Automatic Step Recognition with Video and Kinematic Data for Intelligent Operating Room and Beyond

Chin-Boon Chng
National University of Singapore
Singapore
mpeccbo@nus.edu.sg

Wenjun Lin
National University of Singapore
Singapore
e0546044@u.nus.edu

Yaxin Hu
National University of Singapore
Singapore
e0576042@u.nus.edu

Yan Hu
Southern University of Science and Technology
Shenzhen, Guangdong, China
huy3@sustech.edu.cn

Jiang Liu
Southern University of Science and Technology
Shenzhen, Guangdong, China
liuj@sustech.edu.cn

Chee-Kong Chui
National University of Singapore
Singapore
mpecck@nus.edu.sg

ABSTRACT

With the continuous development of intelligent operating room systems, the segmentation and automatic recognition of surgical workflow have become challenging research fields. In recent years, an increasing number of models have been proposed to address this challenge, with deep learning becoming the mainstream approach. In this paper, we propose a multi-stage network for surgical step recognition by using surgical video and kinematic data. Firstly, a convolutional neural network (ResNet34) is used to extract visual features from video frames. Next, since surgical videos are a form of sequential data, a Temporal Convolutional Network (TCN) is employed as a temporal extractor to process temporal information between video frames for classification. Finally, a multi-stage TCN network, consisting of Encoder-Decoded TCN and Dilated TCN architectures, is used to refine the result. The proposed network is compared against a LSTM network from our prior work and is evaluated on a surgical dataset named MISAW in two modes - video data with and without kinematic data. Experimental results indicate that kinematic data is crucial for robot motion control in the operating rooms of the future. The technology will also find application in robotic labs for the development and optimization of chemical manufacturing processes.

CCS CONCEPTS

• **Computing methodologies** → **Vision for robotics**; **Activity recognition and understanding**; **Neural networks**.

KEYWORDS

Intelligent Operating Room, Step Recognition, Multi-stage Model, Temporal Convolutional Networks, Surgical Robotics, Automation

SOICT 2023, December 07–08, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12.
https://doi.org/10.1145/3628797.3628999

ACM Reference Format:
Chin-Boon Chng, Wenjun Lin, Yaxin Hu, Yan Hu, Jiang Liu, and Chee-Kong Chui. 2023. Automatic Step Recognition with Video and Kinematic Data for Intelligent Operating Room and Beyond. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 07–08, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3628797.3628999

1 INTRODUCTION

The operating room is an advanced-engineered environment that is high-risk and dynamic. The concept of the Intelligent Operating Room is a highly complex and data-rich environment due to the inclusion of numerous advanced technologies. These technologies allow surgeons to perform more complex surgical operations, increase the amount of useful information, and improve patient safety [10]. Intelligent Operating Room includes various directions such as image-guided and robotic surgical systems, augmented reality and visualization, sensing devices, and context-aware systems in computer-assisted interventions (CA-CAI) [2]. Context-aware systems are an essential component of the Intelligent Operating Room. With the increasing usage of technology, it is necessary for an Intelligent Operating Room to interact with various devices and process a large amount of information provided by these devices [22]. Context-aware computer-assisted surgical systems monitor and automatically record the entire surgical process, provide automatic and accurate assistance to the surgical staff, detect the physiological state of the patient, alert surgeons of possible surgical complications, prevent medical errors, and optimize the arrangement of the operating room and surgical staff [19].

The creation of a context-aware system requires a lot of clinical data and information. All kinds of devices involved in surgery are sources of useful information. Cameras not only capture a large amount of information but are also naturally present in many minimally invasive procedures or can be installed without disrupting the surgical workflow. Signals of other devices, such as surgical robots or sensors, can be integrated into the operating room to capture a person's position or tool usage. Due to the development of minimally invasive surgery and surgical robotic systems, video data and kinematic data have become the main sources of information.

Compared with traditional surgery, minimally invasive surgery has many advantages, such as less trauma, less pain, and faster

recovery. Minimally invasive surgery is different from traditional surgery and is typically performed with the use of a camera called an endoscope which is inserted into the body to display the surgical field. Functioning not only to display the surgical procedure being performed, recordings from the endoscope can be used as archive data for many secondary benefits. For example, the videos can be used for training junior surgeons to save the time and manpower of senior medical experts, existing as a detailed medical record of the procedure for further patient briefing and assessing the surgical skills of a surgeon [15].

Recently, surgical robots have been employed to assist surgeons in performing various complex surgical procedures in minimally invasive surgery, simultaneously enhancing the precision of surgical actions. The da Vinci surgical system stands out as a widely utilized robotic system capable of capturing extensive video, visual, and kinematic data [23].

Automatic recognition of surgical workflows represents another crucial aspect in the development of context-aware systems. Such recognition can significantly augment the cognitive understanding of the surgical process. Real-time recognition facilitates the explanation of ongoing specific activities, alerts surgeons to potential impending complications, and provides support for decision-making [5].

The surgical procedure is categorized into multiple granularity levels, as depicted in Figure 1. At the highest level lies the procedure itself, comprising a series of phases. Each phase represents primary types of events occurring during surgery and is composed of one or several steps. A step encompasses a sequence of activities aimed at achieving a surgical objective [11]. An activity corresponds to a physical action executed by the surgeon, involving an action verb describing the gesture, a target affected by the action, and an instrument employed to carry out the specific action [8].

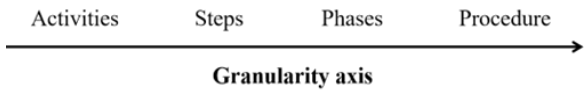

Figure 1: Different granularity levels of a surgical procedure.

Several models have been proposed for segmenting and recognizing the surgical workflow, while most of them rely on vision data only. In recent years, researchers have started to notice the role of kinematic data. An Endoscopic Vision Challenge called MICRO-Surgical Anastomose Workflow Recognition on Training Sessions (MISAW) which focuses on recognizing surgical workflow from surgical videos and kinematic data was held during the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention. We participated in the MISAW challenge and achieved 1st place in Multi Recognition and tied for 1st place in Activity Recognition with Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) network [8].

In this paper, we propose a multi-stage network that uses surgical video and kinematic data to solve the problem of surgical step recognition. Specifically, a backbone network is utilized to extract spatial features of each video frame, followed by Temporal Convolutional Neural Networks (TCN) employed as a temporal extractor to process temporal information from the extracted spatial features and kinematic data respectively. After that, a multi-stage TCN network, consisting of Encoder-Decoded TCN (ED-TCN) and Dilated TCN architectures is proposed to continuously refine the step recognition result.

In conclusion, the main contributions of this work are summarized as follows:

1. We propose a novel surgical step recognition model by using surgical videos and kinematic data. We show that the kinematic data which is important for robot motion control in an intelligent operating room can be fused with video data.
2. We propose a multi-stage TCN architecture to progressively refine step predictions from previous stages.
3. We assess the performance of the models using the MISAW dataset, evaluating the accuracy of the proposed architectures with both video and kinematic data. We illustrate the utility of a stage-stacking architecture in temporal refinement.

2 RELATED WORK

2.1 Surgical Workflow Recognition

Various kinds of information and signals can be obtained during the operation, such as the kinematics data recorded by the remote manipulator, the motion of the surgical instruments, the information contained in the surgeon's eye movements, the surgeon's position displayed by the ultrasound [16], the accelerator information worn by the surgeon, the RFID tag information [17] and so on. In early studies, many methods using this information have been used to try to identify surgical workflows, but these methods rely too much on sensors and can only identify low granularity levels [17]. Moreover, collecting these signals mostly requires additional installation or modification of equipment, which will increase the workload of surgery preparation.

Many methods were proposed based on the surgical video. Usually, the main strategy of these methods is first extracting visual features using RGB and HSV histograms, optical flow, STIP points [3], or CNN networks, then using machine learning algorithms like support vector machine (SVM), bag-of-visual-words, K-nearest neighbors, conditional random fields, or Bayesian networks [14] to classify images based on these extracted features. Due to the sequential nature of the surgical procedure, Hidden Markov Models (HMM), Dynamic Time Warping (DTW) methods or RNN/TCN networks can be used to process the temporal information [6].

There are some typical models designed using different methods. Padoy et al. [18] proposed a hidden Markov model for online recognition of surgical steps based on endoscopic video. Lalys et al. [12] proposed to use SVM classifier to detect surgical phases based on pituitary surgery videos. Twinanda et al. [21] constructed a CNN to capture the visual information in the video and used a Hierarchical Hidden Markov Model (HMM) to process the time information. Jin et al. [9] proposed an end-to-end recurrent convolutional model including a CNN network to capture frame-wise visual information and a LSTM to model the clip-wise sequential dynamics. Lea et al. [13] proposed Temporal Convolutional Network (TCN) that hierarchically captures relationships at low-, intermediate-, and high-level time scales using video or sensor data. Finally, Czempiel

et al. [4] proposed a TeCNO network which is a Multi-Stage Temporal Convolutional Network (MS-TCN) that performs hierarchical prediction refinement for surgical phase recognition.

In the MISAW challenge, we employed the approach described in [10] and extended it to include step and activity recognition. We used EfficientNet [20] as our feature extractor to extract spatial features from each video frame. Recognizing the importance of temporal information in video data, we utilized long short-term memory (LSTM) to model the sequential dependencies. The sequential features were then passed through a fully connected layer to make predictions for surgical workflow. Since the dataset includes kinematic data recorded at 30Hz from encoders mounted on the two robotic arms of the master-slave robotic platform, we hypothesized that these kinematic data are related to the verb and step. As a result, we employed a new LSTM to model the sequential features of the kinematic data. Subsequently, these two types of sequential features were concatenated and sent to fully connected layers to make predictions for the surgical verb and step.

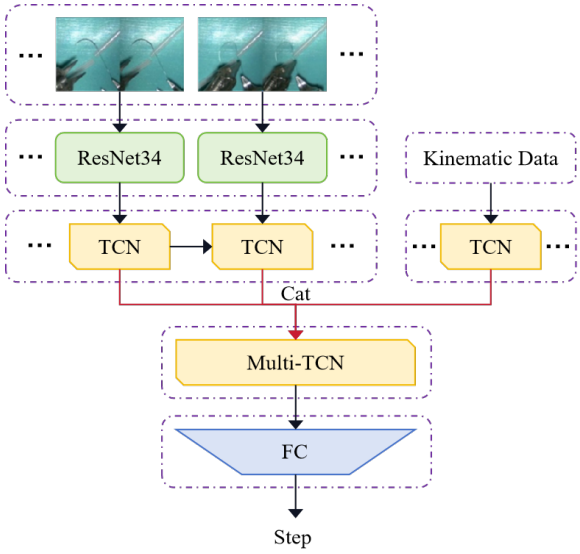

Figure 2: Multi-stage TCN.

2.2 Temporal Convolutional Network (TCN)

Recurrent Neural Network (RNN) models have demonstrated satisfactory performance across various tasks, such as speech recognition, machine translation, text recognition, and sequence prediction. Among the popular RNN architectures, the Long Short-Term Memory (LSTM) stands out, utilizing memory cells and gate functions to model long-term dependencies. However, in practical applications, RNN networks lack extensive parallel processing capabilities since they handle one time step at a time, with subsequent steps dependent on the completion of the preceding step.

In response to these limitations, Temporal Convolutional Networks (TCNs) have been introduced for time series data processing [13]. Just as Convolutional Neural Networks (CNNs) treat images as two-dimensional matrices during image processing, TCNs extend this concept to time series data, treating it as a one-dimensional vector. When processed through a multilayer network structure, TCNs achieve a large receptive field. This fundamental concept forms the basis of TCNs. Unlike RNN models, TCNs lack recurrent connections, enabling parallel sequence processing. This attribute enhances computational efficiency and reduces memory usage. Moreover, TCNs exhibit consistent gradients across all time steps due to their convolutional nature. This property empowers TCNs to learn long-term dependencies without encountering gradient explosions.

In this paper, diverging from previous methods that employ LSTM for temporal feature extraction, we explore the utilization of TCNs for surgical workflow recognition tasks. In comparison to LSTM models, TCNs offer parallel execution and stable gradients. Additionally, TCNs can flexibly adjust their receptive field size, allowing better control over model memory and adaptability to various tasks and domains. We present a multi-stage network that employs a TCN network as a temporal extractor for processing temporal information. Furthermore, we propose a multi-stage TCN network composed of Encoder-Decoder TCN (ED-TCN) and Dilated TCN architectures to refine the outcomes.

3 METHODOLOGY

In this section, we present a novel methodology that adopts a multi-stage framework to recognize surgical steps from the surgical video and kinematic data. The various components of the proposed model will be elaborated upon in the subsequent subsections.

3.1 Overall Structure

As illustrated in Figure 2, the proposed model consists of two branches - a video branch and a kinematic branch. These two branches have different inputs but share parts of the model. The proposed model contains three main networks: 1) ResNet34 is used as a feature extractor to extract spatial features of video frames; 2) two kinds of Temporal Convolutional Networks: Dilated TCN or ED-TCN is used to model the sequential dependencies; 3) a Multi-stage TCN is used to improve and refine the results.

3.2 Backbone

ResNet34 is used as the backbone network to extract spatial features of each video frame. As illustrated in Figure 3, except for the convolutional layer and max pooling at the very beginning and average pooling at the end of the network, the ResNet34 network has many similar units called Residual Block. There are two paths in the Residual Block, one learns the residual mapping function, and the other comes to identity mapping through the shortcut. Finally, the module generates outputs by summing the outputs on the two paths. This approach employed by ResNet mitigates the issue of gradient vanishing and simplifies the optimization process.

3.3 Dilated TCN

To capture long-term temporal information, TCN is utilized as a temporal feature extractor. Two kinds of TCN, Dilated TCN and Encoder-Decoder TCN are explored in this paper. Dilated TCN allows the input to be sampled at intervals during the convolution

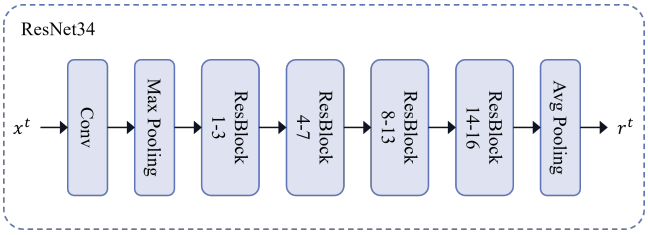

Figure 3: ResNet34 architecture.

process so that large receptive fields can be obtained using fewer convolutional layers.

The top layer of the Dilated TCN architecture is a 1 x 1 convolutional layer, which is used to adjust the spatial feature output from ResNet34 to fit the feature size in Dilated TCN. Next, a dilated convolution layer is used. Dilated convolution has a dilation factor (d), where a larger dilation factor makes the receptive field of the convolution layer larger without the problem of information loss caused by the pooling layer in ordinary convolution. This means that the convolution can handle more historical information. If the dilation factor is equal to 1, the convolution is an ordinary convolution. A zero padding of length (kernel size - 1) is used to keep the input and output length the same for each hidden layer. The receptive field can also be increased by increasing the convolution kernel size. The following expression of the receptive field of dilated convolution is used [1]:

$$\text{Receptive Field} = (\text{kernel size} - 1) * d + 1 \tag{1}$$

In the proposed network, the kernel size is set as 3, with the dilation factor set as 2^i (where i is the number of layers). Figure 4 below shows the variation of the receptive field of the three-layer dilated convolution.

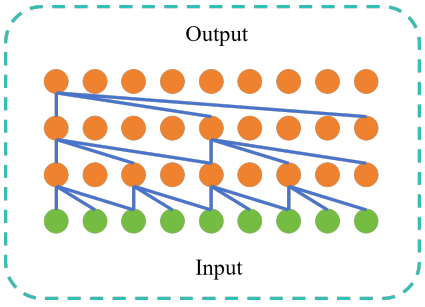

Figure 4: Receptive field of dilated convolution.

Dilated residual block is the key point that Dilated TCN can handle longer time sequences. Increasing the depth of the network can also increase the receptive field, but simply increasing the depth of the network will lead to many model training problems such as gradient explosion, gradient disappearance, or network degradation. Residual blocks can be implemented in the form of a jump-tiered connection, where the input of the cell is directly added to the output of the cell [7]. The residual block can help the forward and back propagation of information, so as to solve to some extent the problems caused by deepening the network.

As illustrated in Figure 5, Dilated residual block architecture has a dilated convolution, a ReLU activation function, a 1 x 1 convolutional layer, and a dropout layer.

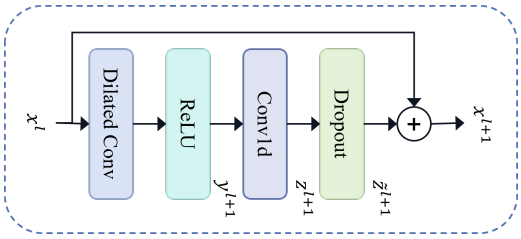

Figure 5: Dilated residual block.

Dilated residual block can be formulated as follows:

$$y_i^{l+1} = \text{ReLU}(w_i^{l+1} * x^l + b_i^{l+1}) \tag{2}$$

$$z_i^{l+1} = w_{i+1}^{l+1} * y_i^{l+1} + b_{i+1}^{l+1} \tag{3}$$

$$r_j^{l+1} \sim \text{Bernoulli}(p) \tag{4}$$

$$\tilde{z}^{l+1} = r^{l+1} * z^{l+1} \tag{5}$$

$$x_i^{l+1} = x^l + \tilde{z}^{l+1} \tag{6}$$

where x^l is the input of residual block, y^{l+1} is the output of dilated convolutional layer activated by a ReLU activation, z^{l+1} is the output of 1 x 1 convolutional layer, r^{l+1} is the vector that Bernoulli's function randomly generates 0 and 1 with probability p. $\tilde{z}^{l+1}$ is the output of dropout, x^{l+1} is the output of the residual block, w_i^{l+1}, b_i^{l+1}, w_{i+1}^{l+1}, b_{i+1}^{l+1} are the weights and bias of dilated convolutional layer and 1 x 1 convolutional layer respectively.

The overall Dilated TCN architecture is illustrated in Figure 6. The final layer of the Dilated TCN uses a 1 x 1 convolutional layer to predict class rather than a fully connected layer. This is to keep the output dimension consistent with the input dimension. In such a manner, end-to-end prediction is realized.

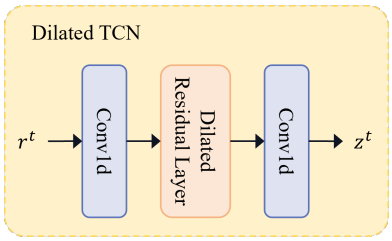

Figure 6: Dilated TCN architecture.

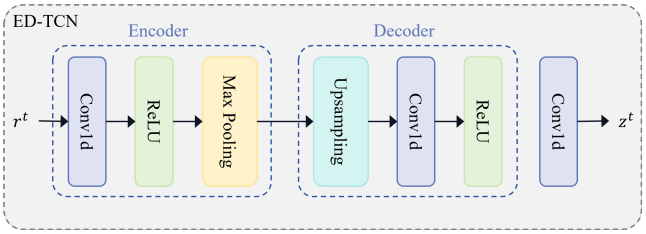

Figure 7: ED-TCN architecture.

3.4 Encoder-Decoder TCN (ED-TCN)

In contrast to the Dilated TCN, which employs dilated convolutions to capture longer temporal relationships, the ED-TCN utilizes max pooling and upsampling functions.

The overall structure of ED-TCN is illustrated in Figure 7. The Encoder of the ED-TCN consists of a 1x1 convolutional layer, a ReLU activation function layer, and a max pooling layer. The 1x1 convolutional layer is used to capture how lower-level features change over time. The max pooling layer reduces the tensor dimension across the time so as to enable the model to efficiently compute activations over longer temporal windows.

The decoder of ED-TCN is similar to the encoder except that upsampling is used first instead of max pooling to maintain the same dimension of the input. Therefore, the order of operations in the decoder of ED-TCN is upsampling, 1x1 convolution, and ReLU activation function. Finally, the last layer of ED-TCN is a 1 x 1 convolutional layer used for classification.

3.5 Multi-stage TCN

The key concept of a multi-stage architecture is to stack several stages on top of each other sequentially, with each stage taking the output of the previous stage as an input. The effect of this stage-stacking strategy is to progressively refine the predictions of previous stages.

After ResNet34, a TCN network is used. The input of the first TCN is the spatial feature extracted by ResNet34. This network contains 5-dilated-residual-block or 2-layer-encoders/decoders to extract temporal features of visual feature and kinematic data respectively. The obtained spatial-temporal visional features and kinematic features are concatenated and fed to a multi-stage model for further refinement and surgical step prediction. This enables the network to process temporal information and gradually refine the previous stage of the prediction. For each stage of the multi-TCN, both the classification result and loss function are predicted and set up separately. Finally, a different weight is set for each loss function to form the total loss function.

The following two expressions represent the prediction of the output of multi-TCN at each stage:

$$P^0 = x_{t-n}, ..., x_t \tag{7}$$

$$P^s = \text{TCNStage}(P^{s-1}) \tag{8}$$

where P^0 is the input of the first TCN stage, P^s is the output of stage s, and TCNStage(·) is a single TCN stage. Except for the first stage, the inputs of all other stages are the prediction results of the previous stage. This enables the network to capture and learn the relationships between the various classes and helps alleviate the problem of over-segmentation.

4 EXPERIMENTAL SETUP

In this section, the following details used in the evaluation of the proposed model are described – 1) the dataset; 2) the data preprocessing used to train our model; 3) the evaluation metrics used to verify the feasibility of the proposed model and 4) the parameters set for training.

4.1 Dataset

The proposed model was applied to a unique dataset for online automatic recognition of surgical workflow on a micro-anastomosis training task. The MIcro-Surgical Anastomose Workflow (MISAW) data set is composed of 27 sequences of micro-surgical anastomosis on artificial blood vessels performed by 3 surgeons and 3 engineering students. The dataset includes video data, kinematic data, and workflow annotations [8]. The kinematics and video data were acquired simultaneously at a frequency of 30 Hz. Kinematic data (which consists of x, y, z, alpha, beta, gamma, and information about the grip and the output grip voltage) is recorded at 30Hz from the encoders mounted on the two robotic arms of the master-slave robot platform. The dataset is randomly divided into 17 videos for training, with 10 videos for test/evaluation. The resolution of the video is 920x540 pixels. Workflow annotations contain labels for each timestamp phase, step, and left-handed and right-handed activities. The step labels are listed in Table 1 and an example video frame is provided in Figure 8. As adjacent frames are very similar, the dataset is subsampled to 5 fps. In this way, the training speed can be accelerated, and computational resources can be saved.

Table 1: Step labels in MISAW dataset.

Step
Needle holding
Suture making
Suture handling
1st knot
2nd knot
3rd knot

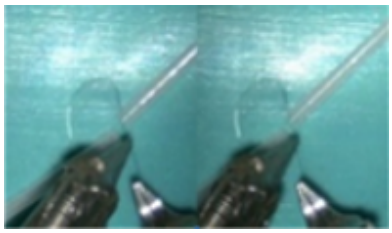

Figure 8: Example video frame in MISAW Dataset [8].

4.2 Data Preprocessing

Video frames are first extracted from the surgical videos in the MISAW dataset. Since there is little change between several successive frames, the sampling rate is set to 5 in order to accelerate the training speed. For training, the images are resized to 256×256 and randomly cropped to 224×224. For validation and testing, the images are resized to 224×224.

4.3 Evaluation Metrics

Accuracy which represents the percentage of correctly recognized frames in the video was used to evaluate the effectiveness of the proposed model.

4.4 Training Details

The network was implemented in Pytorch 1.5.1 and trained on an NVIDIA GeForce GTX 1080 Ti 11GB GPU. Cross Entropy Loss and Adam optimizer with an initial learning rate of 1e-4 for 100 epochs was used. The batch size of both the training set and the validation set was set at 128.

5 EXPERIMENTAL RESULTS

5.1 Experiments on Different Temporal Module

The result of CNN with Dilated TCN and ED-TCN was compared against the CNN+LSTM architecture used in our entry for the MISAW challenge 2020 [8]. In order to better compare the performance of TCN and LSTM, all networks used Resnet34 as the spatial feature extractor. The experimental results are shown in Table 2.

Table 2: Experiments on different temporal modules.

Model	Accuracy (%)
ResNet34 + LSTM	77.39
ResNet34 + Dilated TCN	**79.42**
ResNet34 + ED-TCN	78.05

From Table 2, it can be seen that the performance of TCN is better than that of LSTM. Dilated TCN performed best, outperforming LSTM by 2.04% in accuracy. ED-TCN outperforms LSTM by less than 1% in accuracy. This indicates that dilated convolutions in Dilated TCN can model temporal relationships with less information loss compared to the pooling and upsampling mechanism in ED-TCN.

The better performing Dilated TCN was chosen to form the Multi-Stage TCN architecture.

5.2 Experiments on Different Stages

The performances of different stages of Dilated TCN were tested and the results are shown in Table 3.

The results from Table 3 show that Multi-Stage Dilated TCN architecture can refine the results to some extent. The 2-Stage model shows the greatest improvement. However, further increasing the number of stages only leads to a marginal improvement and even a slight drop in accuracy. This may be due to the fact that the model is deeper and more difficult to train after stacking multiple layers. This outcome is also influenced by the weights assigned to each stage.

6 DISCUSSION AND CONCLUSION

This paper presents a model for surgical step recognition using both visional data and kinematic data based on a TCN network and a multi-TCN architecture. A ResNet34 network is employed to extract visual features, which are then inputted into either a Dilated TCN or an ED-TCN network. Kinematic data is also used to extract temporal features through TCN. Subsequently, a Multi-stage TCN architecture is utilized to fine-tune the results. The experimental results demonstrate the potential of the proposed network for surgical step recognition. It has been verified that an appropriate number of stage stacking can progressively refine the predictions from previous stages.

We have been extending the network to include more datasets for further evaluation of its effectiveness and making additional attempts to optimize the architecture by incorporating dynamic weight allocation of the loss function. Our aim is to incorporate the fusion of both video data and kinematic data into the network. There are three distinct levels of fusion:

1. Early Fusion: This combines vision and kinematic data at the feature level before feeding into the recognition model. This allows the model to learn joint representations.

2. Late Fusion: The vision and kinematic data are processed separately using individual models and then their outputs (decisions or scores) are combined for final recognition steps.

3. Hybrid Fusion: This technique combines features at an intermediate level, allowing the model to leverage joint representations while retaining some independence of the data sources. After the features are extracted and fused, they can be fed into the recognition model, to identify and classify the specific surgical or operational step.

Kinematic data is essential for robot motion control in an intelligent operating room. As we look into the future, robot assistance is posed to become an integral element in the operating rooms of the future. This data, being indispensable for robotic motion control, holds increased significance in advanced manufacturing environments. Consequently, such data-driven robotic systems are set to be foundational in the manufacturing ecosystems of tomorrow.

Building on our surgical foundation, robotic labs can be tailored for the pioneering and refinement of chemical and drug manufacturing processes. In this context, chemical reactions and processes are intrinsically sequential. Our step recognition algorithms, therefore, are primed to autonomously monitor and delineate each phase, ensuring unmatched production consistency. By ensuring meticulous recognition and validation of each manufacturing step, the integrity of the chemical or drug production process is maintained. Beyond pure process enhancement, step recognition also emerges as an instrumental tool for training newcomers in the industry.

Table 3: Experiments on different stages Dilated TCN.

Model	Accuracy (%)
1-Stage	79.42
2-Stage	80.25
3-Stage	**80.39**
4-Stage	80.26

Through this, robotic manipulators can be systematically aligned to the mandated sequence, guaranteeing strict procedural adherence.

In summary, the transition of our step recognition model from surgical theaters to the robotic labs of chemical and drug manufacturing could mark a paradigmatic shift in the industry. The promise lies not just in automation, but also in elevating process robustness and safety standards. With the surgical precision of step recognition in play, chemical, and pharmaceutical entities stand to gain in terms of operational efficiency, unwavering product quality, and diminished operational hazards.

Concluding on a note of emphasis, it is crucial to highlight that harnessing step recognition for the automation of drug and chemical manufacturing within the ambit of intelligent cyber-physical systems remains a central research preoccupation for the first and last authors of this paper at the National University of Singapore.

Figure 9 provides a structured overview of a cyber-physical system in the context of drug and chemical development and manufacturing:

1. Human Supervisor: It represents the human oversight element. There is a bidirectional connection between the human supervisor and the central entity labeled "Cyber." This suggests an ongoing interaction where the human supervisor provides inputs and also receives feedback or data.

2. Cyber: This is a cloud-like entity, representing a cyber system or a computational cloud platform. This system has multiple functions as indicated by the labels:

"Sensing" points towards the connection with the human supervisor, suggesting that the cyber system can recognize and understand inputs from the human supervisor.

"Analyze and Predict", "Optimize and Plan", and "Process Control" indicates the processing capabilities of this system, which aid in controlling the robot.

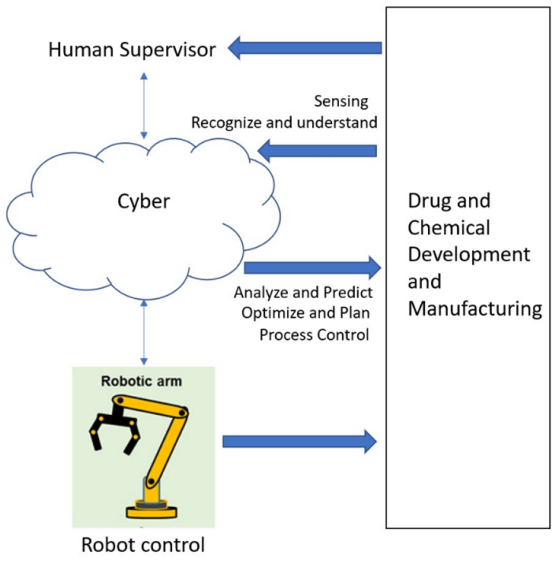

Figure 9: Robot platform for automated drug and chemical development and manufacturing.

3. Robotic Control: It emphasizes its role in physically executing tasks. The robotic arm interacts directly with the cyber system, receiving commands and potentially sending feedback.

4. Drug and Chemical Development and Manufacturing: It indicates the end application or sector where the combined efforts of the human supervisor, cyber system, and robotic arm are directed.

There are arrows from both the cyber system and the robotic arm pointing towards this block, suggesting that both computational decisions and physical actions contribute to the processes in drug and chemical development and manufacturing.

ACKNOWLEDGMENTS

The second author is supported by the Joint NUS-SUSTech PhD Program. The robotic lab work from the first and last authors is supported in part by the National Research Foundation, Singapore, and the National University of Singapore under its Competitive Research Programme (CRP) Grant (NRF-CRP25-2020RS-0002, WBS: A-0008485-02-00).

REFERENCES

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[2] Rasiah Bharathan, Rajesh Aggarwal, and Ara Darzi. 2013. Operating room of the future. *Best Practice & Research Clinical Obstetrics & Gynaecology* 27, 3 (2013), 311–322. https://doi.org/10.1016/j.bpobgyn.2012.11.003 Advances in Gynaecological Surgery.

[3] Katia Charrière, Gwénolé Quellec, Mathieu Lamard, Gouenou Coatrieux, Béatrice Cochener, and Guy Cazuguel. 2014. Automated surgical step recognition in normalized cataract surgery videos. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 4647–4650. https://doi.org/10.1109/EMBC.2014.6944660

[4] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 343–352.

[5] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery* 11 (2016), 1081–1089.

[6] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery* 11 (2016), 1081–1089.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Arnaud HuaulmÃ©, Duygu Sarikaya, KÃ©vin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-SÃ¡nchez, Pablo ArbelÃ¡ez, Wolfgang Reiter, Manoru Mitsuishi, Kanako Harada, and Pierre Jannin. 2021. MIcro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine* 212 (2021), 106452. https://doi.org/10.1016/j.cmpb.2021.106452

[9] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. 2018. SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. *IEEE Transactions on Medical Imaging* 37, 5 (2018), 1114–1126. https://doi.org/10.1109/TMI.2017.2787657

[10] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. 2020. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Analysis* 59 (2020), 101572. https://doi.org/10.1016/j.media.2019.101572

[11] Florent Lalys and Pierre Jannin. 2014. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery* 9 (2014), 495–511.

[12] Florent Lalys, Laurent Riffaud, Xavier Morandi, and Pierre Jannin. 2010. Automatic Phases Recognition in Pituitary Surgeries by Microscope Images Classification. In *Information Processing in Computer-Assisted Interventions*, Nassir Navab and Pierre Jannin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 34–44.

[13] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 47–54.

[14] Benny P. L. Lo, Ara Darzi, and Guang-Zhong Yang. 2003. Episode Classification for the Analysis of Tissue/Instrument Interaction with Multiple Visual Cues. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*, Randy E. Ellis and Terry M. Peters (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 230–237.

[15] Constantinos Loukas, Christos Varytimidis, Konstantinos Rapantzikos, and Meletios A Kanakis. 2018. Keyframe extraction from laparoscopic videos based on visual saliency detection. *Computer methods and programs in biomedicine* 165 (2018), 13–23.

[16] Atsushi Nara, Kiyoshi Izumi, Hiroshi Iseki, Takashi Suzuki, Kyojiro Nambu, and Yasuo Sakurai. 2009. Surgical workflow analysis based on staff's trajectory patterns. In *M2CAI workshop, MICCAI, London*.

[17] Thomas Neumuth and Christian Meißner. 2012. Online recognition of surgical instruments by information fusion. *International journal of computer assisted radiology and surgery* 7 (2012), 297–304.

[18] Nicolas Padoy, Diana Mateus, Daniel Weinland, Marie-Odile Berger, and Nassir Navab. 2009. Workflow monitoring based on 3d motion features. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, 585–592.

[19] Igor Pernek and Alois Ferscha. 2017. A survey of context recognition in surgery. *Medical & biological engineering & computing* 55, 10 (2017), 1719–1734.

[20] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[21] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* 36, 1 (2016), 86–97.

[22] Michael Unger, Claire Chalopin, and Thomas Neumuth. 2014. Vision-based online recognition of surgical activities. *International journal of computer assisted radiology and surgery* 9 (2014), 979–986.

[23] Beatrice van Amsterdam, Hirenkumar Nakawala, Elena De Momi, and Danail Stoyanov. 2019. Weakly supervised recognition of surgical gestures. In *2019 International conference on robotics and automation (ICRA)*. IEEE, 9565–9571.