dataset with the same gestures of the JIGSAWS and test transfer learning techniques applied to the proposed method. We would also like to investigate the possible fusion of kinematic and video to obtain non-redundant information from the camera that could be able to provide environment information suitable for improving manipulation errors detection. Finally, we would like to improve the correlation between the physical meaning of the kinematic and movement recognition.

Acknowledgment

The authors thank BBZ srl, in particular Davide Zerbato and Francesco Bovo, for the support in data acquisition with their virtual training simulator.

References

[1] K. Moorthy, Y. Munz, A. Dosis, J. Hernandez, S. Martin, F. Bello, T. Rockall, and A. Darzi, "Dexterity enhancement with robotic surgery," *Surgical Endoscopy and Other Interventional Techniques*, vol. 18, no. 5, pp. 790–795, 2004.

[2] A. Diodato, M. Brancadoro, G. De Rossi, H. Abidi, D. Dall'Alba, R. Muradore, G. Ciuti, P. Fiorini, A. Menciassi, and M. Cianchetti, "Soft robotic manipulator for improving dexterity in minimally invasive surgery," *Surgical innovation*, vol. 25, no. 1, pp. 69–76, 2018.

[3] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield *et al.*, "The grand challenges of science robotics," *Science Robotics*, vol. 3, no. 14, p. eaar7650, 2018.

[4] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.

[5] G. Lee, M. Lee, I. Green, M. Allaf, and M. Marohn, "Surgeons' physical discomfort and symptoms during robotic surgery: a comprehensive ergonomic survey study," *Surgical endoscopy*, vol. 31, no. 4, pp. 1697–1706, 2017.

[6] D. Zerbato and D. Dall'Alba, "Role of virtual simulation in surgical training," *Journal of visualized surgery*, vol. 3, 2017.

[7] R. DiPietro, C. Rupprecht, N. Navab, and G. D. Hager, "Analyzing and exploiting narx recurrent neural networks for long-term dependencies," *arXiv preprint arXiv:1702.07805*, 2017.

[8] L. Zappella, B. Béjar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Medical image analysis*, vol. 17, no. 7, pp. 732–745, 2013.

[9] J. Rosen, B. Hannaford, C. G. Richards, and M. N. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 579–591, 2001.

[10] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 339–346.

[11] C. Lea, G. D. Hager, and R. Vidal, "An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks," in *Applications of computer vision (WACV), 2015 IEEE winter conference on*. IEEE, 2015, pp. 1123–1129.

[12] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[13] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, 2016.

[14] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4597–4605.

[15] C. Dugast, L. Devillers, and X. Aubert, "Combining tdnn and hmm in a hybrid system for improved continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 217–223, 1994.

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 568–576. [Online]. Available: http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf

[18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Readings in speech recognition*. Elsevier, 1990, pp. 393–404.

[19] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Springer, 1990, pp. 286–297.

[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[21] A. Murali, A. Garg, S. Krishnan, F. T. Pokorny, P. Abbeel, T. Darrell, and K. Goldberg, "Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4150–4157.

[22] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, vol. 3, 2014, p. 3.