



Depth over RGB: automatic evaluation of open surgery skills using depth camera

Ido Zuckerman¹ · Nicole Werner² · Jonathan Kouchly¹ · Emma Huston³ · Shannon DiMarco³ · Paul DiMusto² · Shlomi Laufer¹

Received: 4 March 2024 / Accepted: 22 April 2024
© The Author(s) 2024

Abstract

Purpose In this paper, we present a novel approach to the automatic evaluation of open surgery skills using depth cameras. This work is intended to show that depth cameras achieve similar results to RGB cameras, which is the common method in the automatic evaluation of open surgery skills. Moreover, depth cameras offer advantages such as robustness to lighting variations, camera positioning, simplified data compression, and enhanced privacy, making them a promising alternative to RGB cameras.

Methods Experts and novice surgeons completed two simulators of open suturing. We focused on hand and tool detection and action segmentation in suturing procedures. YOLOv8 was used for tool detection in RGB and depth videos. Furthermore, UVAST and MSTCN++ were used for action segmentation. Our study includes the collection and annotation of a dataset recorded with Azure Kinect.

Results We demonstrated that using depth cameras in object detection and action segmentation achieves comparable results to RGB cameras. Furthermore, we analyzed 3D hand path length, revealing significant differences between experts and novice surgeons, emphasizing the potential of depth cameras in capturing surgical skills. We also investigated the influence of camera angles on measurement accuracy, highlighting the advantages of 3D cameras in providing a more accurate representation of hand movements.

Conclusion Our research contributes to advancing the field of surgical skill assessment by leveraging depth cameras for more reliable and privacy evaluations. The findings suggest that depth cameras can be valuable in assessing surgical skills and provide a foundation for future research in this area.

Keywords Surgical training · Computer vision · Depth camera · Object detection · Open surgery

Introduction

The complexity and high-stakes nature of open surgery necessitate the development of reliable and robust systems for evaluating surgical skills [1]. The evaluation of surgical skills has been an active area of research, with methodologies

ranging from subjective assessments by expert surgeons to objective metrics using sensors and data analytics.

Studies have shown the capability of motion sensors to distinguish between expert and novice surgeons. For instance, novices tend to move their hands with less efficiency, resulting in longer path lengths [2]. Additionally, they exhibit slower movements [3] and employ a more expansive working volume [4]. Unfortunately, hand sensors come with drawbacks such as high costs and discomfort. Furthermore, their integration into the operating room environment poses significant challenges.

The combination of RGB cameras and computer vision provides a new approach for assessing surgical skills. Goldbraikh et al. [5] utilized a standard webcam in combination with object detection to track hand movements, showing significant differences between students and experts. Addi-

✉ Ido Zuckerman
ido.z@campus.technion.ac.il

¹ Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology, Haifa 3200003, Israel

² Department of Surgery, University of Wisconsin-Madison School of Medicine and Public Health, 600 Highland Ave, Madison, WI 53792, USA

³ Clinical Simulation Program, University of Wisconsin Hospitals and Clinics, 600 Highland Ave, Madison, WI 53792, USA

tionally, multiple studies have highlighted the evaluation of surgical skills through the use of RGB cameras. [6] investigated the monitoring of tool usage in surgical videos, while [7] analyzed surgical proficiency using computer vision techniques in laparoscopic surgeries. This technique paves the way for the creation of simple and accessible training systems, providing learners with the opportunity to practice independently and receive objective feedback. However, motion captured by an RGB camera is the 2D in-plane of the camera rather than the actual 3D distance. The measurements can vary significantly if the camera's angle in relationship to the suture area changes. This study aims to investigate the potential of depth cameras to address this limitation, proposing a method that not only resolves this issue but also preserves the simplicity and accessibility of the training systems.

The use of RGB cameras is not limited to object detection and motion analysis. In recent years deep learning techniques have been used for general tasks such as tool detection in laparoscopic surgeries [8] or surgical gesture recognition [9]. Additionally, other studies have harnessed computer vision to formulate task-specific performance metrics [10, 11]. Therefore, RGB cameras may have a broad impact on the quality, efficiency, and safety of surgical procedures.

Nevertheless, using RGB cameras, especially in a clinical scenario, poses several challenges. First, privacy concerns emerge due to factors like capturing facial details and text. Second, lighting in the operating room is very challenging [12], as there is a wide variation in the amount of light in different areas [13, 14]. Depth cameras have been suggested as an alternative to RGB cameras to overcome these issues [15, 16]. They require no contact with the operating environment while still being capable of accurately tracking real-life hand motion data. They may be used to measure pose estimation and gait analysis [17] as well as patient activity recognition [18].

This study introduces an approach that employs depth cameras to automatically evaluate open surgery skills, specifically focusing on hand and tool detection and action segmentation in suturing procedures. We show that depth cameras can achieve comparable results to RGB cameras in a more robust way and provide an alternative approach for assessing surgical skills. The paper's main contributions are: (1) demonstrating that depth cameras are as effective as RGB cameras for object detection and action segmentation; (2) analyzing how the angle between the camera and suture area can affect the accuracy of their results, thus demonstrating the advantage of depth data; (3) introducing a novel metric that relies solely on depth cameras.

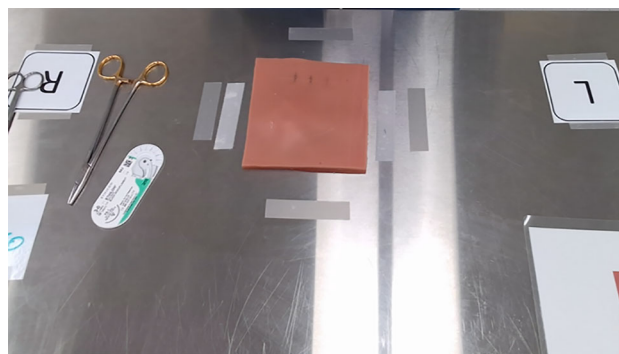


Fig. 1 Suture pad

Methods

The dataset

The study included 28 participants: 22 first-year surgical residents (8 male and 14 female) and 6 attending surgeons (3 male and 3 female) at a Midwestern academic hospital. The residents participated in this study as part of an annual surgical intern simulation series in which all first-year surgical residents complete a selection of basic surgical skills. Each participant was informed of the research prior to the session, and their decision to participate had no influence on the simulation series. One week before the simulation series, each intern received a video showing a faculty member accurately performing each skill. There was no limit on the amount of video views. During the simulation series, each intern is paired with a faculty member in a room within the hospital simulation center. The intern is then given standardized written instructions with scoring metrics and asked to complete each skill using a simulator. After task completion, the faculty member provides feedback to the resident.

The participants were engaged in conducting various surgical tasks utilizing two simulators: a “Suture pad” and a “Fascia Closure”. The execution of these tasks was documented through an Azure Kinect, which features a 4K RGB camera, a Depth Camera, and an IR Camera.

The first simulator, the “Suture Pad” simulator Fig. 1, was made of silicone. It was constructed to resemble human tissue and allows trainees to practice basic suturing techniques, such as creating knots and closing incisions. This simulator is similar to the simulator presented in [19]. In this study, participants executed four tasks using this simulator: simple suture, horizontal mattress suture, vertical mattress suture, and running suture. The goal was to train and assess medical professionals in the technique of suturing wounds. The initial task averaged 54 s, the second task 84 s, the third 81 s, and the final task approximately 206 s.

The second simulator, the “Fascia Closure” simulator Fig. 2, simulated the process of suturing and closing the con-



Fig. 2 Fascia closure

nective tissue layer called fascia during surgical procedures. This simulator design is taken from Mayo Clinic’s Surgical Olympics where it has been used since 2006 [20]. It provides a simulation of the fascia layer, enabling trainees to practice the skills required for successful closure. This simulator focuses on running suture in a distinct type of tissue. This simulator took an average of 379 s.

We focused on several computer vision tasks: object detection of the hands and tools as well as temporal segmentation of the surgical gestures. For object detection, about 900 frames from simulator 1 and an additional 900 frames from simulator 2 were annotated. These frames were drawn from 14 participants, who were chosen randomly, at a rate of one frame every 5 s. The number of participants was limited to 14 due to constraints on labeling resources. In each frame, all present tools, hands, and the simulator itself were marked with bounding boxes. The tools identified include Needle Driver, Tissue Forceps, Dressing Forceps, and Scissors. Additionally, the entire video set was annotated for temporal segmentation. The catalog of gestures consists of “Holding needle with a tool”, “Needle passing”, “Pull the suture”, “Instrumental tie”, “Lay the knot”, “Cut the suture”, “No Gesture”, and “Hand tie”.

We converted the depth matrix from each frame of the depth camera videos into grayscale video to better visualize and analyze the spatial information. In this format, objects nearest to the camera are represented in white, while those at a greater distance appear black, thereby simplifying the representation of depth information. The annotations applied to the RGB videos were also used for these depth videos.

Hardware and software

We conducted all our experiments, including training, testing, and evaluations, using a hardware setup consisting of two NVIDIA RTX A6000 GPUs and a single Intel Core i9-10940X CPU equipped with 28 logical cores. For running these experiments, we employed the PyTorch library, and for experiment tracking, we utilized WANDB [21].

Object detection

For the purpose of object detection, encompassing tools, hands, and the simulator, the YOLOv8 architecture [22] was employed. This architecture was trained using the Ultralytics framework. Notably, four distinct models were trained for each simulator scenario: one for RGB data encompassing all tools, another for depth data encompassing all tools, a third for RGB data focusing solely on hands, and a fourth for depth data dedicated to hands. In the hand-focused models, we had two classes: “Right Hand” and “Left Hand”. While, for the models aimed at detecting all tools, we used: “Right Hand”, “Left Hand”, “Needle Driver”, “Tissue Forceps”, “Dressing Forceps”, “Scissors”, and “Simulator”.

During the training process, several data augmentation techniques were applied to enhance the model’s robustness, including rotations, and brightness adjustments.

Modifications were made to the prediction head of the model to accommodate the aforementioned classes. The evaluation of the model’s performance was carried out using the mean average precision (mAP) based on intersection over union (IoU). Finally, the trained model was applied to extract per-frame bounding boxes, employing a confidence threshold ranging from 0.5 to 0.95, with increments of 0.05, with AdamW optimizer. Every epoch took an estimated one minute to complete, and the model was trained for a total of 300 epochs. The memory footprint was about 19GB.

Action segmentation

In our action segmentation experiments, we employed two architectures: UVASt [23] and MSTCN++ [24]. These architectures were selected because they complement each other effectively. MS-TCN++ is a lighter and less complex model, suitable for online, real-time inference. Conversely, UVASt, being a more feature-rich and complex model, offers greater accuracy but requires longer inference times. Both models were implemented using their original frameworks as described in the cited papers and trained using an Adam optimizer. Both architectures leveraged RGB and optical flow features, extracted using the I3D model [25] trained on the Kinetics 400 dataset [26]. Specifically for depth videos, we initially converted them to grayscale and then adapted them to RGB format by triplicating each frame for compatibility with the I3D model. In terms of training, MSTCN++ was trained over 100 epochs, with each epoch averaging around 5 s, while UVASt was trained 600 epochs, each averaging about one minute. The memory footprint was about 5GB when extracting features and 3GB for model training and prediction.

The UVASt architecture incorporated the Viterbi algorithm [27] during the inference stage. For experiments on simulator 2, due to limited resources and the longer video

Table 1 Suture pad and fascia closure simulators—all tools and hands

Class	Suture Pad Simulator			Fascia Closure Simulator		
	Occurrence	AP_{50-95}^{RGB}	AP_{50-95}^{Depth}	Occurrence	AP_{50-95}^{RGB}	AP_{50-95}^{Depth}
Left Hand	316	0.967	0.964	352	0.935	0.937
Right Hand	306	0.953	0.942	332	0.944	0.973
Needle Driver	295	0.931	0.922	313	0.915	0.882
Tissue Forceps	299	0.648	0.634	246	0.349	0.290
Dressing Forceps	273	0.792	0.819	297	0.646	0.506
Scissors	298	0.932	0.927	287	0.816	0.814
Simulator	309	0.999	0.999	353	0.989	0.981
Average	–	0.890	0.888	–	0.830	0.801

Table 2 Suture pad and fascia closure simulators—only hands

Class	Suture Pad Simulator			Fascia Closure Simulator		
	Occurrence	AP_{50-95}^{RGB}	AP_{50-95}^{Depth}	Occurrence	AP_{50-95}^{RGB}	AP_{50-95}^{Depth}
Left Hand	316	0.980	0.965	352	0.955	0.964
Right Hand	306	0.972	0.961	332	0.933	0.968
Average	–	0.976	0.963	–	0.945	0.966

durations, we limited the hypothesis space to 10,000 during inference.

In the Suture Pad simulator, the first three tasks of the suturing are closely related as they all present the execution of what we term a “stationary knot”—a simple suture, a horizontal mattress suture, and a vertical mattress suture. Consequently, the model was trained on a unified dataset that included these tasks, with each task separated into a distinct video. These videos were then divided into four separate train-test splits, ensuring that all videos from a single participant fell within the same split. The results for this model will be labeled as “Simple Suture”. For the fourth task, in the Suture Pad simulator, the results will be categorized under “Running Suture”, while results from the Fascia Closure simulator will be designated as “Fascia”.

We used distinct action segmentation labels for each simulator. For the suture pad simulator, labels included G0 for holding the needle with a tool, G1 for needle passing, G2 for pulling the suture, G3 for instrumental tie, G4 for laying the knot, G5 for cutting the suture, and G6 for no action. The fascia closure simulator employed similar labels, with the addition of G7 for hand tie.

As previously established in the literature [23, 24], three evaluation metrics were employed. Frame-wise accuracy, segmental edit distance, and F1@k for $k \in \{10, 25, 50\}$. Frame-wise Accuracy assesses the ratio of correctly classified frames to total frames. Segmental Edit Distance, adapted from the Levenshtein distance, compares activity segments and is normalized by the greater length between ground truth and prediction. F1@k calculates the Intersection over Union

(IoU) for each segment, categorizing them as true or false positives based on a threshold k .

3D hand path length

According to [28], the 3D hand path length metric is used to evaluate surgical skills by measuring the efficiency of a surgeon’s movements. Shorter, more direct paths typically indicate higher skill and experience because they reflect a surgeon’s ability to perform movements more efficiently and precisely, making this metric a valuable tool for assessing and improving surgical proficiency.

To quantify the path length traversed by the hands in a three-dimensional space, we employed the object detection algorithm Sect. “Object detection” to identify the hands. Subsequently, we extracted the coordinates of the bounding box’s center. Utilizing the depth camera provided by the Azure Kinect, we transformed the depth information into a point cloud using the Open3D (O3D) library. The coordinates of the bounding box were then used to extract the [x,y,z] coordinates from the point cloud, representing each hand’s spatial location. By aggregating these spatial coordinates across frames, we calculated the total path length using Euclidean distance metrics. For the statistical analysis, we adopted the Wilcoxon rank-sum test to compare the total path lengths between the two groups (experts and residents). The significance level was set at $p < 0.05$.

In our previous work [11], we explored temporal data obtained through action segmentation tools and examined its correlation with skill. In the current work, we extend our investigation into spatial data. Specifically, we introduce a

Table 3 Suture pad and fascia closure simulators—action segmentation results

		Simple Suture			Running Suture			Fascia								
		F1@ {10, 25, 50}	Edit	Acc	F1@ {10, 25, 50}	Edit	Acc	F1@ {10, 25, 50}	Edit	Acc						
RGB	MS-TCN++	80.43	77.38	65.29	77.78	72.40	66.74	61.89	44.96	62.56	62.09	77.53	74.68	62.66	72.21	75.24
	UVAST + Viterbi	83.84	81.39	67.63	81.57	74.92	69.94	65.65	52.04	70.26	64.73	73.66	70.41	59.33	69.47	71.69
Depth	MS-TCN++	82.63	80.40	69.56	79.13	76.75	71.96	68.02	52.33	67.83	66.98	70.22	66.71	53.33	63.36	67.20
	UVAST + Viterbi	84.19	82.55	71.47	79.74	78.22	75.21	72.26	56.69	71.80	70.97	66.20	63.67	51.58	62.90	65.24

novel metric to quantify the average distance a surgeon's hands move during each unique gesture. This approach will facilitate the provision of more focused practice recommendations, honing in on gestures that require further refinement.

2D different angles

In order to investigate the influence of RGB camera angles on measurement accuracy, our approach involved analyzing the movement of hands from different angles, emphasizing how each angle uniquely captures aspects of the movement in 3D space. This approach underscores our aim to demonstrate the superiority of depth cameras, which provide 3D imagery, over RGB cameras that offer only 2D perspectives. To accomplish this, we determined the geometric center of each hand for every frame. We then computed the [x, y, z] coordinates representing the hand's spatial position within the simulator's point cloud, a methodology previously established in Sect. "3D hand path length". These 3D coordinates were then projected onto three orthogonal 2D planes: XY, YZ, and XZ. This projection onto 2D planes serves to mimic the limited perspective of RGB cameras. By comparing these projections, we aim to highlight the constraints of 2D imaging in capturing the full complexity of hand movements in 3D space.

Subsequent to the projection, we quantified the distances covered by the hand within these 2D planes as if they were captured by a 2D camera. This comparison is critical for demonstrating that depth cameras, with their 3D imaging capabilities, provide a more comprehensive and accurate representation of hand movements in 3D space than 2D RGB cameras.

Results

Object detection

This section presents the results of the YOLOv8 algorithm applied to object detection. Tables 1 and 2 provide a detailed overview of the algorithm's performance, specifically in terms of average precision (AP) for each class. The evaluation was conducted on two distinct models: one trained on RGB video data and the other on depth video data. These models were rigorously tested on a separate test set comprising 313 frames for the first simulator and 354 frames for the second from different participants, ensuring that the model's performance was evaluated on previously unseen data.

In the first simulator, the suture pad simulator, for the model trained on RGB video data using all the tools and the hands, we obtained $mAP_{50-95}(RGB)$ of 0.890. Similarly, for the same model trained on depth video data, the corresponding $mAP_{50-95}(Depth)$ was found to be 0.888.

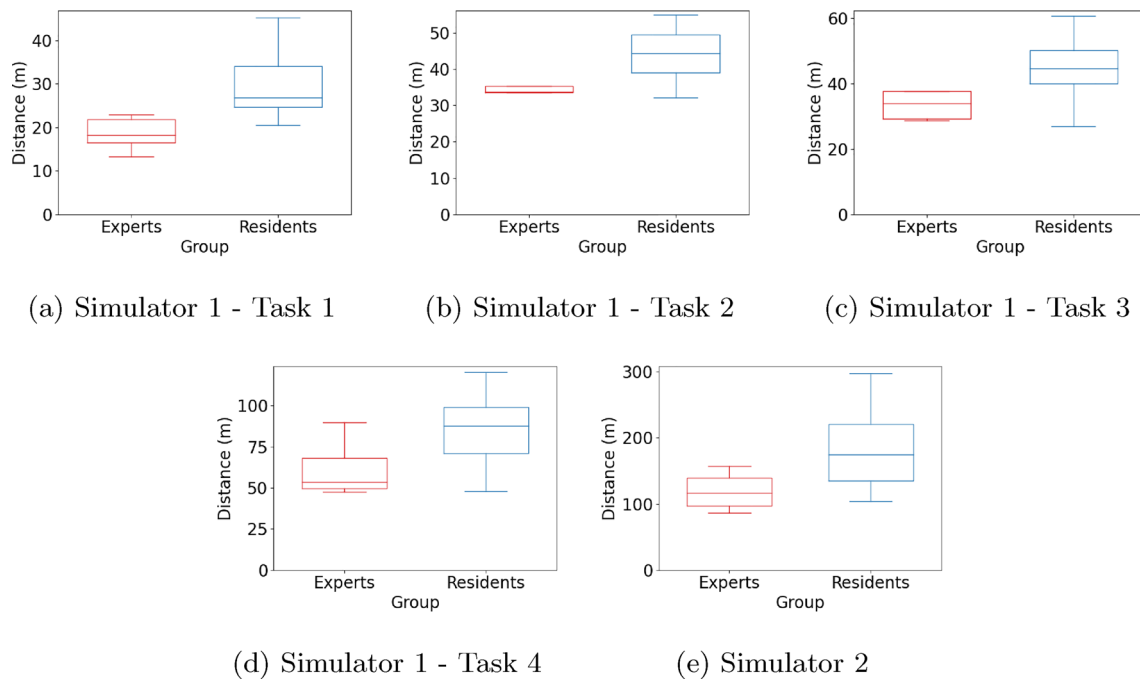


Fig. 3 Task-based box-plot of hand path length for residents and experts

For the models trained only on the hands, we obtained $mAP_{50-95}(RGB)$ of 0.976 and $mAP_{50-95}(Depth)$ of 0.963. These results highlight the consistency in the models' performance despite being trained on distinct data types, emphasizing the similar value that the depth camera gives us.

The second simulator, the fascia closure simulator, the model trained on RGB video data, exhibited performance with a $mAP_{50-95}(RGB)$ of 0.830. The model trained on depth video data for all the tools and the hands achieved a $mAP_{50-95}(Depth)$ of 0.801. For the models trained only on the hands, we obtained $mAP_{50-95}(RGB)$ of 0.945 and $mAP_{50-95}(Depth)$ of 0.966.

Action segmentation

In the case of the Suture Pad simulator, models trained using depth features outperformed others across all evaluation measures, with the sole exception being UVASt's marginally higher edit score in the simple suture task. Using depth features, UVASt attained an accuracy of 78.22% for the Simple Suture and 70.97% for the Running Suture. At the same time, MS-TCN++ achieved accuracies of 76.75% and 66.98% for the same tasks, outperforming their respective RGB-based versions.

In the case of the Fascia Closure simulator, models trained using RGB showcase higher results across evaluation metrics, MS-TCN++ achieving an accuracy of 75.24%, and UVASt achieving an accuracy of 71.69%. Nonetheless, as

indicated in Table 3, these results remain comparable to those achieved using depth features.

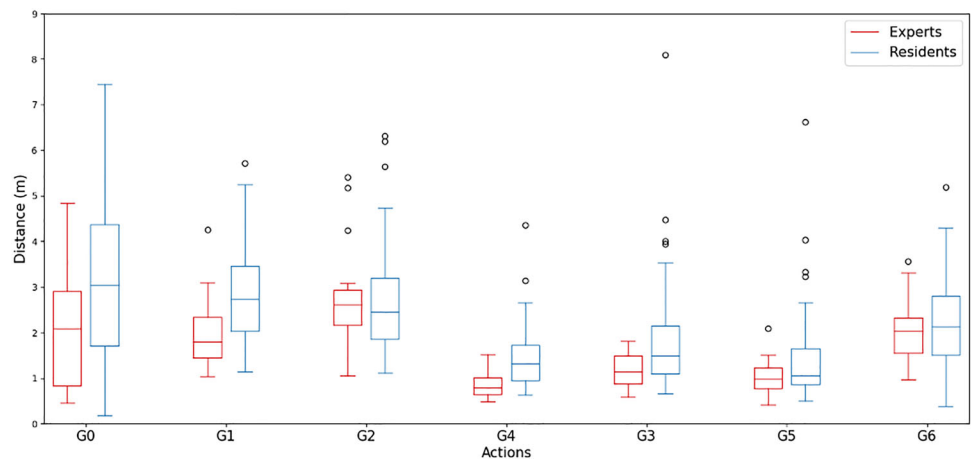
3D hand path length

As we expected, and shown for the 2D case in [5], the box plots in Fig. 3 reveal a consistent pattern across all tasks. Experts consistently navigated a shorter hand path compared to residents. In the suture pad simulator, for Task 1—sub-figure Fig. 3a, the p -value of the Wilcoxon rank-sum test was 0.003, for Task 2—sub-figure Fig. 3b it was 0.038, for Task 3—sub-figure Fig. 3c the p -value was 0.021, for Task 4—sub-figure Fig. 3d the p -value was 0.021, and for the fascia closure simulator two—sub-figure Fig. 3e the p -value was 0.038. All the p -values are $p < 0.05$.

These results indicate that the differences in hand path length between experts and residents are statistically significant, similar to [29], underscoring the value of expertise in surgical efficiency that can be captured using a depth camera. Additionally, it is noteworthy that despite the small sample size in our datasets, we were able to achieve statistically significant p -values. This fact further reinforces the validity of our results, highlighting the robustness of our findings even with limited data.

Figure 4 serves as an initial exploration based on data collected from the Simple Suture simulator and gives us a more nuanced look at the hands' path length, showing which gestures require the most movement. Our results show a statistically significant difference in the distance passed when

Fig. 4 Gesture distance for experts and residents in the simple suture task



passing the needle ($p = 0.001$), tying a knot ($p < 0.001$), laying a knot ($p = 0.008$), and holding the needle ($p = 0.021$). No significant difference was found when pulling the suture ($p = 0.707$), cutting the suture ($p = 0.056$), and for the distance moved when no action is performed ($p = 0.283$). This offers an initial validation for our proposed gesture distance metric.

2D different angles

Our data in Table 4 quantitatively confirmed these visual distortions, revealing a correlation between camera angle and measurement error. We found that certain angles disproportionately amplified or minimized specific types of movement, thereby providing an inaccurate representation of the true hand path. It becomes clear from these results that because the hand's path distance in RGB video is always calculated along a 2D plane, results are inherently subject to variance due to camera angles. This issue can lead to a loss of up to a third of the actual, real-world data, as demonstrated in the XY case under Table 4. This presents an inherent problem when using 2D cameras for applications that demand high accuracy and reliability.

In contrast, 3D cameras offer a solution to this issue by capturing the real-world position of the hands in a three-dimensional space, thereby eliminating the distortions introduced by varying camera angles. This allows for a more authentic and nuanced understanding of hand movements, as it captures the complete spatial relationships between different points in the hand's path.

Discussion and conclusion

Object detection and action segmentation are fundamental tasks in the surgical data science domain. Traditionally these tasks are performed using RGB cameras. However,

RGB cameras have several limitations including sensitivity to OR lighting, positional variations and privacy. These limitations may be overcome, by using depth cameras. This study aimed to determine if depth cameras, which may address these issues, are suitable for these tasks. RGB cameras pose considerable privacy concerns. They capture images that can include faces and confidential text, potentially exposing sensitive information. This becomes particularly problematic in environments where privacy is crucial, such as in healthcare settings. Using depth cameras these issues are overcome. They provide a three-dimensional map of the scene, emphasizing spatial relationships and movement without compromising sensitive visual details, hence overcoming these privacy issues. Motion economy is known indicator of expertise. Previous research has differentiated between experts and novices by detecting the surgeon's hand movements [5]. However, RGB cameras only capture a 2-dimensional representation of the actual hand motion path, which can be affected by the camera's position relative to the simulator. Depth cameras overcome this limitation by providing the true 3-dimensional motion. Specifically, we calculated the motion values as if an RGB camera provided sagittal, coronal, or transverse views of the simulation (corresponding to the YZ, XZ, and XY planes). Our results showed that while the 2-dimensional path lengths of experts and novices are significantly different in each view, the values overlap when compared across different views. For instance, if a study uses a sagittal view to establish baseline performance for experts and novices, a training program using this baseline must ensure identical camera positioning. Depth cameras eliminate this requirement by providing absolute motion data.

It is essential to acknowledge the limitations of our study. The data set size is crucial when using deep learning tools and statistical tests. More accurate results could have been achieved with the availability of a more extensive dataset. Due to the high memory demands of the Viterbi algorithm, we

Table 4 Hand path length in the planes XY, YZ and XZ (Meter) divided to experts and residents

	XYZ		XY		YZ		XZ	
	Experts	Residents	Experts	Residents	Experts	Residents	Experts	Residents
Suture Pad								
Task 1	18.52 ± 3.51	29.95 ± 7.59	11.00 ± 1.65	17.77 ± 5.28	16.06 ± 3.20	26.30 ± 6.88	17.43 ± 3.18	28.11 ± 7.16
Task 2	32.97 ± 6.76	45.94 ± 11.86	19.36 ± 4.07	27.69 ± 9.18	28.77 ± 6.03	40.17 ± 10.71	31.05 ± 6.41	43.00 ± 11.04
Task 3	33.39 ± 3.88	45.70 ± 11.98	19.38 ± 2.43	27.41 ± 8.04	29.41 ± 3.78	40.28 ± 10.79	31.53 ± 3.53	42.45 ± 11.09
Task 4	61.72 ± 15.71	90.20 ± 26.10	57.83 ± 15.01	85.13 ± 23.85	19.38 ± 2.43	27.41 ± 8.04	29.41 ± 3.78	40.28 ± 10.79
Fascia Closure	119.38 ± 27.64	180.71 ± 52.59	74.39 ± 15.81	125.18 ± 37.06	102.37 ± 22.92	154.63 ± 44.04	111.61 ± 25.53	169.57 ± 48.52

had to use a simplified version. This is a balanced approach and a common solution in applications involving the Viterbi algorithm. While the full algorithm might provide slightly more precision in certain cases, the simplified version aligned well with our research requirements. Also, while a larger dataset could potentially offer finer details, rigorous methods were employed to ensure the validity of our study given the available data.

In conclusion, our research contributes to the field of surgical skill assessment. By championing the adoption of depth cameras, we provide a more accurate, privacy-conscious, and robust approach to evaluating surgical proficiency. The advantages of depth cameras, combined with our empirical findings, underscore their potential to alter how surgical skills are assessed and trained, offering a solid foundation for future advancements in this domain.

Acknowledgements This research was funded by a grant from the University of Wisconsin–Madison Department of Surgery. The authors would like to thank Abbie Schafer for assisting in data annotation.

Declarations

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

Ethical approval This study was granted an exemption by the University of Wisconsin–Madison Institutional Review Board.

Informed consent Informed consent was obtained from all individual participants included in the study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Reznick RK, MacRae H (2006) Teaching surgical skills-changes in the wind. *New Engl J Med* 355(25):2664–2669
2. Dosis A, Aggarwal R, Bello F, Moorthy K, Munz Y, Gillies D, Darzi A (2005) Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Arch Surg* 140(3):293–299
3. Smith S, Torkington J, Brown T, Taffinder N, Darzi A (2002) Motion analysis: a tool for assessing laparoscopic dexterity in the performance of a laboratory-based laparoscopic cholecystectomy. *Surg Endosc* 16:640–645

4. D'Angelo A-LD, Rutherford DN, Ray RD, Laufer S, Mason A, Pugh CM (2016) Working volume: validity evidence for a motion-based metric of surgical efficiency. *Am J Surg* 211(2):445–450
5. Goldbraikh A, D'Angelo A-L, Pugh CM, Laufer S (2022) Video-based fully automatic assessment of open surgery suturing skills. *Int J Comput Assist Radiol Surg* 17(3):437–448
6. Al Hajj H, Lamard M, Conze P-H, Cochener B, Quéllec G (2018) Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Med Image Anal* 47:203–218
7. Funke I, Mees ST, Weitz J, Speidel S (2019) Video-based surgical skill assessment using 3d convolutional neural networks. *Int J Comput Assist Radiol Surg* 14:1217–1225
8. Fathabadi FR, Grantner JL, Shebrain SA, Abdel-Qader I (2021) Multi-class detection of laparoscopic instruments for the intelligent box-trainer system using faster r-cnn architecture. In: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMII), pp. 000149–000154 . IEEE
9. Goldbraikh A, Avisdris N, Pugh CM, Laufer S (2022) Bounded future ms-tcn++ for surgical gesture recognition. In: European Conference on Computer Vision, pp. 406–421 . Springer
10. Halperin L, Sroka G, Zuckerman I, Laufer S (2023) Automatic performance evaluation of the intracorporeal suture exercise. *International Journal of Computer Assisted Radiology and Surgery*, 1–4
11. Bkheet E, D'Angelo A-L, Goldbraikh A, Laufer S (2023) Using hand pose estimation to automate open surgery training feedback. *International Journal of Computer Assisted Radiology and Surgery*, 1–7
12. Dascalaki EG, Gaglia AG, Balaras CA, Lagoudi A (2009) Indoor environmental quality in hellenic hospital operating rooms. *Energy Build* 41(5):551–560. <https://doi.org/10.1016/j.enbuild.2008.11.023>
13. Likitlersuang J, Sumitro ER, Theventhiran P, Kalsi-Ryan S, Zariffa J (2017) Views of individuals with spinal cord injury on the use of wearable cameras to monitor upper limb function in the home and community. *J Spinal Cord Med* 40(6):706–714
14. Haque A, Milstein A, Fei-Fei L (2020) Illuminating the dark spaces of healthcare with ambient intelligence. *Nature* 585(7824):193–202
15. Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, Liu J (2023) Human action recognition from various data modalities: a review. *IEEE Trans Pattern Anal Mach Intell* 45(3):3200–3225. <https://doi.org/10.1109/TPAMI.2022.3183112>
16. Yeung S, Rinaldo F, Jopling J, Liu B, Mehra R, Downing NL, Guo M, Bianconi GM, Alahi A, Lee J, Campbell B, Deru K, Beninati W, Fei-Fei L, Milstein A (2019) A computer vision system for deep learning-based detection of patient mobilization activities in the icu. *NPJ digital medicine* 2(1):11
17. Martinez-Martin N, Luo Z, Kaushal A, Adeli E, Haque A, Kelly SS, Wieten S, Cho MK, Magnus D, Fei-Fei L, Schulman K, Milstein A (2021) Ethical issues in using ambient intelligence in health-care settings. *Lancet Digit Health* 3(2):115–123
18. Siddiqi MH, Almashfi N, Ali A, Alruwaili M, Alhwaiti Y, Alanazi S, Kamruzzaman M (2021) A unified approach for patient activity recognition in healthcare using depth camera. *IEEE Access* 9:92300–92317
19. Williams TP, Snyder CL, Hancock KJ, Iglesias NJ, Sommerhalder C, DeLao SC, Chacin AC, Perez A (2020) Development of a low-cost, high-fidelity skin model for suturing. *J Surg Res* 256:618–622
20. Buckarma E (2016) The how to book of low cost surgical simulation. <https://surgicaleducation.mayo.edu/how-to-book/>
21. Biewald L (2020) Experiment tracking with weights and biases. Software available from <https://www.wandb.com/>
22. Jocher G, Chaurasia A, Qiu J YOLO by Ultralytics. <https://github.com/ultralytics/ultralytics>
23. Behrmann N, Golestaneh SA, Kolter Z, Gall J, Noroozi M (2022) Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: European conference on computer vision, pp 52–68 . Springer
24. Li S-J, AbuFarha Y, Liu Y, Cheng M-M, Gall J (2020) Ms-tcn++: multi-stage temporal convolutional network for action segmentation. *IEEE Trans Pattern Anal Mach Intell* 45:6647–6658
25. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
26. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The kinetics human action video dataset. arXiv preprint [arXiv:1705.06950](https://arxiv.org/abs/1705.06950)
27. Forney GD (1973) The viterbi algorithm. *Proc IEEE* 61(3):268–278
28. Chmarra MK, Jansen FW, Grimbergen CA, Dankelman J (2008) Retracting and seeking movements during laparoscopic goal-oriented movements. is the shortest path length optimal? *Surg Endosc* 22:943–949
29. Lefor AK, Harada K, Dosis A, Mitsuishi M (2020) Motion analysis of the JHU-ISI gesture and skill assessment working set using robotics video and motion assessment software. *Int J Comput Assist Radiol Surg* 15:2017–2025

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.