Abed Soleymani, Ali Akbar Sadat Asl, Mojtaba Yeganejou, Scott Dick, Mahdi Tavakoli, Xingyu Li, Surgical Skill Evaluation From Robot-Assisted Surgery Recordings, International Symposium on Medical Robotics, Atlanta, USA, 2021.

Surgical Skill Evaluation From Robot-Assisted Surgery Recordings

Abed Soleymani†, Ali Akbar Sadat Asl†, Mojtaba Yeganejou, Scott Dick, Mahdi Tavakoli, and Xingyu Li
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
{zsoleymani, sadatasl, yeganejo, sdick, mahdi.tavakoli, xingyu}@ualberta.ca

***Abstract*— Quality and safety are critical elements in the performance of surgeries. Therefore, surgical trainees need to obtain the required degrees of expertise before operating on patients. Conventionally, a trainee's performance is evaluated by qualitative methods that are time-consuming and prone to bias. Using autonomous and quantitative surgical skill assessment improves the consistency, repeatability, and reliability of the evaluation. To this end, this paper proposes a video-based deep learning framework for surgical skill assessment. By incorporating prior knowledge on surgeon's activity in the system design, we decompose the complex task of spatio-temporal representation learning from video recordings into two independent, relatively-simple learning processes, which greatly reduces the model size. We evaluate the proposed framework using the publicly available JIGSAWS robotic surgery dataset and demonstrate its capability to learn the underlying features of surgical maneuvers and the dynamic interplay between sequences of actions effectively. The skill level classification accuracy of 97.27% on the public dataset demonstrates the superiority of the proposed model over prior video-based skill assessment methods. The code of this paper will be available on Github at link: `sourceCode`.**

I. Introduction

Quality and safety are reckoned as critical factors in performing surgeries. Surgical trainees need to acquire desirable degrees of proficiency before operating on patients. A lack of proper preparation may have a detrimental effect on the clinical outcome. To assist trainees in surgical skill acquisition, effective training and reliable methods to assess surgical skills are critical.

Conventionally, the trainee performance has been assessed via outcome-based analysis, structured checklists, and rating scales [1]. For example, Martin et al. developed surgical skill evaluations for surgical residents called Objective Structured Assessment of Technical Skill (OSATS), under which the performance was assessed using operation-specific checklists, systematic global ranking forms, and pass/fail judgments of a trainee [2]. The Global Evaluative Assessment of Robotic Skills (GEARS) was proposed by Goh et al. to identify levels of robotic surgery expertise [3]. Qualitative assessment methods like OSATS and GEARS require a large extent of expert monitoring and manual ratings and can be unreliable because of bias and variability in human interpretation. Although crowd-sourced assessment methods of technical skills proved that even a naive person can rate the skill level of a surgical operation with acceptable accuracy [4], these methods seem unreliable for the delicate task of surgical skill assessment due to the vital importance of patients' safety in the critical care procedures. Because of the challenges in minimally invasive surgeries, traditional approaches are no longer sufficient, given the increasing attention to the efficacy of assessment and targeted feedback [5]. Therefore, it can be a rational idea to automate the procedure of surgical skill assessment. In addition to saving time and money, inexperienced surgeons can train effectively with less dependence on a human supervisor using a surgical simulator that is fitted with automated evaluation and feedback properties [6].

With the advent of surgical robot technologies, recording video and kinematics data (i.e., position, velocity, rotation of the robotic joints, etc.) has become available. This large amount of data enables artificial intelligence (AI) based systems to be deployed and utilized in surgical and medical practices. For one thing, AI technology, especially deep learning (DL), can retrieve high-level information from raw data; for the other, the constructed AI models facilitate user training by evaluating and providing feedback to users' expertise levels. In particular, for the surgical skill assessment task, early studies usually follow the statistical machine learning paradigm, where numerical features are manually designed. Recently, various deep learning based models have been proposed and reported very promising results. We briefly discuss pron and cons of prior arts in the next section.

In this paper, we present a DL architecture for video-based surgical skill evaluation. By incorporating prior knowledge on human activity into our system design, we successfully decompose spatio-temporal feature learning into two phases: transfer-learning based on intra-frame local feature extraction and an end-to-end inter-frame temporal feature learning model. This decomposition greatly reduces the model complexity. By effectively and efficiently learning the underlying features of surgical activities and dynamic interplay between sequence of actions over time, the system outperforms the state-of-the-art methods on a public surgery dataset in terms of assessment accuracy and model complex-

† These authors contributed equally to this paper.

‡ This research was supported by the Canada Foundation for Innovation (CFI), UAlberta Huawei-ECE Research Initiative (HERI), the Government of Alberta, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), and the Alberta Economic Development, Trade and Tourism Ministry's grant to Centre for Autonomous Systems in Strengthening Future Communities.

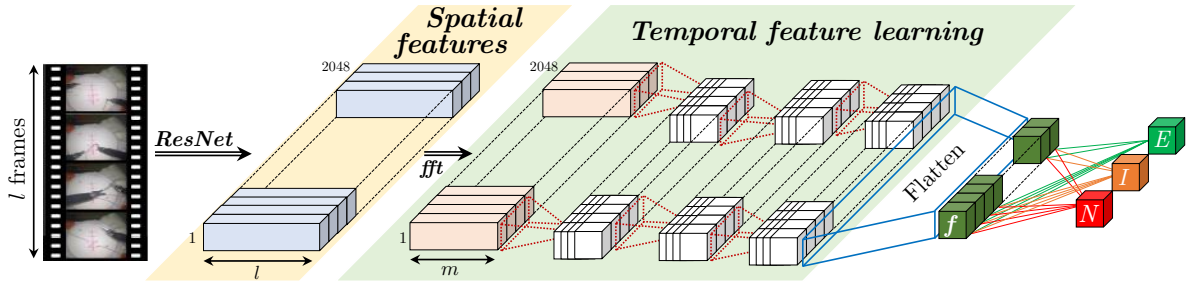


Fig. 1. Network architecture for the proposed video-based skill evaluation method. The spatio-temporal feature representation learning from video recording is decomposed into transfer-learning based intra-frame local feature extraction and an end-to-end inter-frame temporal feature learning. The fast Fourier transform (FFT) operation between the two feature learning phases incorporates prior domain knowledge on robotic surgery in architecture design and helps to reduce the complexity of the model.

ity. It is noteworthy that our method focuses on the analysis of surgery video recording, instead of kinematic or tool motion data, for two reasons. First, video recording is easy to obtain in robot-assisted minimally invasive surgery due to the deployment of laparoscopic camera. Second, though video data is substantially more complicated than sequences of a few motion variables [6], it does contain more comprehensive information on the surgery process.

The structure of the current study is as follows: Section II presents the background of surgical skill evaluation methods and techniques. The proposed skill assessment framework is elaborated in Section III. In Section IV, we evaluate our method on the public JIGSAWS dataset and discuss the advantages and limitations of this study. Finally, conclusions and some recommendations for future study are presented in Section V.

II. Related Work

Early autonomous surgical skill assessment systems were constructed by traditional machine learning techniques. Rosen et al. and others pioneered the idea that the Markov structure of a surgical task was an indicator of user's skill [7]–[9]. Later works extended basic hidden Markov models (HMMs) in a variety of ways by training a segmentation HMM for each skill level [10]–[12]. These endeavors exploit the fact that *"surgeries are composed of several pre-defined surgical gestures and surgeons with a particular level of expertise distinctly perform these gestures"*. Training HMMs requires a large number of gesture annotations that would be very laborious. Therefore, many researchers have attempted manual feature engineering to eliminate the need for gesture annotation [13]. Manual feature engineering is grouped into two main categories: descriptive statistic analysis and predictive modeling-based methods [5]. Descriptive statistic analysis directly calculates clinically-meaningful features as evaluation metrics [13] such as total distance traveled [14], motion jerk [15], movement time [14]–[16], etc. In contrast to descriptive analysis, predictive modeling-based methods extract features and feed them into traditional classifiers or regression models [17], [18]. It should be noted that these feature engineering-based methods usually require specific domain knowledge and are prone to be subjective, time-consuming, and expensive [6].

In recent years, due to the capability of deep learning models, many researchers have used them for autonomous surgical skill assessment. Some deep learning-based approaches take kinematic data as input and use convolutional neural networks (CNNs) to discover skill-related patterns [13]. For instance, Wang and Fey proposed an analytical deep learning system for skill assessment in surgical training. In their study, a deep CNN is used to map multivariate time series data of the motion kinematics to individual skill levels [5]. Moreover, a CNN was proposed by Fawaz et al. to identify surgical skills by extracting latent patterns of kinematics data in the motions of trainees performing robotic surgery tasks [19]. Nguyen et al. presented an autonomous system with a CNN-LSTM neural network model and inertial measurement units (IMU) sensors to systematically classify various levels of expertise in surgical training [20].

While some DL approaches take kinematic data, the majority choose video data that can be captured effortlessly and provides rich contextual details compared to instrument motion. Some video-based deep learning methods formulate surgical skill assessment into classification or regression problems [13]. For instance, Kim et al. proposed a temporal CNN to objectively assess intraoperative technical skills to predict a binary class label (expert/novice) using videos of capsulorhexis [21]. Funke et al. proposed a deep learning-based approach to assess technical surgical skills using video data. For the task of surgical skill classification, they adjusted a pre-trained 3D CNN on stacks of video frames and optical flow fields using the temporal segmentation network [6]. Liu et al. proposed a video-based neural network model that is jointly trained with a supervised regression loss and an unsupervised rank loss to predict users' surgical skill level [22].

Some recent techniques devise a problem of pairwise ranking and learn to characterize relative variation or similarity of skill given a pair of surgeries [13]. For instance, Doughty et al. presented a method that can predict the skill ranking for a set of video data. They suggested a pairwise deep ranking model that utilized both spatial and temporal streams to assess and rate skill in conjunction with a novel

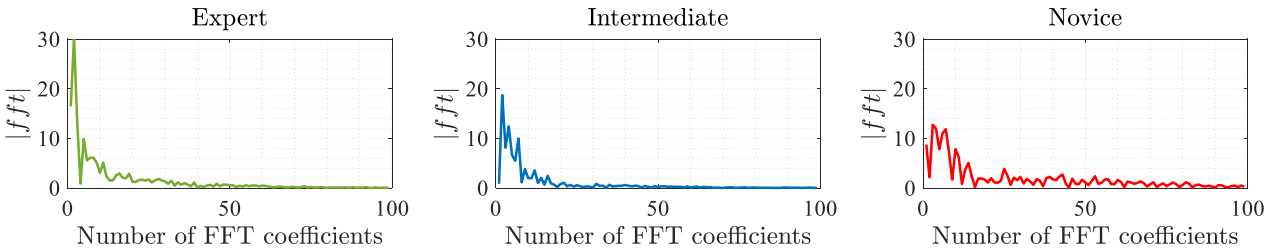


Fig. 2. FFT plots of random trials of three different users with three different level of expertise (i.e., expert, intermediate, and novice) based on their working hours experience with da Vinci Surgical Systems. Expert users typically have dominant lower frequency coefficients and negligible higher frequency activities (i.e., smoother motions). Novice users have negligible low-frequency activities compared to expert and intermediate users. Intermediate users behave between expert and novice.

loss function [23]. [24] presented a new model to determine relative skill level by learnable temporal attention modules from long videos. Li et al. proposed a spatial attention-based approach for skill assessment from videos. They presented a new recurrent neural network (RNN)-based spatial attention model that takes into account accumulated attention state from previous frames as well as high-level information about the progress of an undergoing task [25].

In previous work, authors trained the whole network using an end-to-end learning approach based on a particular dataset (e.g., JIGSAWS). In this paper, we implemented a different approach that uses a pre-trained ResNet50 model to extract spatio-temporal features in the sequence of video frames. Since our video feature extraction procedure is unbiased and naive about our specific dataset, it extracts the salient features contributing to the skill-related behaviors of the user. Using such an unprejudiced approach at the important stage of video feature extraction for a relatively complicated and prone to bias skill evaluation task, makes our model robust and generalized against an upcoming completely new dataset.

III. Methodology

A. Framework overview

Fig. 1 presents the proposed architecture in this study. Particularly, instead of using a large 3-D convolutional neural network, we incorporate prior domain knowledge in the term of Fast Fourier Transfer (FFT) into system design and decompose the spatial-temporal feature learning into two phases: transfer-learning based intra-frame local feature extraction and an end-to-end inter-frame temporal feature learning model. This decomposition helps to construct a lightweight 1-D convolutional model for temporal features learning. The use of FFT explicitly incorporates prior knowledge of robotic surgeries in model design. Basic surgical tasks are inherently repetitive and sequential. Encoding the time-series actions into the frequency domain essentially facilitates retrieving information (e.g., smoothness, jittery motions, abrupt movements, etc.) that differentiates skill levels of surgeons. Additionally, with the prior domain knowledge on surgical activity frequency range, FFT helps to address the issue that surgery recording may have different frame lengths. We will elaborate on the motivation of our design in this section.

Generally, there are two main challenges in skill assessment using video recording data. First, depending on the nature of the surgery, the length of video recording varies, which makes it hard to design a deep learning model. Using methods such as zero-padding or cropping may remove crucial information or adds misleading data [26], [27]. Besides, reshaping in time scale means interfering in the execution speed of the task that is a critical factor in the skill evaluation of a user. Second, since we target to design a task-independent skill assessment method, it is no-trivial to learn the underlying numerical representation corresponding to the nature of motion that meaningfully contributes to the dexterity and level of expertise of the user.

Inspired by the fact that humans can perform tasks with bounded frequency levels (i.e. between 0 and 20 Hz) [28], we observe different statistical patterns among the FFT coefficients associated with activities of surgeons in different skill levels. Intuitively, as illustrated in Fig. 2, expert users typically have dominant lower frequency coefficients and negligible higher frequency activities (i.e., smoother motions). Novice users have negligible low-frequency activities compared to expert and intermediate users. Intermediate users behave between expert and novice. In other words, lower FFT frequency coefficients in Fig.2 are mainly due to the human-robot interaction, while the higher FFT components are attributed to the noise of the robotic device. Based on this fact, we argue that though the length of FFT sequence varies with the length of video recording, usually the first m FFT coefficients are most informative in surgeon skill assessment.

We incorporate FFT in the system to address the two challenges fore-mentioned simultaneously, consequently decomposing the spatial-temporal feature learning into two phases naturally. Specifically, we first apply a pre-trained CNN, such as ResNet50 [29] in Fig. 1, to each video frame for inter-frame spatial features extraction. Several works have shown the effectiveness of transfer learning in image feature extraction [30], [31]. Taking the ResNet50 for example, after feeding a video with l frames, we get a $2048\times l$ feature tensor, where l represents the number of frames in the video recording. To learn the inter-frame temporal features

associating with the surgeon's gestures and maneuvers, we first truncate the $2048\times l$ FFT frequency tensor and keep the first m FFT coefficients for every 2048 features. The new data tensor of $2048\times m$ is then fed to a 1-D CNN model for temporal feature learning. Note, instead of RNN models, we adopt CNN to process the temporal data thanks to its capability of parallel computing. Finally, the obtained deep representations of surgical dexterity levels are passed to a classification head for skill assessment.

B. Model Specification

As shown in Fig. 1, the pre-trained ResNet50 on ImageNet is transferred in the task of skill assessment. In the temporal feature learning model, three 1-D convolutional layers are stacked before the classification head. We construct each convolutional layer following the order of stride 1D convolutional layer - LeackyReLU - batch normalization (BN). All convolutional layers use a kernel size of 3 and the number of filters for the first layer to the third one are 32, 64, and 128 respectively. For the LeakyReLU activation function, the slope of the leak is set to 0.03.

IV. Experimental Evaluation and Discussion

In this study, we evaluate our proposed framework on the public JIGSAWS dataset and compare its performance with state-of-the-art video-based surgical skill assessment methods [23], [6], [32], [21]. In addition, we perform an ablation study on our model, to evaluate its sensitivity against various numbers of FFT coefficients parameter m.

A. Dataset

In this study, we use the public JIGSAWS dataset [33] collected from surgical activities in three different elementary tasks of robot-assisted surgery (i.e., suturing, knot-tying, and needle-passing). Eight right-handed surgeons in three different levels of expertise (i.e., novice, intermediate, and expert) participated in this study. The JIGSAWS consists of 76 dimensional kinematics data, synchronized with two video recordings captured by the left and right endoscopic cameras of the *da Vinci* Surgical System, with the sampling frequency of 30 Hz and the resolution of 640×480 pixels (see Fig. 4). Although the dataset is collected under a simulated environment and cannot exactly recover the real surgical scenes, it demonstrates the axiom of surgical training scenarios very well and thus is widely used by prior arts of surgical skill evaluation. In this work, we only focus on the video data since it is easy to acquire in almost all traditional and modern robot-assisted surgery tasks.

B. Evaluation Protocol and Training details

Since we applied the pre-trained ResNet50 to extract intra-frame spatial features, all video frames were resized and prepossessed accordingly. After applying FFT on the spatial feature tensor, the first top $m{=}50$ FFT coefficients are feed into the temporal feature learning network for the skill classifier network. Note, $m{=}50$ is selected based on the ablation study presented in Fig. 3 (b). Briefly, our model is not sensitive to the number of FFT parameters and is able to generate good performance for any value greater than 20.

All weights in our model are initialized by the random uniform distribution. During training, batch size is set to 4 and the Adadelta optimizer is incorporated with the polynomial decay. To obtain solid results, 10 times of 10-fold cross validation is deployed, considering the relative small number of samples in the dataset. Then all results are averaged and reported.

C. Results and Discussions

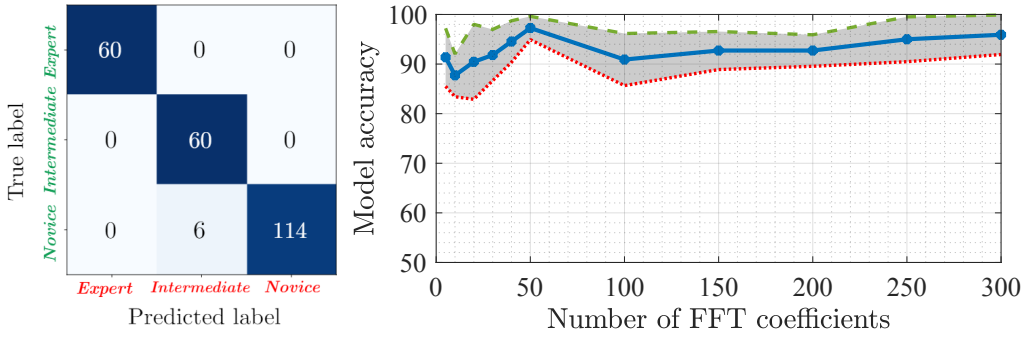
Fig. 3 (a) reports the model's performance in the confusion matrix, which records the statistics in the 10 trials of 10-fold cross-validation. Briefly, the model can successfully classify expert and intermediate surgeons, with the overall accuracy of 97.27 (±2.35).

Table I compares our proposed model to prior arts in terms of classification accuracy and model size. It clearly shows that our model has a smaller size but achieves the top performance. Specifically, to make a fair comparison, 4-fold cross-validation is also performed following the study in Doughty et al. [23]. Compared to the 10-fold cross-validation, the performance degraded a little due to the relatively smaller training set. However, the performance of our model is much better than the one proposed in Doughty et al. [23]. One may notice that we use "-" to indicate the performance of the model proposed by Funke et al. in [6]. Instead of the task-independent skill evaluation approach, Funke et al. [6] proposed to train three different models for each task in the JIGSAWS dataset and reported their performance separately. We argue that though our model and the task-specific models in [6] achieves similar performance, our task-independent model has a better generalization and wider range of application. Besides, we list the performance of prior arts proposed by Lee et al. [32] and Kim et al. [21] here for reference, though both works evaluate their models using other/private datasets.

This study conducted the skill assessment in a retrospective manner, focusing on applications of surgical skill evaluation such as grading the skill levels of surgical trainees. However, this method can be easily adopted in real-time skill assessment platforms. Our recent studies suggest that a 20-second video clip is long enough for the surgical skill assessment. Without any constraints on the length of video data, our method can achieve real-time evaluation by taking the sequence of short video clips of the whole operation as input.

V. Conclusions and Future Work

In this paper, we presented an autonomous deep learning-based framework to objectively assess the level of expertise of surgeons based on the chain of video data frames. To extract the discriminative features for surgical skill assessment, an FFT block was used in the proposed method that decomposes the spatio-temporal representation learning into spatio-feature learning and temporal-feature learning. The two-stage learning in our method reduced the learning



(a) (b)

Fig. 3. Study results for the 10-fold cross-validation. (a) The confusion matrix presents the statistics of 10 times 10-fold cross-validation with the mean accuracy of 97.27%. (b) Ablation study on the effect of the number of FFT coefficients m on the model performance. The model has quite stable performance over a large range of m. The solid blue line represents the average of the model accuracy. Green dashed line and red dotted line represent the higher and lower bound for the model accuracy, respectively.

TABLE I

COMPARISON BETWEEN DIFFERENT SURGICAL SKILL EVALUATION METHODS APPLIED ON THE VIDEO DATA.

Dataset	Authors [year]	Accuracy (%)	model size (# para)
JIGSAWS	Our method (10-fold cv)	97.27 ($\pm$2.35)	220,000
	Our method (4-fold cv)	94.23 ($\pm$2.56)	220,000
	Doughty et al. (4-fold cv) [23] [2018]	76.5 ($\pm$6.5)	16,800,000
	Funke et al. [6] [2019]	-	25,000,000
JIGSAWS	Task-specific evaluation in Funke et al. [6] [2019]		
	Knot-tying task	95.8 ($\pm$1.6)	25,000,000
	Needle-passing task	96.4 ($\pm$0)	25,000,000
	Suturing task	100 ($\pm$0)	25,000,000
Private dataset	Lee et al. [32] [2020]	83	23,000,000
	Kim et al. [21] [2019]	84.8	26,000

Fig. 4. Sample frames of video data in the JIGSAWS dataset for three elementary tasks (From left to right): knot-tying task, needle-passing task, and suturing task.

complexity and led to a lightweight model over 3-D CNNs and 2-D CNN+RNN models in prior arts. We evaluated the proposed method over the JIGSAWS dataset and achieved the skill classification accuracy of 97.27 ($\pm$2.35). Experimentation indicated that our model outperformed state-of-the-arts in terms of generalization and accuracy. Since our model can extract the underlying skill-related features of human activities, it can be used not only in skill assessment but also in other prediction tasks. A good illustration of this is autism recognition. The proposed model can be used to extract underlying features to discriminate between children with autism and normal children.

REFERENCES

[1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[2] JA Martin, Glenn Regehr, Richard Reznick, Helen Macrae, John Murnaghan, Carol Hutchison, and M Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

[3] Alvin C Goh, David W Goldfarb, James C Sander, Brian J Miles, and Brian J Dunkin. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of urology*, 187(1):247–252, 2012.

[4] Carolyn Chen, Lee White, Timothy Kowalewski, Rajesh Aggarwal, Chris Lintott, Bryan Comstock, Katie Kuksenok, Cecilia Aragon, Daniel Holst, and Thomas Lendvay. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of surgical research*, 187(1):65–71, 2014.

[5] Ziheng Wang and Ann Majewicz Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International journal of computer assisted radiology and surgery*, 13(12):1959–1970, 2018.

[6] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.

[7] Jacob Rosen, Blake Hannaford, Christina G Richards, and Mika N Sinanan. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE transactions on Biomedical Engineering*, 48(5):579–591, 2001.

[8] Jacob Rosen, Massimiliano Solazzo, Blake Hannaford, and Mika Sinanan. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. *Computer Aided Surgery*, 7(1):49–61, 2002.

[9] L MacKenzie, JA Ibbotson, CGL Cao, and AJ Lomax. Hierarchical

decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. *Minimally Invasive Therapy & Allied Technologies*, 10(3):121–127, 2001.

[10] Carol E Reiley and Gregory D Hager. Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In *M2CAI workshop. MICCAI, London*, 2009.

[11] Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D Hager, and René Vidal. Sparse hidden markov models for surgical gesture classification and skill evaluation. In *International conference on information processing in computer-assisted interventions*, pages 167–177. Springer, 2012.

[12] Carol E Reiley and Gregory D Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 435–442. Springer, 2009.

[13] Zhiteng Jian, Wenxi Yue, Qiuxia Wu, Wei Li, Zhiyong Wang, and Vincent Lam. Multitask learning for video-based surgical skill assessment.

[14] Timothy N Judkins, Dmitry Oleynikov, and Nick Stergiou. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surgical endoscopy*, 23(3):590, 2009.

[15] Ke Liang, Yuan Xing, Jianmin Li, Shuxin Wang, Aimin Li, and Jinhua Li. Motion control skill assessment based on kinematic analysis of robotic end-effector movements. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1):e1845, 2018.

[16] Ana Luisa Trejos, Rajni V Patel, Richard A Malthaner, and Christopher M Schlachta. Development of force-based metrics for skills assessment in minimally invasive surgery. *Surgical endoscopy*, 28(7):2106–2119, 2014.

[17] Yachna Sharma, Thomas Plötz, Nils Hammerld, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Automated surgical osats prediction from videos. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 461–464. IEEE, 2014.

[18] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Thomas Ploetz, Mark A Clements, and Irfan Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International journal of computer assisted radiology and surgery*, 11(9):1623–1636, 2016.

[19] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(9):1611–1617, 2019.

[20] Xuan Anh Nguyen, Damir Ljuhar, Maurizio Pacilli, Ramesh Mark Nataraja, and Sunita Chauhan. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer methods and programs in biomedicine*, 177:1–8, 2019.

[21] Tae Soo Kim, Molly O'Brien, Sidra Zafar, Gregory D Hager, Shameema Sikder, and S Swaroop Vedula. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *International journal of computer assisted radiology and surgery*, 14(6):1097–1105, 2019.

[22] Daochang Liu, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–484. Springer, 2019.

[23] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2018.

[24] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7862–7871, 2019.

[25] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. Manipulation-skill assessment from videos with spatial attention network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[26] Swarnendu Ghosh, Nibaran Das, and Mita Nasipuri. Reshaping inputs for convolutional neural network: Some common and uncommon methods. *Pattern Recognition*, 93:79–94, 2019.

[27] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[28] Erik K Antonsson and Robert W Mann. The frequency content of gait. *Journal of biomechanics*, 18(1):39–47, 1985.

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

[30] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. *Measurement*, 167:108288, 2020.

[31] Md Nuruddin Qaisar Bhuiyan, Md Shamsujjoha, Shamim H Ripon, Farhin Haque Proma, and Fuad Khan. Transfer learning and supervised classifier based prediction model for breast cancer. In *Big Data Analytics for Intelligent Healthcare Management*, pages 59–86. Elsevier, 2019.

[32] Dongheon Lee, Hyeong Won Yu, Hyungju Kwon, Hyoun-Joong Kong, Kyu Eun Lee, and Hee Chan Kim. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *Journal of clinical medicine*, 9(6):1964, 2020.

[33] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Miccai workshop: M2cai*, volume 3, page 3, 2014.