Convolutional Neural Network based on Temporal Pose Features for Surgical Procedure Recognition

Shota Kishi\textsuperscript{1}, Nozumu Suzuki\textsuperscript{2}, Shota Tsuyuki\textsuperscript{2}, Takio Kurita\textsuperscript{3}, Fujio Miyawaki\textsuperscript{2}, and Akinori Hidaka\textsuperscript{2}

\textsuperscript{1}Graduate School of Science and Engineering, Tokyo Denki University, Japan
\textsuperscript{2}School of Science and Engineering, Tokyo Denki University, Japan
\textsuperscript{3}The Graduate School of Advanced Science and Engineering, Hiroshima University, Japan
E-mail: 20mu03@ms.dendai.ac.jp

Abstract

To compensate for severe shortage of scrub nurses who support surgeons during surgery, Miyawaki et al. have developed a scrub nurse robot (SNR) system \cite{1,2,3,4,5}. One of its current challenges is how to make the SNR recognize surgical procedures which compose a surgical operation and understand/predict surgeons’ intentions.

Therefore, in this paper, we propose a visual recognition system for surgeons’ actions based on convolutional neural network (CNN). We developed a temporal pose feature (TPF) CNN, which is a method to recognize surgical procedures based on the body movements of a surgeon’s stand-in during a simulated surgical operation. We used OpenPose to extract the pose feature vectors from every frame of the short videos filmed our simulated surgery. Besides, we used a matrix of which the pose vectors were chronologically ordered as the input of CNN by considering it as the pseudo-grayscale image.

We show that the TPF CNN was more accurate in the objects of this study than the conventional LSTM, which is used to recognize time series data. The TPF CNN shows higher recognition accuracy with fewer training than LSTM. Our results suggest that surgeons’ body movements may contain much information to be required for recognizing subtle differences in several types of surgical procedures.

1 Introduction

Scrub nurses help surgeons during surgical operations by providing surgical instruments. The collaboration task between surgeons and scrub nurses is exchange of surgical instruments. More specifically, an operating surgeon conducts various types of surgical procedures during one surgical operation and the surgical operation proceeds by surgeon’s replacing an instrument after use with an instrument necessary for the next surgical procedure. A good collaboration, therefore, is essential for a safe and smooth surgical operation. However, because the shortage of scrub nurses has been chronically severe, Miyawaki et al. proposed and have been developing a Scrub Nurse Robot (SNR) system \cite{1,2,3,4,5} in order to compensate for the shortage. The goal for the SNR project is to develop an SNR that is capable of accurately predicting what surgical instrument he/she will request next without their verbal orders and also providing it timely and safely to the surgeon.

In the process of the development, a real-time visual recognition system for SNR was developed \cite{3}. This system was capable of analyzing the surgeon’s motions during exchange of surgical instruments in order to let the SNR know the proper timing for providing the next instrument.

To make the SNR accurately predict the next surgical instrument, Miyawaki et al. developed a system which obtains in real time the information on what instruments are being used during surgery. This system also created a database for surgical instruments used in surgery, and the database enabled the SNR to predict surgical instruments at an accuracy of 70–80\%. The prediction rate will be raised if it is possible to develop another visual recognition system based on deep neural network (DNN) for surgical operations, especially, to recognize the surgeon’s motions during surgical procedures.

Therefore, in this research, we propose convolutional neural network (CNN) using temporal pose features which recognizes each of surgical procedures based on surgeon’s body movements during an operation. We show that our method is more accurate than the conventional long short-term memory (LSTM). Furthermore, our method is adequately fast for real-time processing.

Our results suggest that surgeon’s body movement contain much information to be required for recognizing subtle difference in several types of surgical procedures. Our system will be helpful for predicting surgeon’s next act and preparing required instruments.

The structure of this paper is as follows: Sec. 2 describes the existing methods. In Sec. 3, we introduce our proposal method. Sec. 4 describes our experiment of surgical procedure recognition based on surgeon’s body movement. Finally, we conclude our research in Sec. 5.

2 Related Works

Our recognition system is based on several existing methods. We use OpenPose \cite{6} to recognize a surgeon’s body and extract its movement information. The extracted information is classified by CNN to recognize a surgical procedure that the surgeon is currently conducting. Our results are compared with the recognition results of an existing method, long-short term memory (LSTM). Thus, we introduce these methods in this section.
2.1 OpenPose

OpenPose [6] is the detection tool of humans’ whole-bodies appearing in an input image; it can extract human skeletal structure by finding predefined 18 key points on main body parts, and returns the \((x, y)\) coordinates and the depth information \(s\) of each key point. Fig. 1 shows those key points.

The 36-dimensional vector \((x_0, y_0, x_1, y_1, \ldots, x_{17}, y_{17})\), that is the tuple of coordinates of detected key points, indicates the current spatial pose of a person appearing in a given frame. We refer to it as the pose feature vector. In this research, we use OpenPose to extract surgeon's body movements. The extracted pose feature vectors are used as the input of CNN and LSTM, to recognize the surgical procedure that the surgeon is currently conducting.

![Fig. 1 The key points of OpenPose (COCO model).](image)

2.2 Convolutional Neural Network (CNN)

CNN is a neural network based on the convolutional filtering operations. During the supervised learning using given labeled samples, the convolutional filters in CNN are autonomously optimized. It is known that such filters trained by using natural images can extract various visual patterns similar to the stimuli that the visual cortex in the primates’ brain can respond to [7,8].

Fig. 2 shows a typical structure of CNN. CNN has a hierarchical structure consisting of several types of layers which have specialized functions, e.g., convolution, activation, pooling, normalization, and so on. CNN is known for its extremely high recognition performance in image recognition problems.

Several studies show that CNN has high recognition accuracy not only for image data but also for data in arbitrary two-dimensional format [9,10,11]. As shown in [10], CNN functioned well to classify time series matrices which consist of OpenPose’s outputs. In this research, we also use the recognition results of OpenPose as the input of CNN to classify surgical procedures.

2.3 Long Short-Term Memory (LSTM)

LSTM [12] is a kind of Recurrent Neural Networks (RNNs) which can classify time series data. It is known that classical RNNs have several shortcomings such as gradient explosion and gradient loss problems. LSTM solved these problems by replacing the hidden layers of the RNN with memory units that enable information retention and forgetting.

In this research, we construct LSTM for surgical procedure recognition, by using the sequence of pose feature vectors obtained from OpenPose as the time series input. We compare its recognition performance with our proposed method.

![Fig. 2 The typical structure of CNN.](image)

3 CNN using Temporal Pose Features

In this research, we propose a visual recognition system to classify surgeon's movements into predefined surgical procedures performed in cholecystectomy. In this section, we denote details of our training data and recognition methods.

3.1 Dataset

To train the proposed CNN, we used a dataset obtained from simulated surgical operations for removing a gall bladder (cholecystectomy). To create this dataset, five people performed four to six operations under simulated environment using a surgical training kit (Fig. 3 top). One of the five surgeon’s stand-ins is a retired experienced surgeon, and the remaining four people are university students...
not in the medical field. Each simulated surgery took about 10 to 15 minutes, and it is filmed at 60 frames per second (fps) as a full HD video. Fig. 3 (bottom) shows one of filmed frames. There are 21 videos filmed in total.

During the process of a simulated operation, surgeon's acts can be classified as one of the following four surgical procedures: Dissection (Ds), Ligation (Lg), Division (Dv), and Hemostasis (Hm). Therefore, in this dataset, each frame of each video is assigned one of the procedure labels (Ds, Lg, Dv, or Hm) according to what a surgeon’ stand-in is conducting in the frame, or assigned no label if the stand-in does nothing in the frame.

By using such labeled frames, 2,273 short (10-second) videos are made. Each short video consists of consecutive 600 frames which have the same procedure label. Therefore, we can use the short video as a labeled training sample for the four-class action recognition problem. We extracted such samples (10-second videos) from the full video at an interval of five seconds.

3.2 Temporal Pose Features (TPF)

By using OpenPose, we extracted the pose feature vector from every frame of the short video (which has 600 frames for 10 seconds), and composed a matrix of which the pose vectors were chronologically ordered. This matrix, called temporal pose features (TPF), was used as the input of CNN by considering it as the pseudo-grayscale image [10]. Fig. 4 illustrates an overall procedure of our proposed recognition method.

In our problem setting, the CNN was trained using TPF matrices extracted from surgical procedure videos, so it is called surgical TPF CNN. Our CNN can classify an image sequence of 10-second duration into the four surgical procedures (Ds, Lg, Dv, or Hm). Fig. 5 shows the architecture of TPF CNN we used in this paper. It consists of six convolution blocks which consist of convolution, pooling, the ReLU activation, and batch normalization layers, followed by three fully-connected layers. Note that the 3rd to 5th convolution blocks do not include the pooling layer. The notations such as "285 × 1 × 128" indicate height, width, and the number of channels of the corresponding feature tensors. We did not use any padding option for convolution and pooling operations. The kernel size and the stride of the pooling layers are 2, respectively. Thus, each pooling layer halves the spatial size of the corresponding feature tensor.

4 Experiments

In this section, we describe our experiment of surgical procedure recognition based on surgeons’ body movements. We used OpenPose to extract the pose feature vectors from every frame of the short videos described in Sec. 3.1. The size of each frame is 1920 × 1080 pixels. The outputs of OpenPose include (x, y) coordinates of predefined 18 key points, where 1 ≤ x ≤ 1920 and 1 ≤ y ≤ 1080. We used normalized coordinates (x/1920 and y/1080) for the input of TPF CNN or LSTM.

In our experiments, we used bi-directional LSTM implemented in MATLAB 2020a. We set the number of hidden units in LSTM to 100. For the training of LSTM, we used Adam with a learning rate of 0.001, a number of epochs of 500, and a batch size of 50. Also, for TPF CNN, we used Adam with a learning rate of 1.0 × 10^-4, a number of epochs of 30, and a batch size of 50.

By following two experiments, we evaluated recognition performances of our TPF CNN and traditional LSTM using the pose feature vectors.

![Fig. 4 Overall procedure of the proposed recognition system](image)

4.1 Comparison between LSTM and TPF CNN

Here we describe the experiment using 2/3 of the 2,273 samples selected at random for training and the remaining samples for testing. The numbers of samples for training and testing are shown in the rows indicated as "Train1" and
"Test1" in Table 1. We performed 10 trials of the training and testing of DNNs while remaking the random training and testing sets. Table 2 shows the averaged recognition accuracy of LSTM and TPF CNN. While LSTM could recognize the dissection class and the ligation class with high accuracy, it could hardly recognize the division class and the hemostasis class. On the other hand, TPF CNN showed comparable performance for the dissection and ligation classes while largely improving the accuracy for the division and hemostasis classes.

Focusing on the dissection and ligation classes, both models showed the recognition accuracy of around 90%. Besides, on the division and hemostasis classes, TPF CNN represented a higher recognition accuracy than LSTM. However, the standard deviations of division and hemostasis were larger than those of dissection and ligation in both models. The reason why the accuracy of ligation in both models exceeds 90% may be due to ligation-specific movements observed during the act of tying a string.

Overall, the average recognition accuracy of the average of all classes was higher for TPF CNN than for LSTM. Furthermore, the number of epochs of the training for LSTM was 500 epochs, while that of TPF CNN was 30 epochs. Therefore, TPF CNN showed higher recognition accuracy with fewer epochs than LSTM. However, the standard deviation of the average of TPF CNN was larger than that of LSTM. Although additional validation may be required, the training of TPF CNN is comparatively unstable.

| Class Name | Ds | Lg | Dv | Hm |
|------------|----|----|----|----|
| All        | 792| 1,201| 164| 116|
| Train1     | 528| 801| 109| 77 |
| Test1      | 264| 400| 55 | 39 |
| Train2     | 550| 955| 121| 95 |
| Test2      | 242| 246| 43 | 21 |

Table 2 Correct recognition rates of LSTM and TPF CNN (Average and standard deviation of 10 trials).

|                | LSTM          | TPF CNN       |
|----------------|---------------|---------------|
| Dissection     | 89.2% (2.73)  | 91.0% (4.50)  |
| Ligation       | 90.6% (3.53)  | 93.3% (2.55)  |
| Division       | 35.8% (10.57) | 61.5% (17.09) |
| Hemostasis     | 24.4% (9.91)  | 54.9% (15.68) |
| average        | 60.0% (3.65)  | 75.1% (7.02)  |

4.2 Leave-Videos-Out Evaluation

Compared to the real-world problem, the previous experiment (described in Sec. 4.1) could have a slightly easy experimental condition; Since we extracted each sample (that is a 10-second video) from the consecutive frames which have the same action label at an interval of five seconds, the test samples which were selected at random could be temporally adjoining or half overlapping to samples in the training set. However, in actual operating rooms, it will be necessary to recognize unlearned cholecystectomy scenes which did not exist completely at the training phase.

Hence, we evaluated the recognition performance of the proposed method in such a problem setting. We conducted leave-videos-out experiment. In our dataset, five surgeon’s stand-ins performed the simulated surgical operations four to seven times, and 21 videos were filmed in total. We here used one whole video per each stand-in to make the testing samples, and used the remaining 16 videos to make the training samples. The rows of "Train2" and "Test2" in Table 1 show the number of training samples and testing samples, respectively, that we actually made.

Tables 3 and 4 show the confusion matrix of LSTM and TPF CNN. In the matrix, the diagonal components show correct classification rate for each class. Also, non-diagonal components show misclassification rate. Even considering the standard deviation, both models can correctly recognize dissection and ligation with high accuracy. This indicates that both models can recognize the subtle differences in body movements between both surgical actions. On the other hand, when both models recognize Dv or Hm, they may mistakenly recognize the other class. In particular, LSTM is likely to recognize these classes incorrectly.

4.3 Processing Speed

Since the architecture is a straightforward 2D-CNN, our TPF CNN can run extremely fast; The TPF CNN could recognize over 1,000 samples (where each sample was a 600 × 36 matrix) per second on the PC equipping Ryzen 3900X CPU and Titan RTX GPU. On the other hand, the processing speed of OpenPose for the surgical videos depends on the video resolution and machine performance (especially GPU performance), but was approximately 5-20 fps. Therefore, if we can build a computer environment in which OpenPose runs in real time, we would be able to run our method in real time.

However, since the proposed system used 10-second videos for training, our system has to wait 10 seconds from a certain moment to start the recognition process. To solve this problem, we should use a shorter video; For example, if the TPF CNN is trained by the samples made from a one-second video, the system will be able to recognize the actual surgeon's action with a time lag of about 1-2 seconds.

5 Conclusions

We showed that the TPF CNN was more accurate than the conventional LSTM for the task of the surgical action recognition. Furthermore, TPF CNN showed higher recognition accuracy with fewer epochs than the LSTM. Our results suggest that surgeon's body movements contain much information to be required for recognizing subtle difference in several types of surgical procedures. Furthermore, we showed that the TPF CNN was adequately fast for real-time
processing.

However, since the proposed system used 10-second videos for training, we have to wait 10 seconds from a certain moment to perform the recognition process. Considering the application to actual surgery, such a time lag of the recognition process will be one of the problems we need to solve. Therefore, we would like to learn small videos in a shorter period of time (e.g., 5 seconds or 3 seconds) as a stepping stone to approach the real-time recognition.

Table 3 The confusion matrix of LSTM. The values indicate average and standard deviation of recognition rate in 10 trials.

| Class ave.: 59.9% (4.00) | predicted class |
|--------------------------|-----------------|
| Ds                       | (5.84)          |
| Lg                       | (3.34)          |
| Dv                       | (7.17)          |
| Hm                       | (15.84)         |

Table 4 The confusion matrix of TPF CNN. The values indicate average and standard deviation of recognition rate in 10 trials.

| Class ave.: 64.6% (2.44) | predicted class |
|--------------------------|-----------------|
| Ds                       | (4.28)          |
| Lg                       | (2.84)          |
| Dv                       | (2.51)          |
| Hm                       | (8.53)          |

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP18K11493.

References

[1] F. Miyawaki et al., "Scrub Nurse Robot System - Intraoperative Motion Analysis of a Scrub Nurse and Timed Automata Based Model for Surgery", IEEE Transaction on Industrial Electronics, vol. 52 (5): pp. 1227-1235, 2005.

[2] F. Miyawaki et al., "Development of Real-time Acquisition System of Intraoperative Information on Use of Surgical Instruments for Scrub Nurse Robot", Proc. of the 18th IFAC World Congress, pp. 9458-9463, 2011.

[3] K. Ohnuma et al., "Development and assessment of real-time visual recognition system for scrub nurse robot", International Journal of Assistive Robotics and Mechatronics, vol. 8, pp. 14-24, 2007.

[4] K. Yoshimitsu et al., "Prototype development of scrub nurse robot for laparoscopic surgery", Proceedings of Computer Assisted Radiology and Surgery, pp. 845-850, 2005.

[5] K. Yoshimitsu et al., "Development and evaluation of the second version of scrub nurse robot (SNR) for endoscopic and laparoscopic surgery", Proceedings of Intelligent Robots and Systems, pp. 2288-2294, 2007.

[6] Z. Cao et al., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, no. 1, pp. 172-186, 2021

[7] A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks", Adv. Neural Inf. Process Syst., pp. 1097-1105, 2012.

[8] M. Zeiler and R. Fergus "Visualizing and Understanding Convolutional Networks", ECCV 2014, Part 1, LNCS 8689, pp. 818–833, 2014.

[9] N. Loka et al., "Hilbert Vector Convolutional Neural Network: 2D Neural Network on 1D Data", ICANN 2019: Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation pp. 458-470, 2019.

[10] S. Shuai et al., "Action Classification Based on 2D Coordinates Obtained by Real-time Pose Estimation", Proc. of IWFCV2019, Feb. 2019.

[11] W Wardah et al., "Predicting protein-peptide binding sites with a deep convolutional neural network", Journal of Theoretical Biology, vol. 496, 2020.

[12] S. Hochreiter and J. Schmidhuber, "LONG SHORT-TERM MEMORY", Neural Computation, vol. 9, pp. 1735-1780, 1997.