CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Automated assessment of simulated laparoscopic surgical performance using 3DCNN
Authors	Power, David;Ullah, Ihsan
Publication date	2024
Original Citation	Power, D. and Ullah, I. (2024) 'Automated assessment of simulated laparoscopic surgical performance using 3DCNN', 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1–4. https://doi.org/10.1109/EMBC53108.2024.10782160
Type of publication	Conference item
Link to publisher's version	10.1109/EMBC53108.2024.10782160.
Rights	© 2024, IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Download date	2026-11-16 10:55:02
Item downloaded from	https://hdl.handle.net/10468/17159



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Automated Assessment of Simulated Laparoscopic Surgical Performance using 3DCNN

David Power
ASSERT Centre,
College of Medicine and Health,
University College Cork, Ireland
Email: d.power@ucc.ie

Ihsan Ullah
School of Computer Science
Insight SFI Research Center for Data Analytics,
University of Galway, Ireland

***Abstract*—Artificial intelligence & Computer vision have the potential to improve surgical training, especially for minimally invasive surgery by analyzing intraoperative and simulation videos for training or performance improvement purposes. Among these, techniques based on deep learning have rapidly improved, from recognizing objects, instruments, and gestures, to remembering past surgical steps and phases of surgery. However, data scarcity is a problem, particularly in surgery, where complex datasets and human annotation are expensive and time-consuming, and in most cases rely on direct intervention of clinical expertise. Laproscopic surgical assessment of performance traditionally relies on direct observation or video analysis by human experts, a costly and time-consuming undertaking. A newly collected simulated laparoscopic surgical dataset (LSPD) is presented that will initiate the research in automating this problem and avoiding manual expert assessments. LSPD statistical analyses is given to show similarity and differences between different expertise level (on Stack, Bands, and Tower Skills). Finally, a convolutional neural network is used to predict the experience level of the surgeons, where the model achieved good distinguishing results. The proposed work offers the potential to automate performance assessment and self-learn important features that can discriminate between the performance of novice, trainee, and expert levels.**

I. Introduction

Minimally invasive surgery (MIS) offers numerous benefits, including reduced pain, lower incidence of major complications, less scarring, smaller incisions, less immune system stress, and faster recovery times. However, it is not without risk; 47% of Irish and UK surgeons reported performing a serious error, while 75% know of colleagues who had in the past 12 months [1]. Similarly, in the US 8.9% reported performing a serious error in the past 3 months [2]. Surgical complications are most common during the first 10 procedures performed by trainee MIS surgeons. Mastering MIS skills is challenging and requires skill acquisition beyond open surgery, as it lacks direct tactile feedback and causes a loss of binocularity, amongst other difficulties. For trainees learning these skills is difficult, time-consuming, and prone to error. Acquiring high-quality surgical skills is time-intensive, and requires expert supervision and evaluation. Evaluation of skills is often manual, expensive, time-consuming, and can be subjective [3]. Tools like the objective structured assessment of technical skills (OSATS) have been developed to reduce subjectivity, but it remains a time-consuming manual assessment [4].

Video-based assessment can be a powerful tool for assessing discrete surgical skills and procedures, with a positive correlation between video-based assessment and postoperative patient outcomes [5]. Automation of skills assessment through AI and computer vision (CV) techniques could offer significant benefits in surgery.

A. Background and Related Work

The accessibility of annotated surgical videos is a bottleneck due to the lack of available expert knowledge. On-going efforts to reduce this burden include crowd-sourcing annotations and using supervised, semi-supervised, unsupervised, and reinforcement learning techniques [6]. Various traditional CV techniques and ML techniques have been reported over the years to assess surgical performance, such as optical and magnetic tracking, robotic kinematics, object detection, semantic labeling, histograms of orientated gradients, hidden Markov models, bags of features, support vector machines and random forests [7-10]. The first successful use of deep learning (DL) in MIS was in 2016 with the Convolutional Neural Network (CNN), EndoNet [11]. CNN combined with HMM improves temporal modeling and identification consistency. Combining CNN with LSTM has been more efficient at identifying phases of surgery and tracking surgical instruments. Accuracy rates of 85-90% on surgical phase identification and average precision of 91% for tool detection and tracking using CNN and LSTM, have been reported. However, most studies report these techniques separately. The first successful report using this approach to identify surgical phases and instruments in MIS was by researchers in 2020, with precision of 86.9, recall of 88.0, accuracy of 89.0 for identifying surgical phases, and 89.1 mean average over precision for identifying surgical instruments [12].

B. Limitations and Key Findings

Openly available laparoscopic datasets like Cholec80, SurgAI, Heidelberg Colorectal, LapGyn4, and Dresden Surgical Anatomy are useful for training detection and tracking of instruments, anatomical structures, and surgical phases. However, they are not effective in classifying between performance and skill levels. The JIGSAWS dataset, designed for robotic gesture and skill assessment, was deemed inappropriate for a

simulated laparoscopic surgical setting. A custom dataset was required to discriminate between performance groups. Video datasets are strongly annotated at the video frame level, a bottleneck in AI development in surgery. We did not find a weakly-trained approach to classify between laparoscopic performance groups described in the literature. 3DCNNs have been successfully used to learn volumetric data, recognize surgical workflow, and detect cancer and abnormal growths in endoscopy [13]. However, there seems to be a paucity in the literature on applying the qualities of 3DCNN to learn spatio-temporal features to classify between performance levels with laparoscopic surgical skills.

II. Proposed Methodology

This work acquired and analysed a new dataset (LSPD) for the assessment of simulated laparoscopic surgical performance using a modified 3DCNN. The study aims to classify between novice, trainee, and expert performance of videos of simulated laparoscopic surgical training skills using a desktop surgical simulator. Novices are intern doctors with less than 12 months experience, trainees are those in specialist training programs, and experts are laparoscopic consultants. Three laparoscopic surgical simulation training skills i.e. bands, stacks, and tower were identified using the Laparo™ simulator. Participants watched short videos from the simulator before attempting each skill, they could re-watch videos, ask questions, familiarize themselves with the instruments, or practice a separate skill before attempting the three assessment skills. Researchers often set time limits for attempting skills, which can disadvantage less experienced individuals, introduce measurement bias, and time as a metric is somewhat controversial [14]. Short videos of participants performing basic laparoscopic surgical simulation skills using the Laparo™ system were collected. The three skills varied from relatively easy to very difficult.

A. Laparoscopic Surgical Performance Dataset (LSPD)

The study involved 40 participants, including 8 experts, 12 trainees, and 20 novices. The tasks involved stacking balls, aligning rubber triangles into a tower, and moving elastic bands onto pegs (Fig. 1). Videos were collected from each group, with some performing all three skills multiple times. Some videos were discarded due to obstructed views, to reduce bias in the system. The expert group has 4, 9, and 6 videos for bands, stacks, and tower videos, respectively. The trainee group has 10, 11, and 8 videos from bands, stacks, and tower, respectively. Finally, the novice group has 8, 21, and 19 bands, stacks, and tower videos. This results in a total of 106 videos. OpenCV was used to trim the videos to reduce the noise at the beginning and end while preserving participant actions. The last 60 seconds of the training videos were analyzed due to computational constraints.

Pre-processing: To increase dataset size and diversity, data augmentation was used to generate more training examples. A Gaussian blur sigma of 0.2, brightness alpha of 1.2, contrast alpha of 1.2, salt and pepper noise were added at 0.02 each, and horizontal flipping of the dataset. The final dataset contained 2244 videos in total.

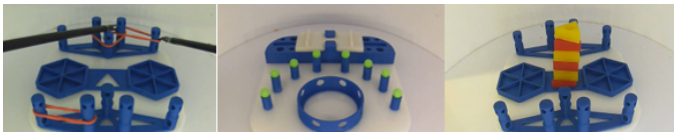


Fig. 1. Skills: Bands (left), Stack (center), Tower (right)

B. Proposed 3DCNN Architecture

A 3DCNN perform convolutions in three dimensions and learns spatiotemporal features, hence, are well-suited to video analysis. The 3DCNN model was built using TensorFlow and Keras with the first, second, third, and fourth layer having 64, 128, 256, and 512, kernels all with 3x3x3 size. The image size was reduced from its original 1280x720 to an input of 128x128 for computational efficiency while retaining essential spatial information. Zero-padding ensured all videos were the same length. The ReLU activation function is used as it is computationally efficient and avoids vanishing gradients. A batch normalization layer was added after each 3D convolutional layer to improve network training and convergence. After each batch normalization layer, a max-pooling layer was used to reduce computational complexity and improve translation invariance by downsampling feature maps in both spatial and temporal dimensions with a window size of 2x2x2. Dropout layers were introduced after each max-pooling layer to prevent overfitting and improve network generalization. The first three dropout layers have small dropout rates of 0.1-0.2, while the final two have a rate of 0.5. The 3DCNN contains three fully connected layers, the first has 1024 neurons, while the second has 512 neurons. Two dropout layers between these layers reduce overfitting and improve generalization. The final fully connected layer is the output layer, which consists of neurons representing each class e.g. trainee, expert, and novice, and is responsible for making classifications.

III. Experimental Outcomes

The dataset is analysed in two ways for classification of the expertise levels i.e. statistical analysis and classification through modified 3DCNN.

A. Statistical Analysis of Videos

All skills were statistically analysed, with the purpose of gauging skill difficulty. First, the median skill completion time of each group was calculated. A basic scoring system was created for each skill, with the stack skill scoring 8/8, with the final score determined by the number of balls remaining on the stacks at the end of the video. The total performance score for the band's skill was 6/6, with 1 for each band moved correctly to the correct pegs and 1 for exact symmetry with the starting position pegs. The total score for the tower skill was 12/12, with 1 for each of the six triangles placed in the correct location and 1 per triangle for the exactness of direction and angle of points.

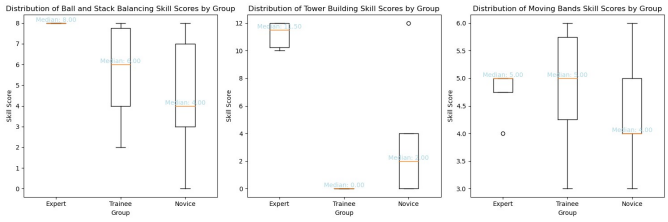


Fig. 2. Distribution of Performance for each skill

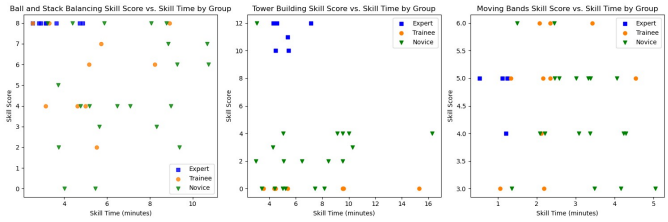


Fig. 3. Performance v Time for each skill

Two healthcare simulation experts marked videos of 30 skills, with inter-rater reliability measured using Cohen's Kappa (k). The stack skill had very high agreement i.e. k=1.0, while the tower skill having k=0.76, and bands having k=0.72 had moderately high agreement. As the aim was to measure model performance, rather than human performance, this was considered robust enough for this study.

The median expert time for the stack skill was 3.47 minutes, compared to 5.11 minutes for trainees and 5.45 minutes for novices. The median expert performance score for the stack skill was 8/8. This compared to a median performance score of 6 for trainees and 4 for novices. The median expert time to complete the tower skill was 4.98 minutes, compared to 7.46 for trainees and 6.47 minutes for novices. The median expert performance score for the tower skill was 11.5/12. This compared to 0 for trainees and 2 for novices. Although it would appear that novices were speedier and performed better, this was not the case. By the time the video ended, some of the trainees had almost erected the tower correctly when they unintentionally toppled it. This skill was the most challenging, and many novices gave up after an attempt to stack only a few triangles. The rudimentary scoring system simply considers the single tower that is upright at the end of the video, not actual participant performance throughout the entire video. The expert median time for the band's skill was 1.17 minutes, compared to 2.17 minutes for trainees and 3.2 minutes for novices. The median expert performance score for the band's skill was 5/6. Compared to 5 for trainees and 4 for novices.

Medians between groups were compared for statistical significance using the Kruskal-Wallis H-statistic. The critical value was calculated at 5.99, and the H-statistic was compared to the critical value from the chi-square distribution. P-values were calculated with a pre-defined significance level of a=0.05. The stack skill has statistically significant medians for both

Skill Metric	Expert	Trainee	Novice	Kruskal-Wallis H Statistic*	p-Valve^
Stack Time	3.47 (1.14)	5.11 (2.0)	6.45 (2.32)	14.08	<0.05 (0.0087)
Stack Score	8.0 (0.0)	6.0 (2.11)	4.0 (2.6)	11.70	<0.05 (0.0028)
Tower Time	4.98 (1.06)	7.46 (4.0)	6.47 (3.34)	1.28	0.52
Tower Score	11.5 (0.98)	0.0 (0.0)	2.0 (2.8)	19.83	<0.05 (0.000049)
Bands Time	1.17 (0.34)	2.17 (0.99)	3.2 (1.01)	12.53	<0.05 (0.00189)
Bands Score	5.0 (0.5)	5.0 (1.14)	4.0 (0.97)	1.858	0.39
*critical value = 5.99, degrees of freedom = 2, ^alpha = 0.05					

Fig. 4. Median, SD, H-statistic, and p-values for skill metrics

the Time and Performance Score metrics with an H statistic of 14.08 (critical value >5.99) and p-value 0.0087 (p-valve <0.05), and an H statistic of 11.70 and p-value 0.0028, respectively. The tower skill's median for the Time metric was not statistically significant between the groups, as indicated by the low H-statistic of 1.28 and high p-value of 0.52. However, there was statistical significance between the means for the Performance Score in the tower skill, with an H-statistic of 19.83 and a p-value of 0.000049. The band's skill demonstrated statistical significance between the means for the Time metric with an H-statistic of 12.53 and a p-value of 0.00189. However, there was no statistical significance between the means of the Performance Score for the band's skill, with an H-statistic of 1.858 and a p-value of 0.39 as shown in Fig. 4. The tower skill was the most difficult and a real discriminator between expert and non-expert performance.

B. 3DCNN Classification Results

The proposed 3DCNN model was tested in two ways i.e. as a multiclass classifier and as a binary classifier. In the first, it is trained to classify among performance classes novice, trainee, and expert over the three skills. However, the model frequently misclassified instances and had poor test accuracy, especially for the tower and bands skills, as seen in the confusion matrices Fig. 5. The testing accuracy of stack (i.e. 79%) is higher than the other two tower (i.e. 49%) and bands (i.e. 54%) skills. This aligns with the results shown in the statistical analysis i.e. tower is one of the difficult task.

In the second test, the model was trained as a binary classifier, with two classes of novice and expert for the same three skills, with much-improved performance as shown in Fig. 6. All three skills were trained with a cross-validation procedure (k=5). The data was divided in training and testing i.e. 80% for training and remaining 20% for testing. From the training data, 20% was used for validation. The data was shuffled for randomness.

The test accuracy of the stack skill was 91%. Expert class precision was 0.94 and recall 0.88. Novice class precision was 0.91. Both classes had an F1-score of 0.91. The expert class has an AUC of 0.92 and the novice class have 0.92. The tower skill's test accuracy was 0.97. Expert class precision was 0.95, and recall 1.0. Novice class precision was 1.0, and recall 0.95. Both classes had an F1-score of 0.97. The AUC

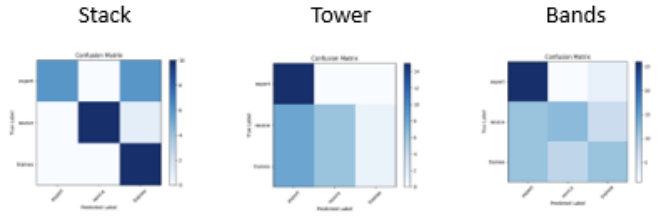


Fig. 5. Multi-class confusion matrices. Test accuracy of stack, tower, and bands is 79%, 49%, and 54%, respectively.

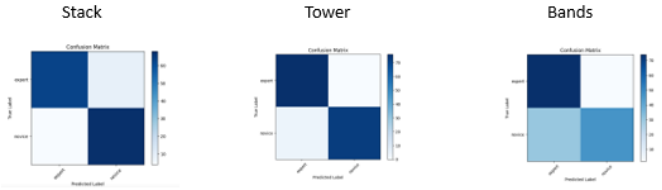


Fig. 6. Binary class confusion matrices. Test accuracy of stack, tower, and bands is 91%, 97%, and 79%, respectively.

for expert and novice classes is 0.99. The model has excellent predictive ability and can discriminate easily between these class instances for both the stack and tower skills. The band skill's test accuracy was 79%. Expert class precision was 0.71, and recall of 0.97. The expert class achieved an F1-score of 0.82. Novice class precision was 0.96, and recall score was 0.61. The novice class has an F1-score of 0.74. The model's AUC for the expert class is 0.86 and 0.79 for the Novice class. The model achieved reasonable performance for this skill however, it struggled to correctly classify all experts as experts and missed significant numbers of novices within the dataset. This is not unsurprising as it was the easiest skill according to the statistical analysis.

IV. Conclusion and Future Work

In conclusion, this study successfully demonstrated automated assessment of laparoscopic surgical performance in a simulated setting using a weakly-trained 3DCNN model on a custom dataset and lays the foundation for a new research area. The model demonstrated excellent performance and predictive ability for both the tower and stack skills, and reasonable performance and predictive ability for the band's skill on unseen test data, as a binary classifier. This approach is viable for classification between expert and novice classes performing basic simulated laparoscopic surgical skills in a desktop laparoscopic surgical simulator. However, this approach could be expanded to other skills beyond simulated laparoscopic skills and has scope for assessment of performance for other procedural and clinical skills. There is significantly less domain expert time needed using this approach as there is no need for frame-level annotation.

However, the model found predicting between three performance skill levels much more challenging, when faced with a multi-class problem. Once the trainee class was removed the model performed excellently. The trainee group needs to be better defined at the onset. The distribution of demonstrated skill level was too wide in this group with some close to novice and some close to expert. In the future, more detailed definitions of each group, particularly an intermediate group such as the trainees are required. The videos were relatively short due to high computational demands, future work could look at augmented approaches of attention mechanisms, temporal pooling, or temporal segmentation to increase the video length sequences and widen the application of use of this approach within the healthcare simulation field.

Acknowledgment

The authors would like to thank the volunteers who gave their time for data collection. In addition, this publication has emanated from research supported by a grant from Science Foundation Ireland under Grant number SFI/12/RC/2289_P2.

References

[1] A. D. White, M. Skelton, F. Mushtaq, T. W. Pike, M. Mon-Williams, J.P. Lodge, R.M. Wilkie. "Inconsistent reporting of minimally invasive surgery errors", Annals of the Royal College of Surgeons England, vol. 97, no. 8, pp. 608-12, 2015.

[2] N. A. Shah, J. Jue & T. K. Mackey. "Surgical Data Recording Technology: A Solution to Address Medical Errors?", Annals of Surgery, vol. 271, no. 3, pp. 431-433, 2020.

[3] J. D. Dimick, A. M. Ryan. "Taking a Broader Perspective on the Benefits of Minimally Invasive Surgery", JAMA Surgery, vol. 48, no. 7, pp.648, 2013

[4] A. Zia, I. Essa. "Automated surgical skill assessment in RMIS training", International Journal of Computer Assisted Radiology and Surgery, vol. 13, pp. 731-739, 2018.

[5] C. P. Scally, O. A. Varban, A. M. Carlin, J.D. Birkmeyer, J. B. Dimick. "Video Ratings of Surgical Skill and Late Outcomes of Bariatric Surgery", JAMA Surgery, vol.151, no.6:e160428, 2016.

[6] T. M. Ward, D. M. Fer, Y. Ban, G. Rosman, O. R. Meireles, D. A. Hashimoto. "Challenges in surgical video annotation", Computer Assisted Surgery, vol. 26, no. 1, pp.58-68, 2021.

[7] R. Nagyn, E. Elek, T. Haidegger et al., "Surgical Tool Segmentation on the JIGSAWS Dataset for Autonomous Image-based Skill Assessment", ICCC 2022, IEEE 10th Jubilee International Conference on Computational Cybernetics and Cyber-Medical Systems, July 6-9, 2022, Reykjavík, Iceland.

[8] Y. Bamba, S. Ogawa, M. Itabashi, H. Shindo, S. Kameoka, T. Okamoto, M. Yamamoto. "Object and anatomical feature recognition in surgical video images based on a convolutional neural network", International Journal of Computer Assisted Radiology and Surgery, vol.16, pp. 2045-2054, 2021.

[9] M. Ríos, M. Molina-Rodriguez, D. Londoño, et al., "Cholec80-CVS: An open dataset with an evaluation of Strasberg's critical view of safety for AI", Nature, vol.10, no.194, pp. 1-7, 2023.

[10] L. Maier-Hein, M. Wagner, T. Ross, et al., (2021). "Heidelberg colorectal data set for surgical data science in the sensor operating room". Nature, vol.8, no.1010, pp. 1-11, 2021.

[11] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.W. Fu, P. A. Heng. "SV-RCNet: Workflow Recognition from Surgical Videos Using Recurrent Convolutional Network", IEEE Transactions on Medical Imaging, 2018 May, vol. 37, no. 5, pp. 1114-1126.

[12] Y. Jin, H.Li, Q. Dou, H. Chen, J. Qin, C.W. Fu, P. A, Heng. "Multi-task recurrent convolutional network with correlation loss for surgical video analysis", Medical Image Analysis vol. 59, e101572, 2020.

[13] N. Ghatwary, M. Zolgharni, F. Janan, X. Ye. "Learning Spatiotemporal Features for Esophageal Abnormality Detection From Endoscopic Videos," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 1, pp. 131-142, Jan. 2021.

[14] T. Jackson. "Does speed matter? The impact of operative time on outcome in laparoscopic surgery", Surgical Endoscopy, vol.25(7), pp.2288-2295, 2011.