Abstract

Purpose: This research aims to facilitate the use of state-of-the-art computer vision algorithms for the automated training of surgeons and the analysis of surgical footage. By estimating 2D hand poses, we model the movement of the practitioner’s hands, and their interaction with surgical instruments, to study their potential benefit for surgical training.

Methods: We leverage pre-trained models on a publicly-available hands dataset to create our own in-house dataset of 100 open surgery simulation videos with 2D hand poses. We also assess the ability of pose estimations to segment surgical videos into gestures and tool-usage segments and compare them to kinematic sensors and I3D features. Furthermore, we introduce 6 novel surgical skill proxies stemming from domain experts’ training advice, all of which our framework can automatically detect given raw video footage.

Results: State-of-the-art gesture segmentation accuracy of 88.49% on the Open Surgery Simulation dataset is achieved with the fusion of 2D poses and I3D features from multiple angles. The introduced surgical skill proxies presented significant differences for novices compared to experts and produced actionable feedback for improvement.

Conclusion: This research demonstrates the benefit of pose estimations for open surgery by analyzing their effectiveness in gesture segmentation and skill assessment. Gesture segmentation using pose estimations achieved comparable results to physical sensors while being
remote and markerless. Surgical skill proxies that rely on pose estimation proved they can be used to work towards automated training feedback. We hope our findings encourage additional collaboration on novel skill proxies to make surgical training more efficient.

**Keywords:** Machine Learning, Computer Vision, Gesture Recognition, Surgical Skill Assessment, Pose Estimation, Surgical Training

1 Introduction

Until this day, the traditional surgical training methodology of “see one, do one, teach one” by Dr. William Halsted [1] remains the most practiced method among medical practitioners. A major drawback of this methodology is the limited availability of expert surgeons. To make the training process more efficient, several works have been done that aim to recognize gestures and estimate surgical skill levels using different input modalities. Those include kinematic sensory data [2, 3], optical flow [4], and RGB videos [5–10].

Sensor-based models offer accurate spatial coordinates of the hands and surgical instruments; however, they are difficult to implement due to the expensive hardware and setup required. Nevertheless, the research of Fawaz et al. [2] and Goldbraikh et al. [3] have shown that this type of input produces insightful motion metrics and statistics that help classify surgical skill.

A more suitable solution is computer vision models that rely on RGB videos as input. This is due to RGB being easy to set up and minimally interfering with the surgical workflow. Some of the popular datasets in this field are the JIGSAWS dataset [11] and the more recent RARP-45 dataset [12]. Existing studies focus on hand detection [13–15], surgical instrument detection [6, 16], tool usage [6, 17], surgical gesture recognition [3–5, 10], and surgical skill classification [2, 7–10].

Existing work on skill assessment focuses on classifying the performance using categorical labels or a numeric score and statistically comparing the performance of novices to experts. Existing frameworks offer feedback that includes (1) highlighting the segments that contributed to the skill classification [2] and (2) providing **global** motion metrics and statistics for novice compared to experts [3, 6, 7]. To the best of our knowledge, a fully automated framework that performs explainable skill assessment with task-specific actionable feedback hasn’t been released yet.

This research aims to close the gap in the implementation of computer vision models in the surgical training environment by exploring surgical skill proxies that automate the expert’s advice. The proposed proxies produce **task-specific** feedback focused on surgical tool utilization for scissors, needle drivers, and forceps, as well as suture holding, and cutting positions.

This paper’s contributions are as follows:
1. A novel 2D hand pose dataset of open surgery simulation with annotations of instruments, hands, tool usage, gestures, and skill levels.

2. State-of-the-art results on the multi-task gesture segmentation problem using the predicted skeletons.

3. A fully automated modular surgical skill assessment framework with actionable and explainable feedback.

2 The Dataset

The dataset used in this research is the Open Surgery Simulation dataset [3]. It contains 100 videos of 25 clinicians with varying levels of expertise. The dataset contains temporal segmentation annotations for tool utilization and gesture recognition. The tools are needle driver, scissors, and forceps. The gestures are ”No Gesture”, ”One Shot Needle Passing”, ”Pull The Suture”, ”Instrumental Tie”, ”Lay The Knot”, and ”Cut The Suture”.

In order to evaluate the added value of pose estimations to surgical skill assessment, 925 frames (0.01\% of the dataset) sampled from 15 videos were annotated with 2D hand poses that include bounding boxes and 21 key-points. This results in a total of 1987 annotated hand instances.

3 Methods

3.1 Object Detection

For the task of hands and tools detection, a real-time YOLO-X [18] object detection model was used. The model was trained using the ”mmdetection” framework [19] for 400 epochs. The dataset was split into 5 folds using cross-user validation (Leave One User Out [11] modified for groups of users). The model’s weights were initialized using a pre-trained model on the One-Hand [20] dataset. The prediction head was modified to detect 2 hands (left and right) and 4 surgical tools (needle driver, scissors, forceps, forceps-not-used). The model was evaluated using the mean average precision (mAP) of intersection over union (IoU), and the final model was used to extract per-frame bounding boxes using a confidence threshold of 0.5.

3.2 2D Pose Estimation

The pose estimation model was trained using the ”mmpose” framework [21]. HRNet [22] was compared to Simple-Baseline ResNet [23], and the latter was used during the succeeding stages due to its speed of 43 items/second compared to HRNet’s 9 items/second on a Tesla V100 GPU. These models were evaluated using the probability of correct keypoint (PCK), area under the curve (AUC), and endpoint error (EPE) metrics. The models were pre-trained on the OneHand [20] dataset and fine-tuned on our dataset for 85 additional epochs. The final model was used to extract per-frame pose estimations using a key point confidence threshold of 0.3. The missing keypoints were imputed using
the last observation carried forward (LOCF) method. Finally, the Savitzky-Golay [24] signal smoothing algorithm was used to minimize the jittering of key-points across time.

3.3 Multi-task Temporal Activity Segmentation

The final skeletons and tool detections are combined to form the **input features** of our spatio-temporal deep learning architecture as seen in Fig. 1. The architecture uses MSTCN++ [25] to predict the performed gesture and tools utilized for each frame. We compare two architectures for multi-task prediction, the first using a modified prediction head, and the other using separate networks for each task, referred to as Late Fusion. Our input features were compared to I3D features [26], and to the combination of our input and I3D features. I3D features were extracted from a model that was pre-trained on the kinetics-400 [27] dataset and fine-tuned on our dataset. Our results were also compared to the current state-of-the-art benchmark on this dataset which uses kinematic sensor data [3]. The hyper-parameter space of MSTCN++ was explored using the random search strategy with 100 trials. The range for learning rate was [0.0001, 0.0005, 0.005, 0.001], for feature maps it was [32, 64, 128, 256], for prediction generation layers it was [3, 7, 11, 15], for refinement layers it was [6, 10, 14], and for refinement stages it was [1, 3, 6, 12].

**Fig. 1**: Data Pipeline
3.4 Surgical Skill Assessment

For the purpose of surgical skill assessment, we propose an approach that relies on the building block of a Surgical Skill Proxy [8]. In our case, a surgical skill proxy is defined as a metric that can be calculated using the tools and skeletons in the video, and that can be explained to the performer in simple words.

In order to come up with meaningful proxies and provide evidence of validity, a domain expert (Author AD) was consulted during the proxy engineering process to provide face validity. The initial proxies were created based on known techniques, and later on, new ones were identified based on the data in an iterative process. Following are the proxies that showed promising feedback and how they model the practitioner’s performance.

3.4.1 Gesture Duration

**Proxy measurement:** This proxy measures the time spent on a particular portion of the operation.

**Clinical relevance:** Previous studies [3] have shown differences in duration between novice and expert when performing different gestures. Furthermore, the background gesture which indicates the time spent on setup and non-specific movement between gestures tends to be longer for residents as observed in figure 4.

3.4.2 Hand Orientation

**Proxy measurement:** This proxy measures the level of pronation of the hand as seen in figure 2. It’s calculated as \( x_{\text{index}} - x_{\text{pinky}} \) where \( x_i \) is the x coordinate of the key point corresponding to the Metacarpophalangeal (MCP) joints of the i’th finger. The proxy produces high positive values when the hand is in a completely pronated position (palm rotated down), values around zero when the hand is directed to the sides (palm to the side and thumb up or down), and low negative values when the hand is fully supinated (palm rotated up).

**Clinical relevance:** Research has shown that pronation and supination of the hand are considered key skills when learning to perform surgical sutures. For example, when cutting the suture, full supination of the hand leads to the scissors forming a 90-degree angle with the suture, increasing the chance of cutting the knot. Therefore, slight supination of the hand is encouraged to form a 45-degree angle instead [28]. When holding forceps, full pronation of the hand indicates an incorrect holding position, whereas slight supination leads to a pencil-like holding position, as is taught by surgical experts [28]. Another example is needle passing, where the literature instructs starting with a pronated position, and ending the gesture with a slightly supinated position [28]. Beginning the gesture in a supinated position in this case limits
the freedom of hand movement, resulting in awkward hand positions with less granularity of control.

3.4.3 Distance Between Thumb and Index Fingers

Proxy measurement: This proxy measures the distance in pixels between the tip of the thumb and index fingers as seen in figure 2.

Clinical relevance: The basic method of holding a needle driver is to straighten the index finger while slightly inserting the thumb through one of the handles. This way, the index finger guides and stabilizes the needle driver while the thumb is used to open and close it [29]. A more advanced technique is to palm the needle driver, by resting one side of the handle on the thenar eminence rather than inserting the thumb into the handle and using the thenar eminence to open and close it. This has the advantage of allowing a wider range of rotational motion of the instrument within the hand [29]. Another example is the suture holding position. When holding the suture, we observed that experienced surgeons tend to hold the suture at the edge of their fingers, allowing them a stronger grip on the suture, whereas novice surgeons end up holding the suture with different parts of their thumb, perhaps due to the larger surface area, allowing easier yet less granular grip of the suture.

3.4.4 Fingers to Tissue Distance

Proxy measurement: This proxy measures the distance in pixels between the tissue and the fingers holding the suture as seen in figure 2.

Clinical relevance: When cutting the suture, the proximity of the hand to the tissue impacts the surgeon’s field of view. A small distance could lead to obscuring the suture, making the cut prone to errors. Oftentimes, assistants cut sutures for the leading surgeon, therefore, holding the suture at distance is required for better surgical performance.

3.4.5 Hand Velocity

Proxy measurement: This proxy measures the velocity of the hand when pulling the suture using a needle driver. The velocity of the hand is measured from multiple key points to account for subtle rotations of the hand. Such movements might be undetectable solely using object detection methods.

Clinical relevance: When pulling the suture, the surgeon needs to leave just the right amount of tail to allow for efficient performance during instrument ties [30]. Too short of a tail does not allow enough suture to complete the knot and too long of a tail makes it cumbersome to pull through the loop to form the knot. Novice surgeons who aren’t familiar with the gesture, need to move their hand slowly while keeping an eye on the tail to leave the right
amount, whereas expert surgeons who are familiar with the gesture move, their hand faster consistently leaving an accurate tail length.

![Surgical Proxy Visualization](image)

**Fig. 2:** Surgical Proxy Visualization

### 4 Experiments and Results

Table 1 shows the pose estimation results of HRNet compared to Simple Baseline. Simple Baseline shows comparable accuracy but offers a significant advantage in speed. The gesture segmentation results of our multi-task network are presented in table 2. The highest accuracy of 88.49% was achieved from the fusion of 2D poses and I3D visual features from multiple angles as seen in figure 3. Figure 5 shows a box plot of mean proxy values of novices compared to experts for the presented proxies during the relevant gestures.

| Model Name                              | PCK  | AUC  | EPE  | Item/s |
|-----------------------------------------|------|------|------|--------|
| Resnet Simple Baseline Pre-trained      | 0.729| 0.567| 15.107| 43.47  |
| HRNet Pre-trained                       | 0.771| 0.603| 13.279| 9.52   |
| Resnet Simple Baseline Fine-tuned       | 0.949| 0.774| 7.178 | 43.47  |
| HRNet Fine-tuned                        | 0.951| 0.776| 7.091 | 9.52   |
5 Conclusion and Discussion

When it comes to gesture segmentation, remote pose estimations showed comparable accuracy to sensors on the frontal view (81.81% vs 82.40%), and better accuracy on the close-up and mixed views (84.16% and 85.39%). Despite being less accurate than I3D features, they offer the advantage of context isolation,
and support for concrete feedback through skill proxies. The late fusion architecture showed up to 0.47% boost in accuracy and up to 1.84% in edit score in comparison to the multi-task head. Furthermore, concatenating key points from multiple views leads to a 1.23% increase in accuracy which leads us to believe that 3D pose estimations could outperform current results. Finally, we see that combining pose estimations and I3D features results in up to 0.46% added accuracy.

An important limitation to note is our use of 2D pose estimations instead of 3D. This limits some of our proxies such as hand orientation 3.4.2 to a single camera angle. Given that we used the videos of the frontal view in our dataset, applying this proxy was successful. When applying it using different camera angles, a 3D pose would provide a more accurate result.

This research aims to bridge the gap in the application of novel computer vision algorithms to the domain of surgical training and monitoring. We reveal how 2D pose estimation can be applied to new open surgery datasets, and how it can be utilized for gesture segmentation and skill assessment. By creating the proxies depicted in section 3.4, and assessing known groups’ validity evidence with our dataset, we demonstrate how the expert’s advice can be automated through the surgical skill proxy methodology. This paves the way to work towards a fully automated surgical training framework requiring only a performance video.

**Declarations**

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

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.
Informed consent  Informed consent was obtained from all individual participants included in the study.

Code availability  The code will be released after publication at https://github.com/edybk/Pose-Estimation-For-Surgical-Training

References

[1] Kotsis, S.V., Chung, K.C.: Application of the "see one, do one, teach one" concept in surgical training. Plastic and Reconstructive Surgery 131(5), 1194–1201 (2013). https://doi.org/10.1097/PRS.0b013e318287a0b3

[2] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A.: Evaluating surgical skills from kinematic data using convolutional neural networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 214–221. Springer, Cham (2018)

[3] Goldbraikh, A., Volk, T., Pugh, C.M., Laufer, S.: Using open surgery simulation kinematic data for tool and gesture recognition. IJCARS 17(6), 965–979 (2022). https://doi.org/10.1007/s11548-022-02615-1

[4] Sarikaya, D., Jannin, P.: Surgical Gesture Recognition with Optical Flow only. arXiv (2019). https://doi.org/10.48550/ARXIV.1904.01143. https://arxiv.org/abs/1904.01143

[5] Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3d convolutional neural networks. IJCARS 14(7), 1217–1225 (2019). https://doi.org/10.1007/s11548-019-01995-1

[6] Goldbraikh, A., D’Angelo, A.-L., Pugh, C.M., Laufer, S.: Video-based fully automatic assessment of open surgery suturing skills. IJCARS 17(3), 437–448 (2022). https://doi.org/10.1007/s11548-022-02559-6

[7] Lavanchy, J.L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., Beldi, G.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Scientific Reports 11(1), 5197 (2021). https://doi.org/10.1038/s41598-021-84295-6. Accessed 2022-03-10

[8] Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9517–9526. IEEE Computer Society, Los Alamitos, CA, USA (2021). https://doi.org/10.1109/CVPR46437.2021.00940. https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00940
[9] Wang, T., Jin, M., Li, M.: Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation. IJCARS 16(9), 1595–1605 (2021). https://doi.org/10.1007/s11548-021-02448-4. Accessed 2022-03-10

[10] Zhang, J., Nie, Y., Lyu, Y., Yang, X., Chang, J., Zhang, J.J.: SD-Net: joint surgical gesture recognition and skill assessment. IJCARS 16(10), 1675–1682 (2021). https://doi.org/10.1007/s11548-021-02495-x. Accessed 2021-12-01

[11] Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., B’ejar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2cai, vol. 3, p. 3 (2014)

[12] van Amsterdam, B., Funke, I., Edwards, E., Speidel, S., Collins, J., Sridhar, A., Kelly, J., Clarkson, M.J., Stoyanov, D.: Gesture Recognition in Robotic Surgery With Multimodal Attention. IEEE Transactions on Medical Imaging 41(7), 1677–1687 (2022). https://doi.org/10.1109/TMI.2022.3147640

[13] Zhang, M., Cheng, X., Copeland, D., Desai, A., Guan, M.Y., Brat, G.A., Yeung, S.: Using Computer Vision to Automate Hand Detection and Tracking of Surgeon Movements in Videos of Open Surgery. AMIA ... Annual Symposium proceedings. AMIA Symposium 2020, 1373–1382 (2020)

[14] Louis, N., Zhou, L., Yule, S.J., Dias, R.D., Manojlovich, M., Pagani, F.D., Likosky, D.S., Corso, J.J.: Temporally Guided Articulated Hand Pose Tracking in Surgical Videos. arXiv (2021). https://doi.org/10.48550/ARXIV.2101.04281. https://arxiv.org/abs/2101.04281

[15] Goodman, E.D., Patel, K.K., Zhang, Y., Locke, W., Kennedy, C.J., Mehrotra, R., Ren, S., Guan, M.Y., Downing, M., Chen, H.W., Clark, J.Z., Brat, G.A., Yeung, S.: A real-time spatiotemporal AI model analyzes skill in open surgical videos. arXiv:2112.07219 [cs] (2021). arXiv: 2112.07219

[16] Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L.: Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. arXiv:1802.08774 [cs] (2018). arXiv: 1802.08774. Accessed 2021-12-01

[17] Basiev, K., Goldbraikh, A., Pugh, C.M., Laufer, S.: Open surgery tool classification and hand utilization using a multi-camera system. IJCARS 17(8), 1497–1505 (2022). https://doi.org/10.1007/s11548-022-02691-3
Pose Estimation For Surgical Training

[18] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO Series in 2021. arXiv. arXiv:2107.08430 [cs] (2021). http://arxiv.org/abs/2107.08430

[19] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv. arXiv:1906.07155 [cs, eess] (2019). http://arxiv.org/abs/1906.07155

[20] Wang, Y., Peng, C., Liu, Y.: Mask-Pose Cascaded CNN for 2D Hand Pose Estimation From Single Color Image. IEEE Transactions on Circuits and Systems for Video Technology 29(11), 3258–3268 (2019). https://doi.org/10.1109/TCSVT.2018.2879980

[21] Contributors, M.: OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose (2020)

[22] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[23] Haiping, Bin, W.Y.X., Wu: Simple Baselines for Human Pose Estimation and Tracking. In: Martial, Cristian, S., Vittorio, W.Y.F., Hebert (eds.) Computer Vision – ECCV 2018, pp. 472–487. Springer, ?? (2018)

[24] Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), 1627–1639 (1964). https://doi.org/10.1021/ac60214a047. eprint: https://doi.org/10.1021/ac60214a047

[25] Li, S., Farha, Y., Liu, Y., Cheng, M.-M., Gall, J.: MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence PP (2020). https://doi.org/10.1109/TPAMI.2020.3021756

[26] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017). https://doi.org/10.1109/CVPR.2017.502

[27] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The Kinetics Human Action Video Dataset. arXiv. arXiv:1705.06950 [cs] (2017). http://arxiv.org/abs/1705.06950
[28] Trott, A.: Wounds and Lacerations: Emergency Care and Closure vol. 4. Elsevier Saunders, ??? (2012). https://doi.org/9780323074186/9780323091329

[29] Kantor, J.: Atlas of Suturing Techniques: Approaches to Surgical Wound, Laceration, and Cosmetic Repair vol. 1. McGraw Hill, ??? (2016). https://doi.org/0071836578/9780071836579

[30] Meloni, S., Mastenbjork, M.: Suture Like a Surgeon vol. 1. Medical Creations, ??? (2019). https://doi.org/1698150857/978-1698150857