Towards Generalizable Surgical Activity Recognition using Spatial Temporal Graph Convolutional Networks

Duygu Sarikaya · Pierre Jannin

Abstract

Purpose
Modeling and recognition of surgical activities poses an interesting research problem. Although a number of recent works studied automatic recognition of surgical activities, generalizability of these works across different tasks and different datasets remains a challenge. We introduce a modality that is robust to scene variation, based on spatial temporal graph representations of surgical tools in videos for surgical activity recognition.

Methods
To show its effectiveness, we model and recognize surgical gestures with the proposed modality. We construct spatial graphs connecting the joint pose estimations of surgical tools. Then, we connect each joint to the corresponding joint in the
consecutive frames forming inter-frame edges representing the trajectory of the joint over time. We then learn hierarchical spatial temporal graph representations using Spatial Temporal Graph Convolutional Networks (ST-GCN).

**Results**

Our experimental results show that learned spatial temporal graph representations of surgical videos perform well in surgical gesture recognition even when used individually. We experiment with the Suturing task of the JIGSAWS dataset where the chance baseline for gesture recognition is 10%. Our results demonstrate 68% average accuracy which suggests a significant improvement.

**Conclusions**

Our experimental results show that our model learns meaningful representations. These learned representations can be used either individually, in cascades or as a complementary modality in surgical activity recognition, therefore provide a benchmark. To our knowledge, our paper is the first to use spatial temporal graph representations based on pose estimations of surgical tools in surgical activity recognition.

**Keywords** Robot-Assisted Surgery · Surgical Activity Recognition · Graph Convolutional Neural Networks · Spatial Temporal Representation · Graph Representation

1 **Introduction**

Modeling and recognition of surgical activities poses an interesting research problem as the need for assistance and guidance through automation is addressed by the community. Although a number of recent works studied automatic recognition of surgical activities, generalizability of these works remain a challenge. Moreover, the need for representations with greater expressive power that we can use not
only to recognize surgical activities but also in control of autonomous systems is growing.

Fig. 1: For each video segment, we construct an undirected spatial temporal graph to form hierarchical representations of the joints over temporal sequences of frames. We construct the spatial graphs by connecting these nodes with edges according to the connectivity of the surgical tool structure. Then, for the temporal part, we connect each joint to the same joint in the consecutive frame forming inter-frame edges representing the trajectory of the joint over time.

Frame based image cues have been widely used for recognition of surgical activities. Although these studies have been tremendously successful in terms of high accuracy, a major setback is that their performances are limited to the dataset that they are modeled on and they are prone to overfitting. The generalizability across different tasks and different datasets remains a challenge. For example, placing a Tie Knot might occur during a task of Suturing on Tissue and also during the more specific and challenging task of Urethrovesical Anastomosis (UVA) that involves stitching and reconnecting two anatomical structures together. If we heavily rely on image cues of the surgical scene, these surgical activities would have very different representations. Kinematic data captured from the surgeon and patient side manipulators are also limited as the inverse kinematics of the end effectors are not sensitive enough. In order to overcome the challenge of generalizability across different tasks and different datasets, we need to define more generic representations of surgical activities that are robust to scene variation.

Pose estimation based skeleton representations suffer relatively little from the intra-class variances when compared to image cues [22]. Although pose estimation of surgical tools has been studied [19][20][21], pose-based skeleton representations
have not been used in surgical activity recognition yet. To our knowledge, our paper is the first to use these representations for surgical activity recognition.

In this paper, we introduce a modality independent of the scene, therefore robust to scene variation, based on spatial temporal graph representations of the surgical tool in surgical activity videos. To show the effectiveness of our modality, we propose to model and recognize surgical activities in surgical videos by first defining the graph representations of the surgical tools, and then learning hierarchical spatial temporal representations using Spatial Temporal Graph Convolutional Networks (ST-GCN) [1]. Figure 1 shows an overview of spatial temporal graph construction.

2 Related Work

2.1 Surgical Activity Recognition

Ahmidi et al. [2] did a comparative benchmark study on the recognition of gestures on JIGSAWS dataset [9]. In this study, in order to classify surgical gestures, three main methods are chosen: Bag of Spatio-Temporal Features (BoF), Linear Dynamical System (LDS) [3][4] and a composite Gaussian Mixture Model-Hidden Markov Model: GMM-HMM [5][6][7]. Studies that use primarily kinematic data have also been suggested. Ahmidi et al. [8] proposed using similarity metrics on the temporal model of surgical tool motion trajectories. More recent works have also been proposed using both video and kinematic data [10][11][12].

More recently, deep learning architectures have been proposed. DiPietro et al. [13] proposed using Recurrent Neural Networks (RNN) trained on kinematic data. Sarikaya et al. [14] proposed a Multi-Modal Convolutional Recurrent Neural Network architecture using video and optical flow. Although optical flow is robust to scene variation, its performance can be affected by the camera zoom and motion. Lea et al. [15] proposed Temporal Convolutional Network (TCN), that hierarchically captures temporal relationships at low, intermediate, and high-
level time-scales, and Convolutional Action Primitives for multimodal time-series of video and kinematics. Funke et al. learned 3D convolutional neural networks to capture spatiotemporal cues on video data, however 3D CNNs are known to have problems in training and they only marginally improve the frame based models [17].

2.2 Representations of Joints and Skeletons

Representations of human body joints and skeletons, and their dynamics have been widely used in human activity recognition as they are robust to illumination change and scene variation [18]. Although pose estimation of surgical tools has been studied [19][20][21], these representations have never been used for surgical activity recognition.

3 Dataset

3.1 JIGSAWS

The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [9] provides a public benchmark surgical activity dataset. In this video dataset, 8 surgeons perform 3 surgical tasks on the daVinci Surgical System (dVSS®): Suturing, Needle Passing and Knot Tying. The dataset provides 15 gesture labels $< G_1, G_2, ..., G_{15} >$, which are the smallest action units such as Reaching for needle with right hand. We performed our experiments on the Suturing task which is composed of 10 different gestures.

3.2 Annotation of Surgical Tool Poses

We labeled the surgical tool poses in a subset and then we trained a deep residual network (ResNet50) [23] to estimate the poses in the rest of the dataset. We first extracted frames, and then we clustered these frames using a simple $k$ –
means clustering algorithm based on frame similarity. We picked the 20 most
distinguishable frames based on these clusters in order to use for annotation. In
other words, for each video we labeled only 20 frames. We defined 5 joints to
capture the structure of a surgical grasper tool with respect to joints (the arm,
the joint that connects the arm and the tool, the tool and its end effectors).
We used DeepLabCut [24] to both annotate the joints and to train the ResNet
with transferred weights learned from ImageNet. Then, using the learned model,
we estimated the pose coordinates for the rest of the frames. Please note that,
we intentionally used frames from the same videos in both training and testing
in this step, for efficient labeling purposes with minimal effort. Automatic pose
estimation of surgical tools are one of the better solved problems in surgical video
understanding and they achieve high accuracy [19,20,21]. Since pose estimation is
not the focus of our paper, we opted for an efficient, quick solution.

3.3 Preprocessing videos as Input to ST-GCN

We used pose estimations and confidence scores of each joint that we collected
using our trained annotation network, to construct our spatial temporal graph
representations. So, for each frame, we have the pose estimations and the confi-
dence scores as input. We defined video segments $V_t = (v_{t15}, ..., v_{t1}, v_{t})$ of
t = 90 consequent frames at 30 fps which equals to 3 seconds. We set the gesture
label of this segment of activity as the gesture label of the frame at time $t = 90$.
We collected these video segments in a sliding window manner with a step size
of 3 frames (10 fps), and we used these segments as data input. For the initial
segments, we pad the frames to the beginning of the video segment until it reaches the size of 90 frames, by copying the first frame.

4 Material and Methods

4.0.1 Spatial Temporal Graph Construction

For each video segment of 90 consequent frames (3 seconds), we construct an undirected spatial temporal graph $G = (V, E)$ to form hierarchical representation of the joints over temporal sequences of frames. First, for each frame, we define nodes corresponding each joint. We construct the spatial graphs by connecting these nodes with edges according to the connectivity of the surgical tool structure (skeleton). Then, for the temporal part, we connect each joint to the same joint in the consecutive frames forming inter-frame edges representing the trajectory of the joint over time.

4.0.2 Spatial Temporal Graph Convolution Network

After a spatial graph based on the joints of the surgical tool and the temporal edges between corresponding joints in consecutive frames are defined, a distance-based sampling function is proposed for constructing the graph convolutional layer, which is then used to build the spatial temporal graph convolutional network (ST-GCN) \cite{1}. Multiple layers of spatial temporal graph convolutions on the neighbouring spatial and temporal nodes are applied; a process similar to the workings of Convolutional Neural Networks (CNN) assuming an image as a regular 2D grid graph. ST-GCN exploits the natural graph structure of the skeleton instead of representing them as a vector sequence.

The pose estimation and the adjacency matrix connecting the joints of the surgical tools is used as input to our ST-GCN. The ST-GCN model is composed of 9 layers of spatial temporal graph convolution operators (ST-GCN units). In order to apply convolutions, ST-GCN proposes partitioning strategies for constructing
convolution operations on graphs. We use spatial configuration partitioning where the nodes are labeled according to their distances to the surgical tool skeleton gravity center. The gravity center of the surgical tool skeleton is chosen as the average coordinate of all joints in the skeleton at a frame. According to this partitioning strategy, nodes that are spatially closer to the skeleton gravity center, compared to the node where the convolution is applied (root), have shorter distances, while nodes that are further have longer distances. Using these convolutions, hierarchical representations which capture the spatial and temporal dynamics of surgical activities are learned. Following multiple layers of graph convolutions and pooling, a soft-max layer is applied to label video segments.

5 Experiments and Evaluation

We carried out our experiments with a TITAN X (Pascal architecture) GPU and an Intel Xeon (R) CPU E5 3.50 GHz x8 with a 31.2 GiB memory. All experiments are conducted on the PyTorch deep learning framework. We trained the ST-GCN for 30 epochs with stochastic gradient descent (SGD) optimization algorithm with a base learning rate of 0.01 and then we decreased the learning rate using a step approach by diving the learning rate by 10 at every 10 epochs, we set the weight decay to 0.0005. In order to avoid overfitting, we used a random dropout with 0.5 probability. We also performed data augmentation; firstly, we performed random affine transformations which apply random combinations of different angle, translation and scaling factors on the skeleton sequences of all consequent frames. Secondly, we randomly sampled fragments from the skeleton sequences of consequent frames.

For test, we used the Leave-one-user-out (LOUO) experimentation split set which is provided by JIGSAWS. In the LOUO setup for cross-validation, there are eight folds, each one consisting of data from one of the eight subjects. We reported the average accuracy of all eight folds. We predicted the gesture label at every 3 frames that is, 10 frames per second. We compared the results of our
Table 1: We compared the results of our model with the JIGSAWS Benchmark and the more recent Convolutional Neural Network based studies.

| JIGSAWS Benchmark [2] | Average Accuracy |
|------------------------|------------------|
| GMM-HMM (kinematic)    | 73.95            |
| KSVD- SHMM (kinematic) | 73.45            |
| MsM-CRF (kinematic)    | 67.84            |
| MsM-CRF (video)        | 77.20            |
| MsM-CRF (kinematic + video) | 78.98 |
| SC-CRF (kinematic)     | 81.74            |
| SC-CRF (kinematic + video) | 81.60 |
| CNN based models (Evaluation at 10 fps) | Delay | Average Accuracy |
| S-CNN (video) [15]     | 1 s              | 74.0            |
| ST-CNN (video) [15]    | 10 s             | 77.7            |
| 2D ResNet-18 (video) [23] | 0 s | 79.5 |
| 3D CNN (K) + window (video) [16] | 3 s | 84.3 |
| ST-GCN (Evaluation at 10 fps) | Delay | Average Accuracy |
| Ours (2D joint pose estimations (X,Y coordinates)) | 0 s | 67.86 |

Saturing task of JIGSAWS dataset has the chance baseline for gesture recognition of 10% (there are 10 different gestures available). Our results demonstrate 68(67.86)% average accuracy on this dataset which suggests a significant improvement. Our experimental results show that learned spatial temporal graph representations of surgical videos are informative and they perform well in terms of recognizing low-level surgical activities (gestures) even when used individually.

6 Conclusion

Modeling and recognition of surgical activities poses an interesting research problem as the need for assistance and guidance through automation is addressed by the community. Although a number of recent works studied automatic recognition of surgical activities, generalizability of these works across different tasks and different datasets remains a challenge. In order to overcome the challenge of gen-
eralizability across different tasks and different datasets, we need to define generic representations of surgical activities that are robust to scene variation. Pose-based joint and skeleton representations suffer relatively little from the intra-class variances when compared to image cues [22].

In this paper, we introduced a modality independent of the scene, therefore robust to scene variation, based on spatial temporal graph representations of surgical tool structure and joints. To our knowledge, our paper is the first to use spatial temporal graph representations based on pose estimations of surgical tools for surgical activity recognition. To show the effectiveness of the modality we introduce, we modeled and recognized surgical activities in videos using this modality. We first constructed a spatial temporal graph of surgical tool joints representing the surgical tool skeleton. We then learn hierarchical temporal relationships between the joints over time using Spatial Temporal Graph Convolutional Networks (ST-GCN) [1] which exploits the natural graph structure of skeleton data and the structural connectivities of joints.

Our experimental results show that learned spatial temporal graph representations of surgical videos are informative and they perform well in terms of recognizing low-level surgical gestures even when used individually. We experiment our model on the “Suturing” task of the JIGSAWS dataset where the chance baseline for gesture recognition is 10% (there are 10 different gestures available). Our results demonstrate 68% average accuracy on this dataset which suggests a significant improvement. These learned representations can be used either individually, in cascades or as a complementary modality in surgical activity recognition, therefore provide a benchmark for future studies. Moreover, the expressive power of these graph representations can potentially be coupled in control of autonomous systems bridging the gap between recognition and control.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human partici-
pants or animals performed by any of the authors.

**Informed consent** This articles does not contain patient data.

**Acknowledgements** This work was supported by French state funds managed within the Investissements d’Avenir program by BPI France (project CONDOR).

**References**

1. Yan S, Xiong Y, Lin D (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, The Association for the Advancement of Artificial Intelligence (AAAI)

2. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD (2017) A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. Transaction of Biomedical Engineering

3. Lin HC, Shafran I, Murphy TE, Okamura AM, Yuh DD, Hager GD (2005) Automatic detection and segmentation of robot-assisted surgical motions. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 3749:802-810

4. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery 11:220-230

5. Leong JJH, Nicolaou M, Atallah L, Mylonas GP, Darzi AW, Yang GZ (2006) HMM assessment of quality of movement trajectory in laparoscopic surgery. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 4190:752-759

6. Yang GZ, Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager G (2009) Data-derived models for segmentation with application to surgical assessment and training. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 5761:426-434

7. Balakrishnan Varadarajan (2011) Learning and inference algorithms for dynamical system models of dextrous motion. PhD thesis Johns Hopkins University

8. Ahmidi N, Gao Y, Vedula SS, Khudanpur S, Vidal R, Hager GD (2013) String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery. Medical Image Computing and Computer-Assisted Interventions (MICCAI), 26-33

9. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Bejar B, Yuh DD, Chen CCG, Vidal R, Khudanpur S, Hager GD (2014) The JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. In: Proc. Modeling Monitor. Comput. Assist. Interventions (MCAI)

10. Tao L, Zappella L, Hager GD, Vidal R (2013) Surgical Gesture Segmentation and Recognition, Medical Image Computing and Computer-Assisted Interventions (MICCAI), 8151:339-346
11. Zappella L, Benjamín B, Hager GD, Vidal R (2013) Surgical gesture classification from video and kinematic data, Medical Image Analysis 7(17):732-745
12. Lea C, Hager GD, Vidal R (2015) An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks, IEEE Winter Conference on Applications of Computer Vision, 1123-1129
13. DiPietro R, Lea C, Malpani A, Ahmidi N, Vedula SS, Lee GI, Lee MR, Hager HG (2016) Recognizing surgical activities with recurrent neural networks. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 551-558
14. Sarikaya D, Khurshid AG, Corso JJ (2018) Joint Surgical Gesture and Task Classification with Multi-Task and Multimodal Learning, arXiv(cs.CV):1805.00721
15. Lea C, Vidal R, Reiter A, Hager GD, Hua G, Jégou H (2016) Temporal Convolutional Networks: A Unified Approach to Action Segmentation, European Conference on Computer Vision (ECCV) Workshops, 47-54
16. Funke I, Bodenstedt S, Oehme F, von Bechtolsheim F, Weitz J, Speidel S (2019) Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video, International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI), 467-475
17. Wang H. and Schmid C, Action recognition with improved trajectories (2013) International Conference on Computer Vision
18. Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R (2011) Real-time human pose recognition in parts from single depth images, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)
19. Bouget D, Allan M, Stoyanov D, Jannin P (2017) Vision-based and marker-less surgical tool detection and tracking: a review of the literature, Medical image analysis, 35:633-654
20. Kurmann T, Marquez-Neila P, Du X, Fua P, Stoyanov D, Wolf S, Sznitman R (2017) Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery, CoRR abs/1710.06668, arXiv:1710.06668
21. Rieke N, Tan DJ, Alsheakhali M, Tombari F, di San Filippo CA, Belagianiss V, Eslami A, Navab N (2015) Surgical Tool Tracking and Pose Estimation in Retinal Microsurgery, Medical Image Computing and Computer-Assisted Interventions (MICCAI), 266-273
22. Yao A, Gall J, Van Gool L (2012) Coupled action recognition and pose estimation from multiple views, International Journal of Computer Vision (IJCV), 100(1):16-37
23. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition, CoRR abs/1512.03385, arXiv:1512.03385
24. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M (2018) DeepLabCut: markerless pose estimation of user-defined body parts with deep learning, Nature Neuroscience