Recognition of Endovascular Manipulations using Recurrent Neural Networks

Rui-Qi Li[1,2], Xiao-Hu Zhou[1,2], Gui-Bin Bian[1,2], Xiao-Liang Xie[1,2] and Zeng-Guang Hou[1,2,3]

***Abstract*— The ability to accurately recognize elementary surgical gestures is a stepping stone to automated surgical assessment and surgical training. In this paper, a long short-term memory (LSTM) recurrent neural network is applied to the task of recognizing six typical manipulations in percutaneous coronary intervention (PCI). The manipulation mentioned above is referring to the atomic surgical operation, also called surgeme in many research. Instead of using the video data or kinematic data of surgical instruments, we propose to use the kinematic data of the operator's hand acquired by our wearable data glove to recognize the manipulations. To establish a baseline for comparison, a method based on Hidden Markov Model (HMM) is applied because HMM is frequently used in the tasks of surgical sequence learning. Two cross-validation schemes are used in our experiments, they both illustrate that our LSTM-based method far outperforms the HMM-based method. To our knowledge, this is the first paper to apply the LSTM recurrent neural network in the field of PCI.**

I. INTRODUCTION

Recognition of surgical manipulations is an important prerequisite for some higher-level surgical tasks, such as objective assessment of surgical skills and surgical training. For percutaneous coronary intervention (PCI), in which surgeons need to manipulate guidewires in fragile blood vessels, although the surgeon's manipulations on the guidewire only have two degrees of freedom, complex tool-tissue interactions still require the surgeon to undergo more hours of training to acquire dexterous surgery skills. Because of the complexity of PCI manipulations, as the prerequisite of some high-level tasks, recognition of surgeon's manipulations in real time is of great significance. It can not only help us better understand the action intention of the surgeon during the surgery [12], but also provide targeted feedback to the novice in time during the training [2]. It can also help design the transmission structure of the PCI surgical robot [1].

Specifically, there are six types of manipulation applied to the guidewire: (1) Advancement (AV), (2) Retracement (RT), (3) Rotation Clockwise (RC), (4) Rotation Counterclockwise (RCC), (5) Advancement and Rotation Clockwise (ARC), (6) Advancement and Rotation Counterclockwise (ARCC). There is no Retracement and Rotation Clockwise or Retracement and Rotation Counterclockwise because these two types of manipulation are unnecessary.

From the perspective of data acquisition, the most straightforward way is to place a positioning sensor on each surgical instrument, then manipulations can be recognized through the trajectory of the surgical instruments. However, because of the long and thin structure of the guidewire, it is impossible to attach any sensor to the guidewire. Hence, hand motion data is proposed to be used and there are three reasons for this. First, the type of manipulation can be recognized by observing the hand and fingers movements. Second, hand motion data can be easily acquired by the sensors fixed on the wearable glove. Third, there is almost no interference to the surgery process.

The recognition task is actually a sequence labeling task, so the recognition model to be used must have the ability to learn the time dependence within the time series. Up to now, the recognition models used in most of the related works are based on various variants of Hidden Markov Model (HMM), in which each surgical process is simplified into a Markov process. However, this simplification is actually inappropriate, and as a result, the recognition accuracy of the HMM-based methods will decrease when the sample diversity increases.

In this paper, a long short-term memory (LSTM)-based method is proposed to recognize PCI manipulations. This is the first time to introduce the LSTM-based method into the field of PCI. We compare our LSTM-based method with the conventional HMM-based recognition method, and find that our LSTM-based method outperforms the HMM-based method when the number of samples is relatively abundant. This means that the surgery process should not be simplified as a simple Markov process, and our LSTM network can better learn the complex time correlation during the manipulation.

II. RELATED WORK

A. LSTM

LSTM was first introduced in [3] as an improved architecture of the recurrent neural network (RNN) for sequence learning. Unlike the traditional RNN, LSTM can address the vanishing gradient problem so that the LSTM network is easy to train and suitable to learn long-term time dependencies. Since the original LSTM was introduced, several variants called modern LSTM have been proposed, including adding forget gates [4] and adding peephole connections [5] to the original LSTM cell. In recent years, methods based on LSTM

[1]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China

Emails: {liruiqi2016, zhouxiaohu2014, guibin.bian, xiaoliang.xie, zengguang.hou}@ia.ac.cn

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

have achieved state-of-the-art results in a wide range of supervised and unsupervised machine learning tasks [6].

B. Surgical Sequence Learning

There are very few researches aim to recognize surgical manipulations using the segmented time series. C.E.Reiley *et al.* [7] used several statistical models to recognize the gesture in suturing surgery performed by the da Vinci system in order to verify the applicability of the statistical models when data variability increased. L.Zappella *et al.* [8] also used a variety of methods to recognize surgical gestures from video data and kinematic data, which in order to prove the combination of both kinematic and video data outperformed any other algorithm based on one type of data alone. Our previous work [9] used an HMM framework with a Gaussian mixture model (GMM) as continuous observations to recognize manipulations using the same segmented sequences.

Most researches on surgical sequence learning mainly focus on two tasks: workflow segmentation and surgical skill evaluation. Prior work in workflow segmentation based on variants of HMMs [10] and conditional random field (CRF) [11]. In recent years, several research studies have used the LSTM networks [12] achieving the state-of-the-art segmentation results. As for surgical skill evaluation, variants of HMMs are widely used in most of the researches. N.Ahmidi *et al.* [13] used discrete HMM to classify seven surgical tasks and two levels of the surgical skills in functional endoscopic sinus surgery. J.Leong *et al.* [14] used HMM with GMM to classify skill levels in laparoscopic surgery. The differences between these surgical skill evaluation researches are mainly reflected in the types of surgery and the data sources used.

III. DATASET

A. Data Glove

Our modified data glove is used to acquire the hand kinematic data, the glove is shown in Fig. 1. This glove contains seven sensors, including three 6-DOF electromagnetic (EM) sensors (two of them fixed on the forefinger tip and thumb tip respectively, one put on the wrist), and four fiber-optic sensors (FOS) (placed in four knuckles of the thumb and forefinger). As shown in Fig. 1, the glove only acquires the kinematic data of thumb and forefinger. This is because the rest of the fingers are useless during the PCI.

Fig. 1. (left) The prototype of modified data glove. (right) the position of fiber-optic sensors and EM sensors.

Each of the 6-DOF EM sensors can record the data of its three-dimensional position and Euler angles, so from these EM sensors, the posture of the wrist and two fingertips can be fully acquired. Each of the fiber-optic sensors can detect the degree of bend of each knuckle, the more bent the knuckle is, the larger the output value will be. So from the fiber-optic sensors, the shape of the fingers can be inferred. Four fiber-optic sensors output 4-dimensional data, and three EM sensors output 18-dimensional data, so the dimension of all the time series samples in the dataset is 22. Besides, the sampling frequency of our glove is approximately 25 Hz.

B. Acquisition Process

The data acquisition setup is shown in Fig. 2. Eight operators' data (2 experts, 6 novices) is collected in our experiments. Each operator wears our modified data glove on his right hand and performs all six types of the manipulations on the platform of our 3D vascular model. In each data acquisition process, the operator can only consistently complete the corresponding manipulation, cannot pause or mix other manipulations. Each type of manipulations is repeated 10 to 20 times for each operator. All the manipulations must be carried out by the thumb and forefinger of the right hand. In order to increase the diversity of the data, there is no time limit or action limit in our acquisition process.

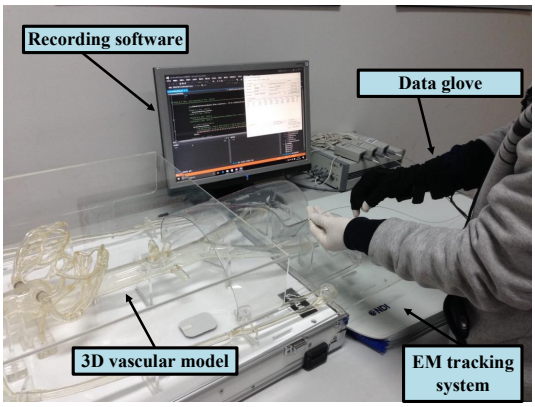

Fig. 2. The experimental setup

C. Data Processing

After acquiring the original sequential data, some processing is needed.

1) Purify the data: Useless actions are inevitably recorded at the beginning and the end of the original data. All these useless data need to be deleted to ensure the purity of data.

2) Noise filtering: A median filter is used to eliminate the noise of the acquired data.

3) Crop data: The duration of data varies between 2 to 9 seconds because there is no time limit. However, the average duration of each manipulation during the surgery is about 1 second, it is meaningless to recognize the manipulations with long durations. So the original data is cropped and the average duration of the cropped results is at about 1 second. According to the sampling frequency of 25Hz, the average length of all new samples is about 25 (range from 13 to 40). To verify the robustness of classifier we still keep the

sequences of different lengths. It's worth noting that one extra advantage of cropping samples is that the number of samples can be greatly increased, which is very suitable for LSTM training.

Finally, kinematic data of all six manipulations performed by eight operators has been acquired, and after the data processing, the dataset contains a total of 2979 available samples.

IV. METHODS

A. LSTM

LSTM networks have been applied successfully to many diverse sequence-modeling tasks. Because of its powerful sequence learning ability, LSTM is suitable for the task of sequence labeling. Our proposed LSTM network use the memory cell with forget gates [4] but without peephole connections. A schematic diagram of the LSTM cell architecture is shown in Fig. 3.

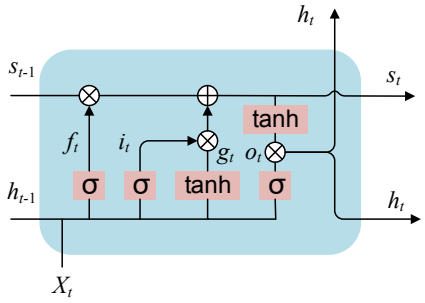

Fig. 3. Schematic of an LSTM cell.

The following equations give the update for each cell at a given timestep.

$$i_t = \sigma(W_i X_t + U_i h_{t\ \ 1} + b_i) \quad (1)$$

$$f_t = \sigma(W_f X_t + U_f h_{t\ \ 1} + b_f) \quad (2)$$

$$o_t = \sigma(W_o X_t + U_o h_{t\ \ 1} + b_o) \quad (3)$$

$$g_t = tanh(W_g X_t + U_g h_{t\ \ 1} + b_g) \quad (4)$$

$$s_t = g_t \odot i_t + s_{t\ \ 1} \odot f_t \quad (5)$$

$$h_t = tanh(s_t) \odot o_t \quad (6)$$

In these equations, X_t is the input, i_t, f_t, and o_t represent the value of the input, forget, and output gates respectively. g_t represents the update value to the hidden state, s_t is the current hidden state. h_t is the output of the network. σ stands for an element-wise application of the sigmoid function, and $\odot$ is the Hadamard (element-wise) product.

The LSTM network used in our experiments only has one hidden layer, and the network architecture unrolled over time is shown in Fig. 4. Although our LSTM network has output at every timestep, only the output at the final timestep h_T is used as network output. Following the final output h_T, a fully connected layer is used and the output of the layer is a 6-D vector, each dimension represents a category of manipulations. The category corresponding to the dimension with the largest value is the recognition result of the input sequence. Categorical cross entropy loss is used as the loss function to train the network.

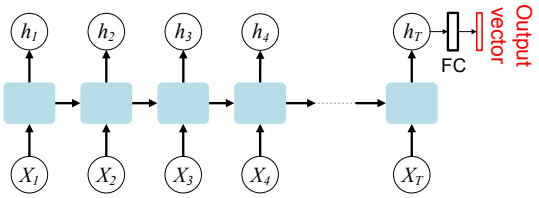

Fig. 4. Schematic of an LSTM network unrolled over time.

B. Cross-Validation

Two cross-validation schemes are used in our experiments. The first one is the 5-Fold cross-validation scheme. The dataset is randomly divided into 5 folds. Each fold is used for testing once, while the remaining 4 folds are used in training. In order to ensure the dispersity of samples, first, all samples are divided into 30 small folds (samples of each manipulation are divided into 5 small folds), then these 30 small folds are combined into five folds. This way ensures that the number of samples of each manipulation in each fold is consistent.

Noting that great degree of similarity in a given operator's set of samples of a given manipulation, in order to test the ability of the classifiers to generalize to new operators, a leave one user out (LOUO) cross-validation scheme is also applied. Each operator's samples are used for testing once, while the remaining samples are used for training. This cross-validation scheme is more indicative of the generalization and robustness of the method.

V. EXPERIMENTS

A. Implementation Details

We use a Keras implementation of the LSTM network. Since we don't have a very large number of training samples, as few parameters as possible should be used in case of overfitting. A one-layer architecture with 64 hidden nodes is used in our experiments, we use the Adam optimizer with a batch size of 32, dropout of 0.2. The LSTM network is trained for 500 epoch with the learning rate of 0.01.

The HMM-based method is followed the HMM framework described in our previous work [9] with GMM as continuous observations. The hyperparameters of the HMM-based method are optimized using random search.

B. 5-Fold Cross-validation

The dataset is divided into 5 equal folds as introduced before. Our LSTM network and the HMM-based method are validated using the same folds to ensure the experimental conditions are consistent. All test results are summarized as a confusion matrix. Confusion matrices of the HMM-based method and LSTM-based method are shown in Fig. 5 (a),(b).

From the results, the average recognition accuracy of the HMM-based method is 92.72%. Our LSTM-based method outperforms the HMM-based method, achieving an average recognition accuracy of 99.13%. From the confusion matrices, we find that the HMM-based method is occasionally confused between the RC and RCC, and between the ARC and ARCC, while our LSTM network almost recognizes all kinds of manipulations correctly.

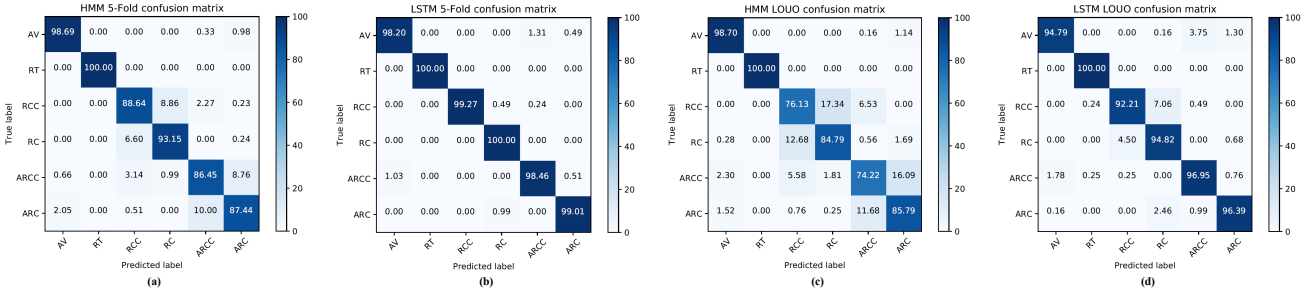

Fig. 5. Confusion matrices of all cross-validations.

C. LOUO Cross-validation

To verify the generalization performance of two methods, a leave one user out (LOUO) cross-validation scheme is used. Confusion matrices of the HMM-based method and LSTM-based method are shown in Fig. 5 (c),(d).

The average recognition accuracy of the HMM-based method is 87.06%, and the average recognition accuracy of our LSTM-based method is 95.97%, still significantly better than the HMM-based method. The results prove that our method has better generalization performance.

Clearly the results in LOUO cross-validation scheme are overall worse than the results in 5-Fold cross-validation scheme. This phenomenon is reasonable. Because everyone has his own fixed mode of manipulation, so when an operator repeats a manipulation several times, the collected data is very similar. Therefore, the 5-Fold scheme can ensure the sample distributions of the training set and testing set are consistent. For each sample in the testing set, we can find similar samples in the training set. By contrast, the LOUO scheme cannot guarantee the consistency of sample distribution, as a result, the recognition accuracy in LOUO scheme is worse than the accuracy in the 5-Fold scheme.

Even though the results are bound to get worse, we can observe that, compared with the results in the 5-Fold scheme, the average accuracy of the HMM-based method decreases more than our LSTM-based method, which illustrates the robustness of our LSTM-based method.

VI. CONCLUSIONS

In this paper, an LSTM-based method is proposed for endovascular manipulations recognition using the kinematic data of the operator's hand. In order to verify the superiority of the method, we compare our method with the conventional HMM-based method. The experimental results show that our method far outperforms the HMM-based method in both recognition accuracy and model robustness. It can also be explained that simplifying the surgical process into a Markov process is not an appropriate approach, but a compromise in the case of insufficient data. If having enough data to train the model, obviously LSTM network can build a better model of the surgical process than HMM.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (Grants 61533016, U1713220, U1613210) , by the National Key Research and Development Program of China under Grant 2017YFB1302704, by the Strategic Priority Research Program of CAS under Grant XDBS01040100.

REFERENCES

[1] G. Bin Bian et al., "An enhanced dual-finger robotic Hand for Catheter manipulating in vascular intervention: A preliminary study," *2013 IEEE International Conference on Information and Automation (ICIA)*, Yinchuan, pp. 356-361, 2013.

[2] W. Chi et al., "A learning based training and skill assessment platform with haptic guidance for endovascular catheterization," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, pp. 2357-2363, 2017.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[4] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451-2471, 2000.

[5] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Como, Italy, pp. 189-194 vol.3, 2000.

[6] Lipton, Zachary C., John Berkowitz, and Charles Elkan. "A Critical Review of Recurrent Neural Networks for Sequence Learning," arXiv preprint arXiv:1506.00019 (2015).

[7] C. E. Reiley et al., "Automatic recognition of surgical motions using statistical modeling for capturing variability," *Studies in Health Technology and Informatics*, vol. 132, pp. 396-401, 2008.

[8] L. Zappella, B. Bejar, 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] X. Zhou, G. Bian, X. Xie, and Z. Hou, "An HMM-Based Recognition Framework for Endovascular Manipulations," *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, pp. 3393-3396, 2017.

[10] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation," *International conference on Information Processing in Computer-Assisted Interventions*, Springer, Berlin, Heidelberg, pp. 167-177. 2012.

[11] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, pp. 1642-1649, 2016.

[12] R. DiPietro et al., "Recognizing Surgical Activities with Recurrent Neural Networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 551-558, 2016.

[13] N. Ahmidi, G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii, "Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, Heidelberg, pp. 295-302, 2010.

[14] J. J. H. Leong, M. Nicolaou, L. Atallah, G. P. Mylonas, A. W. Darzi, and G.-Z. Yang, "HMM assessment of quality of movement trajectory in laparoscopic surgery," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, Heidelberg, pp. 752-759, 2006.