Chen *et al. EURASIP Journal on Image and Video Processing* (2018) 2018:76
https://doi.org/10.1186/s13640-018-0316-4

EURASIP Journal on Image
and Video Processing

RESEARCH Open Access

Semi-supervised spatio-temporal CNN for recognition of surgical workflow

Yuwen Chen[1,2,3], Qi Long Sun[3] and Kunhua Zhong[1,2,3*]

Abstract

Robust and automated surgical workflow detection in real time is a core component of the future intelligent operating room. Based on this technology, it can help medical staff to automate and intelligently complete many routine activities during surgery. Recognition of surgical workflow based on traditional pattern recognition methods requires a large number of labeled surgical video data. However, the labeled surgical video data requires expert knowledge and it is difficult and time consuming to collect a sufficient number of labeled surgical video data in the medical field. Therefore, this paper proposes a semi-supervised spatio-temporal convolutional network for the recognition of surgical workflow based on convolutional neural networks and temporal-recursive networks. Firstly, we build a spatial convolutional extraction feature network based on unsupervised generative adversarial learning. Then, we build a bridge between low-level surgical video features and high-level surgical workflow semantics based on an unsupervised temporal-ordered network learning approach. Finally, we use the semi-supervised learning method to integrate the spatial model and the temporal model to fine-tune the network, and realize the intelligent recognition of the surgical workflow at a low cost to efficiently determine the progress of the surgical workflow. We performed some experiments for validating the mode based on m2cai16-workflow dataset. It shows that the proposed model can effectively extract the surgical feature and determine the surgical workflow. The Jaccard score of the model reaches 71.3%, and the accuracy of the model reaches 85.8%.

Keywords: Semi-supervised, Surgical workflow, CNN

1 Introduction

According to the Statistical Yearbook for health and family planning in China [1], in 2016, the total number of patients treated by Chinese medical and health institutions was 7.932 billion, of which 227 million were hospitalized and 50.822 million were inpatients, among which the mortality rate was 0.4%. See Fig. 1 for the detailed data.

With the development of modern artificial intelligence medical technology, whether scientific and technological progress can be used to improve the efficiency of surgery. In computer-aided surgery (CAS), intelligent recognition of surgical procedures is an important issue in recent years, which has attracted widespread attention from researchers in the field of computer vision [2]. In recent years, operating rooms (ORs) have undergone tremendous changes with the increase of available technology to support and assist surgical teams. One of the targeted goals is the development of context-aware systems [3] that continuously monitor the activities performed in the ORs in order to provide an accurate and reliable support. The key challenge in developing these new methods is to process the data coming from sensors and real-time detection systems in order to provide useful information and support decision-making. This task is challenging because of the complexity of the ORs' environment and the high variability of surgical interventions due to patient abnormalities, surgeon experience, and ORs' specific constraints. But it is very important, because it is the basis for realizing the intelligence of surgery and related activities [4]. For example, based on this basic technology, it can achieve surgical task detection [5], early warning of critical events [6], surgeon skills assessment [7], automatic indexing of surgical video [8], automatic generation of surgical records [9], surgery remaining time estimate [10], and so on. Intelligent perception of surgeons' behavior and accurate and efficient determination of the progress of the

---

* Correspondence: qingchens7@sina.com
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Chengdu Information Technology of Chinese Academy of Sciences COLTD, Chengdu, China
Full list of author information is available at the end of the article

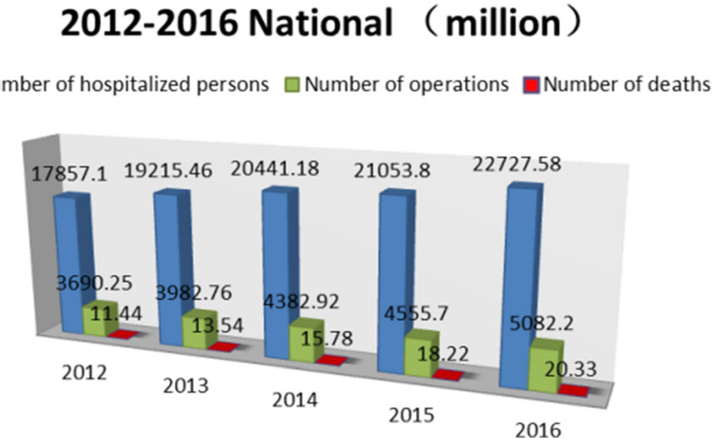


Fig. 1 2013–2017 national medical service data

operation process can effectively reduce the risk of surgery, improve the efficiency of the doctor's operation, and better save the lives of patients. Many methods have been proposed to solve the problem of automatic recognition of surgical procedures. In [11–14], the authors use instruments and sensor data directly to recognition of surgical activities. However, these methods require some special sensors and usually connected to a surgical instrument or a surgeon's hand, which may interfere with the normal operation of the operation. In the literature [15], surgical workflow was identified by integrating surgical instruments, anatomical organs, and surgical behavior. However, these features require manual design and cannot adapt to different procedures. Recently, the deep convolution neural network (DCNN) has made historic progress in the computer vision problem of image classification [16] and semantic segmentation [17], using deep learning to identify the process workflow [18, 19]. Although avoiding feature engineering and achieving good results, training the model requires a large number of training datasets, which requires a lot of manpower and material resources to pre-label the surgical data. Surgical video is more massive unlabeled data and a small number of labeled data exist simultaneously. Therefore, how to automate the recognition of surgical procedures by using only a small amount of labeled data and a large amount of unlabeled data is particularly important.

Therefore, this paper proposes a semi-supervised spatio-temporal convolution network for recognition of surgical workflow, taking laparoscopic cholecystectomy surgical video data as the research object, based on deep learning theory spatio-temporal convolution network model as the research foundation, and adopting unsupervised generative adversarial network learning methods to non-structured surgical video data structured to construct a spatial convolution feature network, using an unsupervised temporal recursive network learning approach to construct a bridge between low-level surgical video features and semantics of high-level surgical procedures. We try to achieve intelligent detection of surgical video processes at a low cost. The model is shown in Fig. 2 and described in detail below. The result of the recognition is shown in Fig. 3.

2 Review of related work

Within this field, the development of new methods for analyzing procedures is an important issue. Many researches have been conducted for developing methods for recognition of surgical workflow. Bardram et al. [20] proposed a system using embedded and body-worn sensor data to train a decision tree in order to predict surgical phases. They studied sensor significance in order to identity the most important features for surgical phase prediction. Stauder et al. [21] used Random Forest (i.e., a bag of decision trees) to predict surgical phases from sensors measurement. Other models like hidden Markov model (HMM) were also considered by Padoy et al. [22, 23] for online recognition of surgical steps. In this work, surgical activities were extracted using image processing techniques on laparoscopic camera. Similarly, Bouarfa et al. [24] used HMM with a pre-processing on the input sensor data in order to improve the detection of high-level surgical tasks. SVM classifier was also considered by Lalys et al. [25] to detect phases and low-level surgical tasks using cameras in pituitary surgery. Varadarajan et al. [26] used HMM to recognize and segment surgical gestures for surgical assessment and training. Learning the topology of an HMM is however still challenging and improving this step continues to be investigated [27].

This paper is mainly based on video analysis and understanding of surgical procedure. The following is a brief review of progress related to this field. In [28], Padoy roughly classified laparoscopic cholecystectomy into six stages and use evolutionary reinforcement

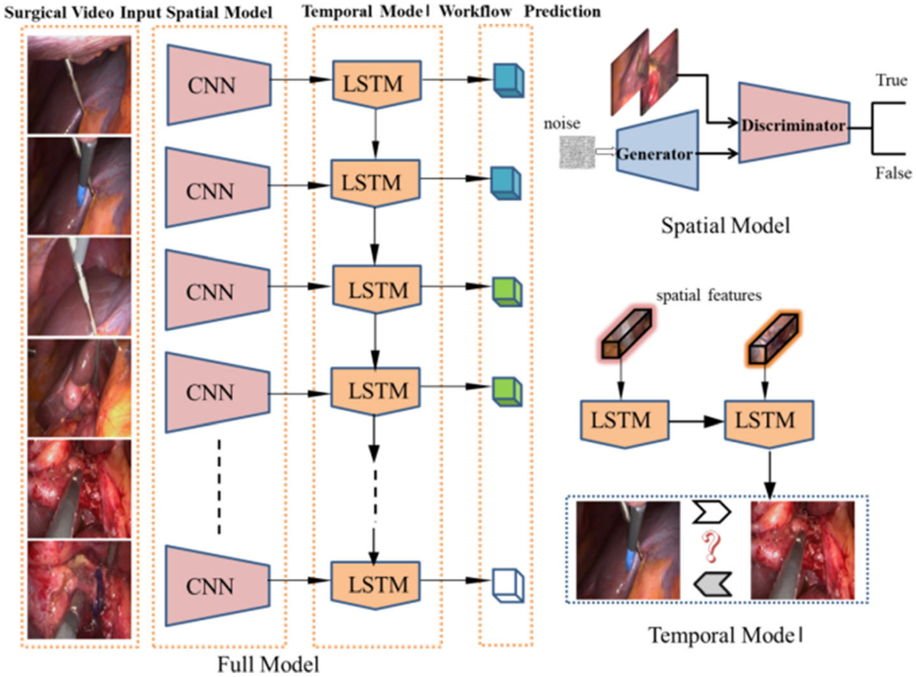


Fig. 2 (Left) Our model contains three components. The spatial, temporal, and fine-tune. (Top right) The spatial component of our model. (Bottom right) The temporal component of our model

learning to extract feature for recognition for the first time. The accuracy was about 50%. Blum et al. [29] of the Technical University of Munich, Germany, took pictures from the video of laparoscopic surgery and adopted dimensionality-reduced features, based on hidden Markov model (HMM), dynamic time warping algorithm (DTW), and other methods for phase detection. DTW algorithm produces the best performance detection accuracy 76.8%. Dergachyova et al. [30] based on the dataset of laparoscopic cholecystectomy [2] combined surgical instrument data to detect surgical procedures. This method firstly models the surgical process, performs feature extraction on visual and surgical instruments, classifies the features using AdaBoost, and finally generates the decision using a hidden Markov model. Based on the visual features, the accuracy of the algorithm is close to 68%, and the accuracy of the fusion surgical instruments is close to 90%. Recent studies [2] proposed the Endonet framework, a CNN based on the AlexNet architecture, to identify the online and offline learning processes. This method is still based on laparoscopic cholecystectomy performed on two large datasets (Cholec 80 and EndoVis) and achieves better performance. Offline analysis has the highest average accuracy of 92.2% (Cholec80) and 86% (EndoVis).

The surgical procedure detection based on the supervised learning method described above needs to learn

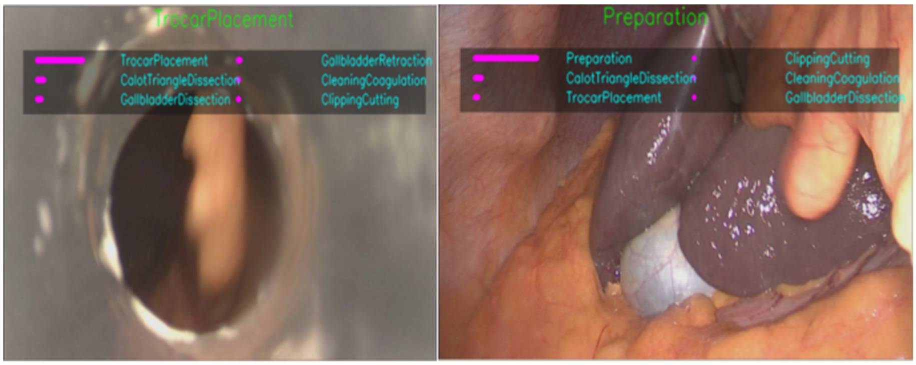


Fig. 3 This figure shows the experimental results of the model proposed in this paper. The ground truth displayed on top of the picture, and the length of the small rectangular box indicate the probability of recognition for each phase

from a large amount of data. In order to train this method, a large amount of labeled surgical video data is required. However, the labeling of surgical video data requires expert knowledge and it is difficult and time consuming to collect a sufficient number of labeled surgical video data in the medical field. Therefore, we propose a semi-supervised method for surgical workflow. The rest of the paper is composed of Section 2 where we present our semi-supervised spatio-temporal CNN method, Section 3 where we evaluate the proposed method, and Section 4 which concludes this study and gives the directions for the future work.

3 Method

This section describes the method proposed for surgical workflow recognition. This paper attempts to adopt an unsupervised feature extraction method to achieve automatic recognition of the surgical process phase in order to efficiently determine the progress of surgery, reduce the risk of surgery, and provide the core algorithms and techniques for computer-assisted surgical systems. The proposed model consists of unsupervised spatial feature extraction and temporal feature extraction. Finally, the full model is completed by merging two parts. Each section is described in detail as follows.

3.1 Unsupervised spatio generative adversarial learning

Generative adversarial networks (GANs) [2] a clever new way to leverage the power of discriminative models to get good generative models. At their heart, GANs rely on the idea that a data generator is good if people cannot tell fake data apart from real data. There are two pieces to GANs. First off, the method needs a network that might potentially be able to generate data that looks just like the real thing. The authors call this the generator network. The second component is the discriminator network. It attempts to distinguish fake and real data from each other. Both networks are in competition with each other. The generator network attempts to fool the discriminator network. At that point, the discriminator network adapts to the new fake data. This information, in turn is used to improve the generator network, and so on.

Generator maps vectors z from the noise space N^z with a known distribution P_z to the image space N^x.The generator's goal is to model the distribution P_{data} of the image space N^x (in this work, P_{data} is the distribution of all possible surgical workflow images).

Discriminator The discriminator's goal is to distinguish real surgical workflow images coming from the image distribution P_{data} and synthetic images produced by the generator.

In short, there are two optimization problems running simultaneously, and the author adjust parameters for G to minimize log(1 -D (G (z)) and adjust parameters for D to minimize log (D (X)),as if they are following the two-player min-max game with value function V (G, D). See Fig. 4 in detail. The loss function of the model contains two parts. Based on the backpropagation algorithm, the model is continuously optimized by updating z.

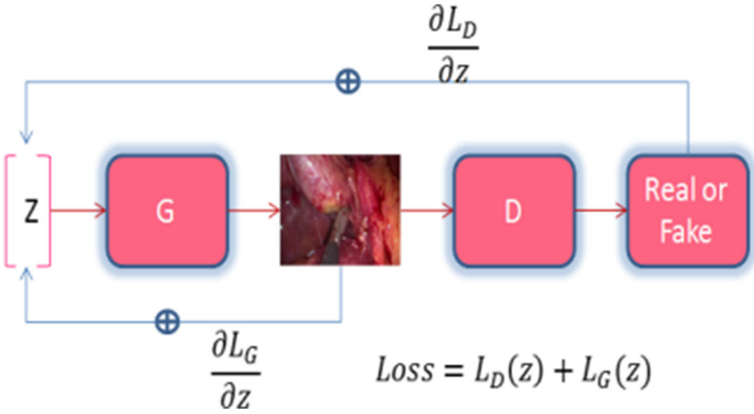
$$\min_{G} \max_{D} V(D,G) = E_{x\sim pdata}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

The recognition of surgical workflow based on deep learning requires a large number of labeled data. However, surgical video is more massive unlabeled data and a small number of labeled data. By generative adversarial networks, we can not only effectively utilize a large number of unlabeled surgical video to pre-train the models, but also generate surgical video samples [31]. This paper draws on the idea of unsupervised generative adversarial networks, trains a generative network based on surgical data sets, and then uses the discriminant network as the spatial feature extraction model in this paper. The spatial feature of the surgical workflow is mainly the image feature of the surgical video. By training the generative adversarial networks, the discriminant network can effectively understand the image features of the surgical workflow, so the network can be used as spatial feature extraction for surgical workflow. Our implementation is mainly learned from the [31], and details are shown in the experimental part of this paper.

3.2 Unsupervised temporal context learning

The temporal information is very important for recognition of surgical workflow; in order to distinguish between different phases of surgery, we need to combine context to make logical decisions. Surgery has a relatively stable sequence in logic. An operation process must precede or follow a certain surgical process phase. In this section, we present our method for training unsupervised temporal model using unlabeled videos. We accomplish this by solving a task that requires the long short-term memory (LSTM) [32] to sort two given frames into the correct temporal order. LSTM is a temporal-recursive neural network, which is suitable for processing and predicting time series tasks. We assume that the features learned while solving the sorting task enable the LSTM to distinguish frames based on their temporal context. See Fig. 5.

The long short-term memory network can capture temporal information well, and the surgical workflow is a logical process with temporal sequence. See Fig. 5. Given two surgical video frames, firstly, we extract the spatial feature from spatial model according to the sequence and then send them to the LSTM network for sorting tasks and finally enter a sigmoid function to determine. If the order is



$\frac{\partial L_D}{\partial z}$
z
G
D
Real or Fake
$\frac{\partial L_G}{\partial z}$
$Loss = L_D(z) + L_G(z)$

Fig. 4 Given a GAN model trained on real images, update z to find the closest mapping on the latent image manifold, based on the designed loss functions

true, otherwise, it is false. The network identifies the surgical workflow by the sequence of the operation, which is judging the correctness of the operation process through the network and training the model. The process is performed in an unsupervised manner as well as the extraction of surgical spatial features. Training process is generated by unlabeled surgical video sequences, and the details are shown in the experimental section.

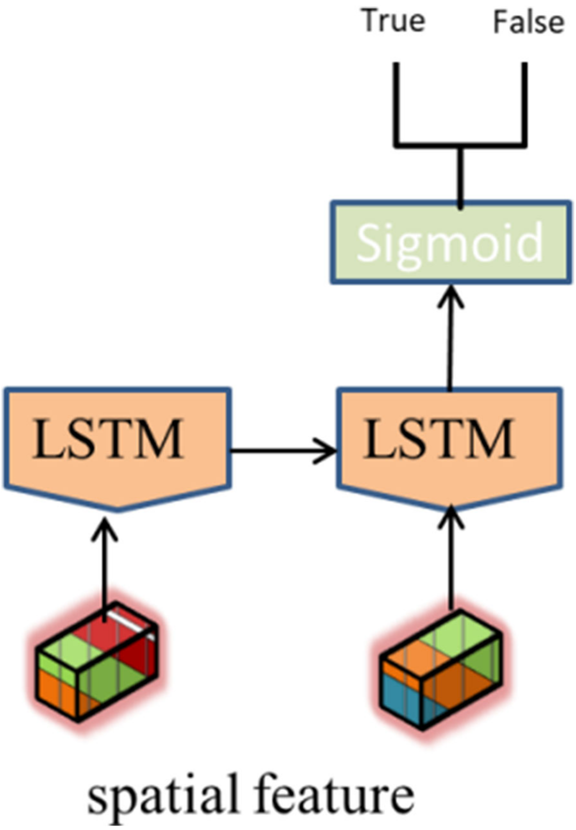


Fig. 5 Temporal model

3.3 Fine-tuning model

Surgical video can be well represented by the extraction model of spatial and temporal features. The fusion of two features can be a good classification of the surgical video. In this section, we combine spatial model and temporal model to fine-tune the surgical phase recognition. Each type of surgery can be decomposed into different stages according to the granularity. Each stage is in a different state of operation and doctors treat patients differently. The recognition of intelligent surgical workflow is based on the analysis of surgical video to determine which stage the operation is in. Therefore, the recognition of surgical workflow can be regarded as a multiple classification problems. After the extraction of spatial and temporal features, multiple classifications are performed. We use the softmax multinomial logistic function, which is an extension of the cross-entropy function, to compute the loss. The function is formulated as:

$$L = \frac{-1}{N_i} \sum_{i=1}^{N_i} \sum_{p=1}^{N_p} l_p^i \log\left(\phi\left(w_p^i\right)\right)$$

where $p \in \{1, , , , , , N_p\}$ is the phase index and $N_p = 8$ is the number of phases, $l_p^i \in \{0, 1\}$ and w_p^i are respectively the ground truth of the phases and the output of layer of the model corresponding to phase p and image i, and $\phi(\cdot) \in [0, 1]$ is the softmax function.

Surgical video representation is extracted by pre-trained spatial and temporal feature model, and the features are input into multi-classification functions to identify and judge the surgical workflow. This part integrates surgical video feature extraction and recognition to fine-tune training. The specific process is shown in Fig. 1, and the details of implementation are shown in the experimental section.

4 Experimental results and discussion

4.1 Dataset

The experiment in this paper is based on m2cai16-workflow dataset. It contains 41 videos of cholecystectomy procedures from University Hospital of Strasbourg/IRCAD (Strasbourg, France) and Hospital Klinikum Rechts der Isar (Munich, Germany). The dataset is split into two parts: training subset (containing 27 videos) and testing subset (14 videos). The videos are recorded at 25 fps. All the frames are fully annotated with eight defined phases: (1) trocarplacement, (2) preparation, (3) calot triangle dissection, (4) clipping and cutting, (5) gallbladder dissection, (6) gallbladder packaging, (7) cleaning and coagulation, and (8) gallbladder retraction. The list of phases in the dataset is shown in Table 1. The distribution of the phases in dataset is shown in Fig. 6.

4.2 Evaluation metrics

There are many different stages in each surgical workflow and each stage is continuous, so we use the Jaccard score to evaluate the recognition model, which is computed as follows:

$$J(GT,P) = \frac{GTI \cap P}{GTY \cup P}$$

where GT and P are respectively the ground truth and prediction for each phase. There are eight stages in our paper. In addition to that, we will also show the accuracy and recall of the methods. The accuracy is the percentage of correct samples in a process stage is the percentage of positive samples. The recall is the percentage of correct samples in a process stage and is the percentage of all positive samples.

4.3 Experiments

Firstly, we train the spatial model and downsampled the original 25 fps video into 1 fps and resized them into the resolution of 64×64. The images were further augmented with cropping and mirroring before input to the model. And normalize all pixel values to the [0,1] range. Paper use binary cross entropy and L1 loss as loss functions. L1 loss can be used to capture low frequencies in images. The model was trained using Adam optimizer with mini-batches of size 32 and initialized the network's parameters by sampling from a normal distribution with standard deviation 0.02 and then train the temporal model through the unsupervised sorting task. The model was trained using RMSprop optimizer with mini-batches of size 32 and initialized the network's parameters by sampling from a normal distribution with standard deviation 0.01 based on spatial feature and finally integrated the space model and the temporal model to fine tune the overall model, which used SGD optimizer and fixed space model and temporal model weights.

Table 1 List of phases in the dataset

ID	Phase
P0	Trocar placement
P1	Preparation
P2	Calot triangle dissection
P3	Clipping and cutting
P4	Gallbladder dissection
P5	Gallbladder packaging
P6	Cleaning and coagulation
P7	Gallbladder retraction

In order to verify the validity of the model, we are going to compare the performances of the following networks based on datasets. The relevant experimental results are shown in Table 2 and Fig. 7

cnn-lstm-net: the model does not use unsupervised learning for pre-training, but direct supervision training and the network structure of the model is the same as the spatio-temporal model
Spatial-net: spatial model we proposed in this paper
spatio-tempora-Net: full model we proposed in this paper

We show the surgical recognition results in Table 2 and Fig. 7. From Table 2, it can be seen that the spatio-tempora-Net yield significantly better results than cnn-lstm-net and spatial-net. The Jaccard score of the model reaches 71.3%, and the accuracy of the model reaches 85.8%. This shows that our unsupervised pre-training method is effective. It also shows that through unsupervised spatial feature learning, the model can learn the spatial pixel-level features of surgical videos. Through unsupervised temporal feature learning, the model can learn the temporal features of surgical procedures. There is a decrease in performance for preparation and cleaning and coagulation phase. This might be due to the fact that these two phases have the smallest amount of training data (see Fig. 4), as it only appears shortly in the surgeries.

From Fig. 7, it can be observed that through unsupervised learning, the fine-tuning model converges faster and after the model iterates 10 times, the model starts to converge. This also shows the effectiveness of unsupervised pre-training on the model. To illustrate the effectiveness of the space model, we generated a surgical video image using a generative network; as shown in Fig. 8, the generators can generate clear picture on the datasets, which shows that our model is feasible.

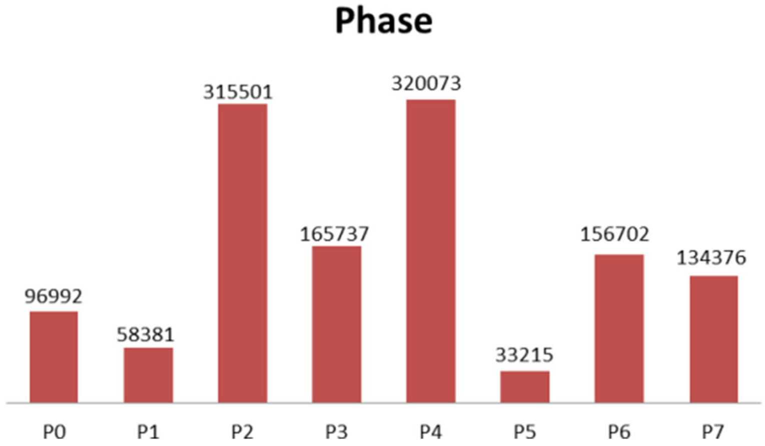


Fig. 6 Phase distribution (training data at 1 fps)

Through the experiment results, we can draw the following conclusions: Unsupervised pre-training can greatly improve the recognition effect of the model, and it can accelerate the convergence of the model; unlabeled surgical video data can be effectively utilized through unsupervised learning; recognition of surgical workflow requires fusion of spatial and temporal features. Only spatial features cannot capture the logical features of surgery, which leads to poor recognition results.

5 Conclusions

This paper proposes a semi-supervised spatio-temporal convolutional network for the recognition of surgical workflow based on convolutional neural networks and temporal-recursive networks. Firstly, we build a spatial convolutional extraction feature network based on unsupervised generative adversarial learning. Then, we build a bridge between low-level surgical video features and high-level surgical workflow semantics based on an unsupervised temporal-ordered network learning approach. Finally, we use the semi-supervised learning method to integrate the spatial model and the temporal model to fine-tune the network, and realize the intelligent recognition of the surgical workflow at a low cost to efficiently determine the progress of the surgical workflow. We performed some experiments for validating the mode based on m2cai16-workflow dataset. It shows that the proposed model can effectively extract the surgical feature and determine the surgical workflow. The Jaccard score of the model reaches 71.3%, and the accuracy of the model reaches 85.8%. The medical surgery scene has special spatial information. The doctors, nurses, etc. in the operating room rarely move during

Table 2 Phase recognition results

Model type	cnn-lstm-net			Spatial-net			spatio-tempora-Net		
Phase	Jacc	Prec	Rec	Jacc	Prec	Rec	Jacc	Prec	Rec
P0	51.4	67.0	70.3	53.2	65.4	75.3	72.5	87.3	78.2
P1	38.9	43.8	71.4	52.3	67.7	65.3	67.1	80.3	80.2
P2	59.0	68.5	67.8	68.3	82.2	70.3	72.6	95.5	73.2
P3	57.3	62.2	73.5	62.1	76.6	76.7	72.7	83.7	78.6
P4	54.1	61.2	76.6	63.4	70.2	70.2	64.6	85.2	74.3
P5	42.1	52.8	78.3	64.7	79.3	74.5	75.9	93.0	68.4
P6	51.4	61.4	62.8	53.9	65.6	77.3	66.3	73.2	70.3
P7	61.2	59.5	67.4	79.4	73.6	74.3	71.6	83.1	76.5
Average value	52.8	60.8	72.2	64.4	73.4	72.9	71.3	85.8	74.9

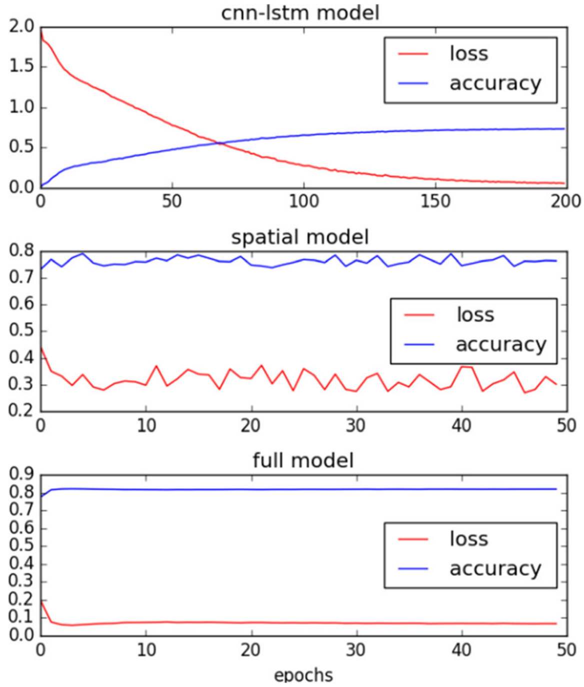


Fig. 7 Loss and accuracy of three different models

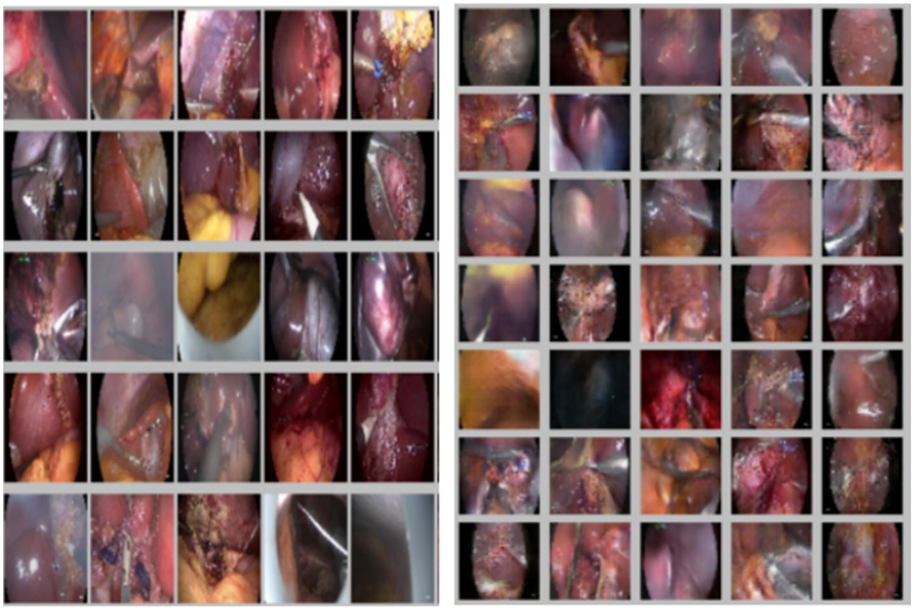


Fig. 8 (Left) Original dataset cholecystectomy surgical workflow images. (Right) Generative cholecystectomy surgical workflow images

the entire operation. The spatial model proposed in this paper is feature extraction of the entire scene, not focusing on the specific behavior of doctors and nurses. Therefore, it is not enough to capture the features of complicated movements during the surgery in subtle scenes with changing background appearance. In view of this, the future work of this paper will be based on the understanding of fine-grained surgical movements to detect surgical procedures. First, the basic operation of the surgery is decomposed, such as cutting, threading, suturing, and hooking. Then, each surgical procedure is detected and understood based on the movement. It is hoped that this method can improve the understanding of the surgical procedure and the detection accuracy. The doctors can be operated efficiently and save the patient's life time.

Abbreviations

CAS: Computer-aided surgery; DCNN: Deep convolution neural network; DTW: Dynamic time warping algorithm; GANs: Generative adversarial networks; HMM: Hidden Markov model; ORs: Operating rooms

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

This work is supported by National Defense Science and technology innovation project (No. 17-163-12-ZT-002-070-01) and Technical innovation project of Southwest Hospital (No. SWH2017ZDCX4102).

Availability of data and materials

Please contact the authors for data requests.

Authors' contributions

CYW and ZKH conceived and designed the study. CYW, SQL, and ZKH performed the experiments. CYW wrote the paper. ZKH and SQL reviewed and edited the manuscript. All authors read and approved the manuscript.

Authors' information

Chen Yuwen (1985–present) is a male, master assistant research fellow whose main research directions are computer vision and video understanding.
Sun Qilong (1984–present) is a male associate researcher and has long been engaged in the research of supercomputing application technology, data mining, machine learning, and other fields.
Zhong Kunhua (1984–present) is a male assistant researcher and has been engaged in machine learning, data mining, and statistical learning for a long time.

Ethics approval and consent to participate

Approved.

Consent for publication

Approved.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]University of Chinese Academy of Sciences, Beijing, China. [2]Chengdu Information Technology of Chinese Academy of Sciences COLTD, Chengdu, China. [3]Chongqing Institute of Green and Intelligent, CAS, Chongqing, China.

Received: 27 June 2018 Accepted: 10 August 2018

References

1. National Statistical Bureau, *China Statistical Yearbook* (China Statistics Press, Beijing, 2013–2017)
2. AP Twinanda, S Shehata, D Mutter, et al., EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
3. N Bricon-Souf, E Conchon, *Context awareness for medical applications. Medical applications of artificial intelligence*, vol 355 (2013)

4. K Cleary, HY Chung, SK Mun, in *CARS, volume 1268 of International Congress Series*. Or 2020 workshop overview: operating room of the future (2004), pp. 847–852
5. M Guggenberger, M Riegler, M Lux, in *1st ACM international workshop on human centered event understanding from multimedia*. Event Understanding in Endoscopic Surgery Videos[C]//HuEvent 2014 ACM MM (ACM, Orlando, 2014), pp. 17–22
6. N Padoy, T Blum, SA Ahmadi, H Feussner, MO Berger, N Navab, Statistical modeling and recognition of surgical workflow. Med. Image Anal. **16**(3), 632–641 (2012)
7. C Loukas, Video content analysis of surgical procedures. Surg. Endosc. **3**, 1–16 (2017)
8. K Schoeffmann, C Beecks, M Lux, et al., in *SPIE medical imaging: Image-guided procedures, robotic interventions, and modeling*. Content-based retrieval in videos from laparoscopic surgery[C]//SPIE Medical Imaging, 97861 vol. (San Diego, 2016), pp. 1–10
9. SK Agarwal, AJ et Tim Finin. Context-Aware System to Create Electronic Medical Encounter Records. PhD thesis (University of Maryland, Baltimore County, 2006), p. 10
10. I Pernek, A Ferscha, A survey of context recognition in surgery. Med. Biol. Eng. Comput. **1-6**, 2–4 (2017)
11. R Stauder, E Kayis, N Navab. Learning-based surgical workflow detection from intra-operative signals. 2017
12. JE Bardram, A Doryab, RM Jensen, et al. Phase recognition during surgical procedures using embedded and body-worn sensors. IEEE International Conference on Pervasive Computing and Communications. IEEE Comput. Soc. **8**, 45–53 (2011)
13. A Nara, C Allen, K Izumi, in D Griffith, Y Chun, D Dean, editors. *Advances in Geocomputation. Advances in Geographic Information Science*. Surgical Phase Recognition using Movement Data from Video Imagery and Location Sensor Data (Springer, Cham, 2017)
14. O Dergachyova, D Bouget, A Huaulmé, et al., Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist. Radiol. Surg. **11**(6), 1–9 (2016)
15. X Du, M Allan, A Dore, et al., Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery. Int. J. Comput. Assist. Radiol. Surg. **11**(6), 1109–1119 (2016)
16. A Krizhevsky, I Sutskever, GE Hinton, in *Advances in Neural Information Processing Systems (NIPS)*. Imagenet classification with deep convolutional neural networks (2012), pp. 1097–1105
17. J Long, E Shelhamer, T Darrell, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Fully convolutional networks for semantic segmentation (Boston, IEEE Computer Society, 2015), pp. 3431–3440
18. AP Twinanda, D Mutter, J Marescaux, et al. Single- and multi-task architecture for surgical workflow at M2CAI 2016. 2016
19. P Jannin, X Morandi, Surgical models for computer-assisted neurosurgery. Neuroimage **37**(3), 783–791 (2007)
20. JE Bardram, A Doryab, RM Jensen, PM Lange, KL Nielsen, ST Petersen, in *IEEE International Conference on Pervasive Computing and Communications*. Phase recognition during surgical procedures using embedded and body-worn sensors (2011), pp. 45–53
21. R Stauder, A Okur, L Peter, A Schneider, M Kranzfelder, H Feussner, N Navab, in *Information Processing in Computer-Assisted Interventions*. Random forests for phase detection in surgical workflow analysis (Springer, 2014), pp. 148–157
22. N Padoy, T Blum, H Feussner, MO Berger, N Navab, in *AAAI*. On-line recognition of surgical activity for monitoring in the operating room (2008), pp. 1718–1724
23. N Padoy, D Mateus, D Weinland, MO Berger, N Navab, in *IEEE International Conference on Computer VisionWorkshops*. Workflow monitoring based on 3d motion features (2009), pp. 585–592
24. L Bouarfa, PP Jonker, J Dankelman, Discovery of high-level tasks in the operating room. J. Biomed. Inform. **44**(3), 455–462 (2011)
25. F Lalys, L Riffaud, X Morandi, P Jannin, in N Navab, P Jannin, editors. *Information processing in computer-assisted interventions. IPCAI 2010. Lecture Notes in Computer Science*. Automatic phases recognition in pituitary surgeries by microscope images classification, vol 6135 (Springer, Berlin, Heidelberg, 2010), pp. 34–44
26. B Varadarajan, C Reiley, H Lin, S Khudanpur, G Hager, in GZ Yang, D Hawkes, D Rueckert, A Noble, C Taylor, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009. MICCAI 2009. Lecture Notes in Computer Science, vol 5761*. Data-derived models for segmentation with application to surgical assessment and training (Springer, Berlin, Heidelberg, 2009), pp. 426–434
27. Y Shi, A Bobick, I Essa, in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Learning temporal sequence model from partially labeled data, vol 2 (IEEE, 2006), pp. 1631–1638
28. U Klank, N Padoy, H Feussner, N Navab, Automatic feature generation in endoscopic images. Int. J. Comput. Assist. Radiol. Surg. **3**, 331–339 (2008)
29. T Blum, H Feussner, N Navab, Modeling and segmentation of surgical workflow from laparoscopic video. Lect. Notes Comput. Sci. **6363**, 400–407 (2010)
30. O Dergachyova, D Bouget, A Huaulmé, X Morandi, P Jannin, Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist. Radiol. Surg. **11**, 1081–1089 (2016)
31. Y Chen, K Zhong, F Wang, in *International conference on artificial intelligence and big data*. Surgical workflow image generation based on generative adversarial networks (China, IEEE, 2018), p. 4
32. S Hochreiter, J Schmidhuber, Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)