Surgical Phase Recognition Method with a Sequential Consistency for CAOS-AI Navigation System

Shoichi Nishio
Graduate school of Engineering
University of Hyogo
Himeji, Japan

Belayat Hossain
Graduate school of Engineering
University of Hyogo
Himeji, Japan

Naomi Yagi
Himeji Dokkyo University
Himeji,Japan

Manabu Nii
Graduate school of Engineering
University of Hyogo
Himeji, Japan

Takafumi Hiranaka
Takatsuki General Hospital
Takatsuki, Japan

Syoji Kobashi
Graduate school of Engineering
University of Hyogo
Himeji, Japan

2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech) 978-1-7281-7063-3/20/$31.00 ©2020 IEEE 10.1109/LifeTech48969.2020.1570619203

Abstract—The procedure of orthopedic surgery is quite complicated, and many kinds of equipment have been used. Operating room nurses who deliver surgical instruments to surgeon are supposed to be forced to incur a heavy burden. There are some studies to recognize surgical phase with convolutional neural network (CNN) in minimally invasive laparoscopic surgery only. Previously, we proposed a computer-aided orthopedic surgery (CAOS)-AI navigation system based on CNN. However, the work propose a method to improve accuracy of phase recognition by considering temporal dependency of orthopedic surgery video acquired from surgeon-wearable video camera. The method estimates current surgical phase by combining both temporal dependency and convolutional-long-short term memory network (CNN-LSTM). Experimental results shows a phase recognition accuracy of 59.9% by the proposed method applied in uni-comapartmenatal knee arthroplasty (UKA).

Keywords—Deep Learning, Computer-aided Orthopaedic Surgery, Operating Room Nurse, Phase Recognition

I. INTRODUCTION

Artificial knee replacement is performed to restore the normal functioning of the knee by replacing the damaged part of the knee due to osteoarthritis (OA) or rheumatoid arthritis (RA) [1] with a prosthetic implant. The number of cases in Japan is increasing year by year, exceeding 80,000, making it one of the major orthopedic surgeries. Knee replacement is mainly classified into total knee replacement (TKA) and unicompartmental knee arthroplasty (UKA). UKA replaces a part of the slightly damaged knee joint surfaces. Surgery consists of many surgical techniques, and many surgical instruments and manual assembly of instruments during surgery are also required. Most hospitals have surgical tools of multiple implants manufacturer and different model number. The surgical techniques and surgical instruments Surgical procedures and surgical tools differ for depending on the model number and instrument manufacturer.

It is a heavy burden for instrumentation nurses in charge of multiple types of surgery to grasp these complex procedures and surgical instruments. Therefore, it is desired to introduce a computer-aided orthopedic surgery-artificial intelligence navigation system (CAOS-AI Navigation System) [1] that automatically assists medical staff, including nurses, who take out instruments during surgery, by informing them the status of surgery

In previous study, Jin proposed SV-RCNet [2], which consisted of ResNet-50(CNN) and LSTM (long short term memory), also achieved 81.7% in the cholecystectomy surgery (Cholec80-dataset). However, the surgical video taken by the surgeon's wearable device differs from the surgical video of the Cholec80 dataset in that there are many differences between the surgery, such as angle, distance, and lighting environment. Therefore, it is difficult to directly apply the existing method to knee arthroplasty.

In this study, we propose a surgical phase recognition method based on temporal dependency with convolutional-LSTM Network targeting 11 types of surgical phase (11 classes) that appear frequently in UKA surgery.

II. PROPOSE METHOD

A. Convolutional-LSTM Network for Surgical Phase Recoginition

Figure 1 shows our phase recognition model using convolutional-LSTM Network. As shown in Figure 1, the model consists of multi ResNet-50 (convolutional layers), single LSTM layer, classification layer including fully connected layer and output layer (Softmax).

At time step T, N (= 10 is used in this study) previous frames x_{T-10}, x_{T-9}, …, x_T with 224×224 pixels, called video clip, are feed to multi ResNet-50. Multi ResNet-50 has N ResNet-50 expanded in time direction, and they share same weights parameter. The ResNet-50 extracts 512-D feature vector from each frame, and the extracted feature vector (512×10) is served to LSTM network as input, then it outputs feature vector (512-D) to fully-connected and Softmax layer. Finally, Softmax layer outputs vector p_t $(p_0\ p_1 \ldots p_9\ p_{10})$ with probability of 11 classes (*i.e.*, phases). The class with the highest probability is selected as the most probable surgical phase. CNN and LSTM models are trained all at once based on video clip

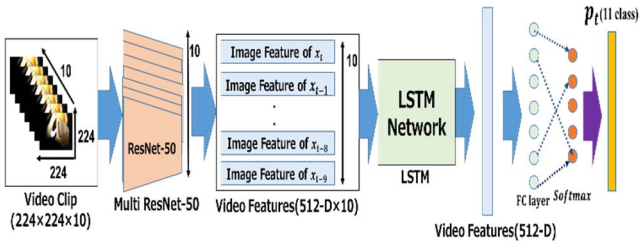

Fig. 1 Convolutional-LSTM Network with Multi ResNet-50

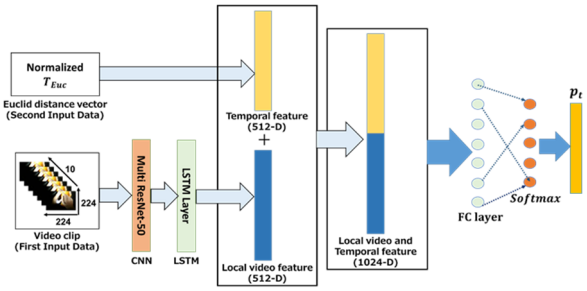

Fig. 2 Method of inserting $\boldsymbol{T_{Euc}}$ as second input data to CNN-LSTM network

data. One video clip consists of 10 frames, and the phase number is given to each video clip. The obtained vector is used to calculate the loss function value, and to update parameter of both CNN and LSTM at the same time. The initial weight of ResNet-50 is learned using ImageNet.

B. Multimodal Training Method with Temporal Feature $\boldsymbol{T_{Euc}}$ and Local video feature

UKA surgery is basically performed according to established surgical procedures. Therefore, we can use temporal information with local video feature (spatial information) extracted from video clips. We define a temporal feature (t_{els}) expressing the elapsed time of the current surgical procedure by the following equation (1), where t_{els}: elapsed time, t_{cur}: current time, and t_{str}: start time.

$$t_{els} = t_{cur} - t_{str} \qquad (1)$$

The start time t_{str} is defined as the time when the incision (*Phase 1*) is started, and the current time t_{cur} is defined as the elapsed time of the last frame of video clip to be recognized in real-time. We calculate μ_i as average of all t_{els} each surgical phase i ($i = 0, 1, ..9, 10$) from training dataset. We define $\boldsymbol{\mu}$ as statistical temporal information by following equation (2).

$$\boldsymbol{\mu} = \sum_{i=0}^{10} \mu_i \qquad (2)$$

Next, we use both t_{els} and $\boldsymbol{\mu}$ for implementing temporal feature from surgical video clips. We firstly extract t_{els} from surgical video (*Phase* i), then calculate euclidean distance (T_{Euc_i}) between t_{els} and μ_i by following equation (3).

$$T_{Euc_i} = |t_{els} - \mu_i| \qquad (3)$$

T_{Euc_i} is calculated in all phase, so 11 number of T_{Euc_i} are generated from one surgical video clip. All T_{Euc_i} are define as temporal feature vectors ($\boldsymbol{T_{Euc}}$) based on euclidean distance by following equation (4).

$$\boldsymbol{T_{Euc}} = \sum_{i=0}^{10} T_{Euc_i} \qquad (4)$$

It is necessary to normalize $\boldsymbol{T_{Euc}}$ to 0.0 to 1.0 for inserting $\boldsymbol{T_{Euc}}$ to CNN-LSTM network by following equation (5).

$$T_{Euc_i} = \frac{1}{1 + \alpha * T_{Euc_i}} \qquad (5)$$

α is a constant parameter in equation (5), also α is determined by searching parameter using machine learning method (logistic regression). First, We use normalized $\boldsymbol{T_{Euc}}$ as feature vector representing video clip for training machine learning method. Based on training accuracy results, we determine best parameter α in equation (5).

Normalized $\boldsymbol{T_{Euc}}$ is used as second input data for recognizing phase with surgical video clip (first input data). The method of inserting $\boldsymbol{T_{Euc}}$ to CNN-LSTM network is shown in Fig. 2. $\boldsymbol{T_{Euc}}$ have 11 normalize euclidean distance value, it can be treated as temporal feature information.

The proposed CNN-LSTM network can extract local video feature (512-*D*) from surgical video clip throughout CNN and LSTM. We duplicate 11 value of $\boldsymbol{T_{Euc}}$ to 512 value to enhance importance of $\boldsymbol{T_{Euc}}$ towards fully-connected layer. Duplicated $\boldsymbol{T_{Euc}}$ is linked to local video feature (512-*D*) outputted from LSTM layer, so local video and temporal feature is generated as new feature vectors (1024-*D*). That feature are served to fully-connected and Softmax layer, estimated most probable phase.

C. Feature Series Vetor Adaptive LSTM Network (FSVAL Network)

We regard the extracted feature vector from each video clip using CNN-LSTM model as video clip's feature vector $\mathrm{v_t}$. Actually, The $\mathrm{v_t}$ is extracted from fully-connected layer of CNN-LSTM model, and it has 512 value. At time t_{cur}, we have multiple v_t extracted from all video frames between t_{str} and t_{cur}. We define multiple v_t as feature series vectors $\boldsymbol{V_t}$ showed in equation (6).

$$\boldsymbol{V_t} = [v_{t_{str}} \; v_{t_{str+1}} \quad \cdots \quad \mathrm{v_{t_{cur-1}}} \; \mathrm{v_{t_{cur}}}] \qquad (6)$$

$\boldsymbol{V_t}$ consists of t_{cur} components in time direction. t_{cur} changes in the range of t_{str} (start time of surgery) to t_{end}(end time of surgery). The end time t_{end} is defined as the time when the suture (Phase 11) is finished. We propose feature series vectors adapted LSTM network (FSVAL). It has 3 LSTM layers with dropout and fully-connected layer, Output layer (Softmax). Figure 3 shows an overview of recognition flow of FSVAL network.

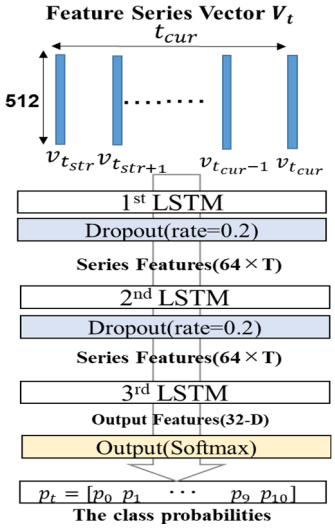

Fig. 3 Feature Series Vector Adapted LSTM Network (FSVAL)

III. EXPERIMENTAL RESULTS & DISCUSSION

In order to evaluate the effectiveness of temporal feature t_{els}, experiments were conducted in two ways- with and without t_{els} for effectively training CNN-LSTM network. We evaluated proposed method using 12 UKA surgery videos by calculating accuracy, precision, and recall on test videos following 3-fold cross validation (exp. 1, exp. 2 and exp. 3).

We extracted 25395 clips extracted from 12 UKA videos. V_t is consist of 6204 vectors extracted at 0.5 sec intervals. They were used for evaluating CNN-LSTM+$\mathrm{t_{els}}$

with FSVAL network. The parameter α in Eq. (5) was set between 0.004 to 0.009.

TABLE I shows the evaluation metric of all methods. Accuracy of CNN-LSTM with $\boldsymbol{T_{Euc}}$ is 53.3%, so temporal feature $\boldsymbol{T_{Euc}}$ contributes to the improvement of accuracy by 15%. Accuracy of CNN-LSTM with $\boldsymbol{T_{Euc}}$ and FSVAL network was 59.9 %, so FSVAL network contributes to the improvement of accuracy about 20% by using additionally FSVAL network with $\boldsymbol{T_{Euc}}$. From perspective of these metrics, method with temporal feature is more effectively than method without that.

TABLE I. Evaluation metric of all methods

(a) Accuracy in percentage

# of Exp.	CNN-LSTM	CNN-LSTM+$\boldsymbol{T_{Euc}}$	CNN-LSTM +$\boldsymbol{T_{Euc}}$ with FSVAL
Exp. 1	46.1	54.5	54.5
Exp. 2	37.0	54.7	68.8
Exp. 3	33.1	50.7	56.4
Avg.	38.0	53.3	59.9

(a) Recall in in percentage

# of Exp.	CNN-LSTM	CNN-LSTM+$\boldsymbol{T_{Euc}}$	CNN-LSTM+$\boldsymbol{T_{Euc}}$ with FSVAL
Exp. 1	40.9	50.7	43.1
Exp. 2	33.4	57.4	62.2
Exp. 3	26.6	41.2	37.1
Avg.	33.6	49.8	47.5

(b) Precision in in percentage

# of Exp.	CNN-LSTM	CNN-LSTM+$\boldsymbol{T_{Euc}}$	CNN-LSTM+$\boldsymbol{T_{Euc}}$ with FSVAL
Exp. 1	49.9	49.4	42.5
Exp. 2	42.6	56.5	59.1
Exp. 3	30.1	48.4	44.6
Avg.	40.9	51.4	48.7

However, CNN-LSTM +$\boldsymbol{T_{Euc}}$ and CNN-LSTM + $\boldsymbol{T_{Euc}}$ with FSVAL are 51.5% and 51.5% as harmonic average of 3 metrics. The significant difference in these method was not appeared in this experiment.

IV. CONCLUSION

In this study, we proposed a surgical phase recognition method based on temporal dependency. Experimental results in UKA surgeries showed that the proposed method achieved a phase recognition accuracy with 59.9%. We could show that our proposed method with temporal feature is more effectively than method without temporal feature. However, the practical use of navigation system in operating room requires high reliability -almost equal to or better than medical staff's judge. Thus, system's accuracy should be around 95% for the practical use, which is left for future work.

REFERENCES

[1] S.Nishio, M. Hossain, B. Hossain, M. Nii, T. Hiranaka, and S. Kobashi, "Real-time Orhtopaedic Surgery Procedure Recognition Method with Video Images from Smart Glasses Using Convolutional Neural Network," presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics(SMC), 2018.

[2] Y. Jin, Q. Dou, H. Chen, K. Yu, J. Qin, C. W. Fu, and P. A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," IEEE Trans. on medical imaging vol. 37, no. 5, 1114–1126, 2018.