Gaze behavior is related to objective technical skills assessment during virtual reality simulator-based surgical training: a proof of concept

Soline Galuret, Nicolas Vallée, Alexandre Tronchot, Herve Thomazeau, Pierre Jannin, Arnaud Huaulmé

Gaze behavior is related to objective technical skills assessment during virtual reality simulator based surgical training: a proof of concept

Soline Galuret[1], Nicolas Vallée[1,2], Alexandre Tronchot[1,2], Hervé Thomazeau[1,2], Pierre Jannin[1], Arnaud Huaulmé[1]

Affiliations

[1] Univ. Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France
[2] Orthopedics and Trauma Department, Rennes University Hospital, F-35000 Rennes, France

E-mails (ORCID number)

soline.galuret@univ-rennes1.fr (0000-0001-6589-9051); nicolas.vallee@chu-rennes.fr (0000-0002-2885-2091); alexandre.tronchot@chu-rennes.fr (0000-0002-3092-3282); herve.thomazeau@chu-rennes.fr (0000-0003-1730-7597); pierre.jannin@univ-rennes1.fr (0000-0002-7415-071X); arnaud.huaulme@univ-rennes1.fr (0000-0002-7844-5259)

Abstract.

Purpose: Simulation-based training allows surgical skills to be learned safely. Most virtual reality based surgical simulators address technical skills without considering non-technical skills, such as gaze use. In this study, we investigated surgeons' visual behavior during virtual reality based surgical training where visual guidance is provided. Our hypothesis was that the gaze distribution in the environment is correlated to the simulator's technical skills assessment.
Methods: We recorded 25 surgical training sessions on an arthroscopic simulator. Trainees were equipped with a head-mounted eye-tracking device. A U-net was trained on two sessions to segment three simulator-specific areas of interest (AoI) and the background, to quantify gaze distribution. We tested whether the percentage of gazes in those areas was correlated to the simulator's scores.
Results: The neural network was able to segment all AoI with a mean Intersection over Union superior to 94% for each area. The gaze percentage in the AoI differed among trainees. Despite several sources of data loss, we found significant correlations between gaze position and the simulator scores. For instance, trainees obtained better procedural scores when their gaze focused on the virtual assistance (Spearman correlation test, N=7, r=0.800, p=0.031).
Conclusion: Our findings suggest that visual behavior should be quantified for assessing surgical expertise in simulation-based training environments, especially when visual guidance is provided. Ultimately visual behavior could be used to quantitatively assess surgeons' learning curve and expertise while training on VR simulators, in a way that complements existing metrics.

Keywords: Eye-tracking; Non-technical skills; Deep learning; Surgical training; Orthopedic surgery

Statements and Declarations

Conflict of Interest
The authors declare that they have no conflict of interest.
Ethical approval
All procedures involving human participants were conducted 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
This article does not contain patient data.

Acknowledgement

This study is part of the French network of University Hospitals HUGO ("Hôpitaux Universitaires du Grand Ouest"). It was made possible thanks to the VirtaMed Society.

1. Introduction

Surgical simulators recreate surgical situations for trainees to learn and practice [1]. Virtual reality (VR) simulators offer high fidelity reusable environments [2] and the possibility to work on several procedures in a single device. VR simulation-based learning has been reported in many disciplines such as thoracic [3], urologic [4] and orthopedic [5] surgery. It has been demonstrated that VR training on arthroscopy significantly improves performance (e.g. time and camera alignment) after six months of training [6].

VR simulators usually give feedback on strategy anticipation and technical skills through quantitative metrics such as the time of completion and gesture accuracy [7]. However, non-technical skills (NTS), defined as cognitive (decision-making, planning, situational awareness), social (teamwork, communication, leadership) and personal resources (stress, fatigue) factors, are equally important to patient outcomes [8]. There is even evidence that the lack of NTS may lead to more frequent surgical incidents than surgical technique errors [9].

NTS are usually assessed using subjective scales such as Non Technical Skills for Surgeons [10], Non TECHnical skills [11] and Observational Teamwork Assessment for Surgery [12]. Some initial works recently proposed objective assessment of NTS relying on sensors, such as measuring posture via surgeon's joints kinematics [13] or workload via pupillometry [14].

Eye-tracking (ET) devices have been extensively used in the medical field. A recent review of the broad use of ET in surgical research reported 46 articles on skill assessment, 25 on visual attention, 19 on workload and 11 on skills training [15].

ET can be used as a non-invasive tool for assessing workload in clinical settings through pupil responses, gaze patterns and blinks [14]. ET has been used for training purposes [16]. It enables the observation of gaze and the use of visual cues of a surgeon, highlighting attentional gaze patterns. These can be used to draw the novices' visual attention to the locations looked at by the experts to improve skills acquisition [17].

Visual behavior is also used to assess the expertise level of surgeons [18]. Comparing eye movements and fixations showed that novice surgeons need more visual feedback regarding their tools' locations to complete a task [19]. Moreover, novices split their attention between the target and the surgical instruments, while experts tend to maintain their gaze on the target even when manipulating instruments [19]. It has also been demonstrated in orthopedic surgery training that participants with more surgical experience had a shorter duration and fewer fixations on the operative site [20]. However, the total number of participants was quite small (3 novice surgeons, 5 surgeons with an intermediate level and 4 experienced surgeons), and all participants did not use the same tools to perform the simulation. Moreover, it assumes that the surgeons with more experience have better technical skills in this simulation environment but it is not objectively measured.

Few studies attempted to correlate surgical skills with ET metrics. Pupillometry data was correlated with performance assessments of the Global Evaluative Assessment of Robotic skills (GEARS) and the numeric psychomotor test score (NPMTS) in both real tissue and a simulator on da Vinci surgical robot systems [21]. During laparoscopic cholecystectomy, time spent fixing on areas of interest (AoIs) was correlated with the Objective Structured Assessment of Technical Skill (OSATS) [22]. GEARS, NPMTS and OSATS are peer assessments and carry the risk of subjectivity. One study overcame this bias using a network of multiple camera sensors to assess the surgical trainees' performance. The predicted assessment was correlated with several gaze metrics such as the time spent looking at the surgical display [23].

The study of visual behavior during surgical training should be a required assessment, especially in a VR environment where simulators can provide visual guidance. This assistance is useful in a training context, but carries the risk of skewing the skills transfer to a real environment if the trainee is not able to perform the task without the visual aid. Accordingly, the gaze position should be quantitatively measured to assess trainee visual behavior during simulation-based training.

In this paper, we report on a study investigating surgical trainees' visual behavior during VR arthroscopic training. Using a head-mounted ET device, we aimed to determine what the trainees were looking at. We trained a deep neural network to segment the environment into 4 simulator-specific areas. We then quantified gaze distribution in each area. Finally, we tested whether such gaze behavior was correlated to the simulator's technical skills assessment.

2. Materials and methods

2.1. Data acquisition

2.1.1. Surgical simulator

Data were acquired from 25 initial training sessions on the VR simulator VirtaMed (ArthroS™, VirtaMed, Schlieren, Switzerland) performed by 23 participants. The participants came from 6 Western French hospital centers and were post-graduate year 4 to 6. Seven of them wore eyeglasses. The 25 sessions were performed in three simulation centers in a controlled environment (a closed room with only the participant and the supervising surgeon to avoid environmental distractions, and not during conferences).

The simulator consisted of a screen on which the simulation is performed (Figure 1), and a physical module in which the surgical instruments are inserted. Each session began with three instrument handling exercises, followed by a diagnostic task, and a double row cuff repair. It consisted of reattaching the tendon connecting the head of the humerus to the scapula using wired anchors. This is a minimally invasive surgical procedure where an arthroscope (a camera) is inserted into the shoulder joint [24]. We focused our study on this exercise since it involves multiple steps and multiple instruments.

The sessions lasted approximately 35 minutes, with an average of 15 minutes devoted to the cuff repair. All sessions were overseen by a single supervising surgeon, responsible for instrument preparation and data acquisition.

2.1.2. Eye-tracking device

The participants were equipped with the eyeglass-compatible ET device Pupil Core (Pupil Labs, Berlin, Germany). The device consisted of three cameras: two directed toward the eyes (120Hz, 192x192 pixels) and one placed on the forehead to record the user's point of view (world camera, 30Hz, 1280x720 pixels). The calibration was performed at the beginning of each session by the single marker choreography provided by Pupil Labs (i.e., the user fixes on a moving target). The ET device uses the position of the pupils (relative to the eyeball) from the images of the eye cameras to determine the position of the gaze on the image of the world camera at a given time. These gaze position estimations, as well as their confidence, are extracted in real time during the recording [25]. Both gazes (240Hz) and world recorded frames are associated with a timestamp, enabling the association of each gaze with the corresponding environment image.

2.2. Segmentation of simulator-specific areas

To study gaze distribution in the environment during a training exercise, we defined several AoIs in the world camera field of view (FoV), focusing on the simulator's screen. We used deep learning segmentation to automate the AoI recognition on the world camera video frames.

2.2.1. Dataset

The data consist of the images recorded by the world camera during the cuff repair. 888 images were sampled from the videos of two sessions (alpha and beta, Table 1) and were randomly distributed in training (70%) and validation datasets (30%).

Table 1: Number of images (before data augmentation) from sessions alpha and beta in the training and validation datasets. Percentages are shown relative to the total number of images of the training and validation datasets.

	Session alpha	**Session beta**	**Total**
Training	329	293	622 (70%)
Validation	134	132	266 (30%)
Total	463 (52%)	425 (48%)	888 (100%)

The test dataset was composed of 2,720 images sampled from the videos of the remaining 23 sessions. We choose to put the emphasis on the test dataset to assess whether the neural network was skewed toward session alpha's and beta's conditions (e.g. brightness, blurriness). In doing so, we ensured that the neural network performed equally for the sessions not used in the learning process.

720 images were extracted from the videos of 18 sessions, with a rate of 40 images per session. The 2,000 other images came from the video of 5 sessions with a rate of 400 images per session.

2.2.2. Annotations

Annotations were all performed using the CVAT annotation software. Three AoIs were annotated:

- "Arthroscope": the arthroscope point of view inside the articulation (left side of the screen);
- “Virtual shoulder" (outside view): internal and external views of the articulation and relative placement of the surgical instruments (center of the screen);
- “Information" (sidebar): list of tasks and access to instruments (right side of the screen).

A fourth area "Background" was set for the rest of the image. All the annotations of an image are called a segmentation mask (Figure 1 right panels). When the AoIs were partially covered by the supervising surgeon’s hand (Figure 1 upper panels), the hand was not contoured.

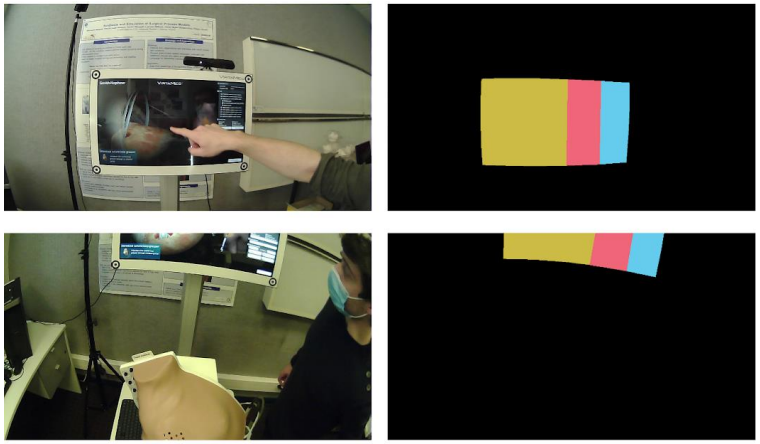

Figure 1: Original images from the frontal world camera (left panels) and their annotated segmentation mask (right panels). The Arthroscope area is in yellow (left side of the screen), the Virtual shoulder area is in red (center of the screen), the Information area is in blue (right side of the screen), and the Background area is in black.

Annotation was performed by 6 annotators. They annotated 30 supplementary images randomly taken to test the inter-annotator variation. The mean Intersection over Union (IoU) quantifying the similarity between two annotations was superior to 90% for all segmented areas between each pair of annotators. For each segmented area of each image, we also computed a consensus annotation based on the 6 annotators’ annotations using the Simultaneous Truth and Performance Level Estimation [26]. The mean IoU between the consensus annotations and each annotator’s annotations was superior to 96% for all segmented areas. Therefore, the variation between annotators was considered to be negligible.

Since the ET device is head-mounted, AoIs could be fully or partially out of the frame (Figure 1, bottom panels). Therefore, we ensured that each segmented area was correctly represented in each dataset (Table 2).

Table 2: Number of annotated images of each area to be segmented for each dataset. Percentages given are relative to the 3,608 annotated images.

	Training (622 frames, ≈17.2%)	**Validation (266 frames, ≈7.4%)**	**Test (2,720 frames, ≈75.4%)**
Background	622	266	2,720
Arthroscope	618	265	2,689
Virtual shoulder	619	265	2,688
Information	621	266	2,689

2.2.3. Learning

Before being processed by the network, each image of the training and validation datasets was augmented 5 times with random brightness, contrast and saturation adjustments. After augmentation, the training set consisted

of 3,732 (622×6) images, and the validation set of 1,596 (266×6) images. The images were also resized to 256×512 pixels.

The segmentation learning task was performed using a U-Net (resNet-34) pre-trained on ImageNet to extract spatial data [27] using an ADAM optimizer, a cross-entropy loss, a learning rate of 10^{-4} and a batch size of 50, with an early stopping after 10 non-improving epochs. The training was performed with a processor Intel Core i9-9940X (14 Cores, 3.30 GHz), RTX 5000, using Python 3.6.9, Cuda 11.4 and Pytorch 1.10.1.

The performance was evaluated using the mean IoU, also known as the Jaccard Index, of each class. The IoU is the overlapping area (i.e., intersection) between the prediction and the ground truth, divided by the union area between the prediction and the ground truth.

2.3. Gaze distribution in the environment

To study the distribution of the trainees' gaze in the simulation environment, we computed the percentage of gazes (%gaze) in each AoI.

2.3.1 Gaze selection

a) Gaze confidence

To ensure gaze detection quality, we used the detection confidence provided by the ET software. We used a Gaussian Mixture Model (GMM) to split the gaze data into several populations of confidence, the goal being to use the Gaussian parameters to set a confidence threshold high enough to make a proper selection while keeping sufficient data to analyze.

Based on the gaze confidence distribution of all trainees and the shape of the associated probability density function (Figure 2a), we estimated that the confidence distribution could be modeled as the sum of 4 Gaussians, whose parameters were estimated via fitting a 4-component GMM (Figure 2b). From the parameters of the 3rd component (mean μ=0.905 and standard deviation σ=0.077), we fixed our threshold at μ-σ on this component. The confidence threshold for estimating the gaze position was then set at 0.828 (μ-σ=0.905-0.077, Figure 2a, vertical line).

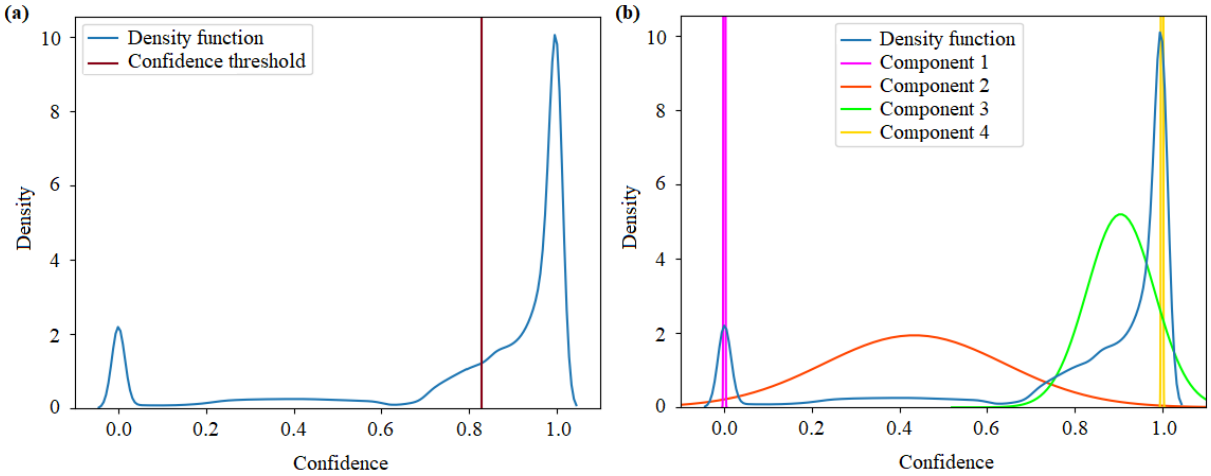

Figure 2: (a) Density function of the number of gazes according to the associated confidence and (b) 4-component GMM on this distribution.

b) Gazes and world camera matching

During acquisition, the world camera could be subject to freezes (i.e., the camera stopped recording new frames), skewing correct matches between the gaze and the closest world image. Since the ET device is head-mounted, this matching could be unrepresentative of the real gaze (Figure 3b). To ensure that the matching was not biased, we only kept the gaze whose timestamps belonged to a window of 1/30 second centered on the timestamp of each image of the world camera (Figure 3c). This window is the theoretical maximum temporal gap between 2 images of the world camera.

(a) Ground truth

Gaze T.	World T.
…	
877.013	
877.017	877.0
877.021	
877.025	
877.029	
877.033	
877.037	877.0
877.041	
877.045	
877.049	
877.053	
877.057	
877.061	877.0
877.065	
…	

(b) With world camera freeze without threshold

Gaze T.	World T.
…	
877.013	
877.017	877.0
877.021	
877.025	
877.029	
877.033	
877.037	
877.041	
877.045	
877.049	
877.053	
877.057	
877.061	877.0
877.065	
…	

(c) With world camera freeze with matching threshold

Gaze T.	World T.
…	
877.013	
877.017	877.0
877.021	
877.025	
877.029	
877.033	
877.037	Unassigned
877.041	
877.045	
877.049	
877.053	
877.057	
877.061	877.0
877.065	
…	

Figure 3: World camera freezes consequence on the pairing between a gaze and the *a priori* matching world camera image. (a) Ground truth, (b) without the matching threshold and (c) when applying the matching threshold. T. = timestamp.

c) Out-of-frame gazes

The gaze coordinates that were not in the FoV of the world camera image were not included in the analysis since it was impossible to know which area the trainee was looking at.

2.3.2. Simulator-specific areas

The AoIs were the Arthroscope, Virtual shoulder, Information and Background areas (see Section 2.2.2.). For each session, images were extracted from the world camera video, and processed by the network for segmentation.

2.3.3. Metric: percentage of gazes in simulator-specific areas

Using the timestamps, each gaze position was placed on the corresponding image of the environment in order to deduce the area looked at. The %gaze per AoI was computed for each session. We only considered the gazes with a sufficiently high confidence (confidence threshold, see Section 2.3.1.a), which could be reliably attributed to a world camera image (matching threshold, see Section 2.3.1.b) and with coordinates in the FoV of the world camera image (see Section 2.3.2.c).

2.4. Link between gaze position and simulator metrics

2.4.1. Variables

For each simulation, the simulator returns 20 metrics (e.g. procedure time, instrument path lengths, cartilage scratching) and 26 scores (based on the metrics compared to a target value) to objectively assess the surgical exercise. To investigate a link between gaze position and the metrics and scores computed by the simulator, we considered the %gaze within the Arthroscope (only view of the joint in real surgery) and the Virtual shoulder (a one-time visual aid in a simulator learning setting) areas.

Correlation tests were applied between the %gaze on the Arthroscope or on the Virtual shoulder areas, and each metric and score for the cuff repair exercise. We performed Pearson correlation tests for data with a normal distribution (Shapiro-Wilk test, $p > 0.05$) and Spearman correlation tests otherwise (Shapiro-Wilk test, $p \leq 0.05$).

2.4.2. Training sessions

To consider the whole process (segmentation included), both sessions used to train and validate the segmentation network were excluded from the correlation tests (23 sessions left).

To apply the correlation tests on data that were representative of the exercise, we placed a filter on the %gaze not included in the gaze distribution analysis (with a confidence ≤ 0.828, not assigned to, or outside the world camera FoV, see Section 2.3.1). The 15 sessions for which the gaze selection removed more than 30% of gaze data were not included in the analysis (8 sessions left).

Due to calibration issues, we excluded 1 session for which the %gaze on the Background area was higher than 50% as the surgeon was unlikely to complete the exercise without looking at the screen. Finally, 7 out of 23 sessions were considered for the correlation tests.

3. Results

3.1. Segmentation of simulator-specific areas

The cross-entropy loss did not show any overlearning (Figure 4a). The IoU averaged 98% on the training dataset and 96% on the validation dataset. Independently, each class reached an IoU of at least 94% (Figure 4b).

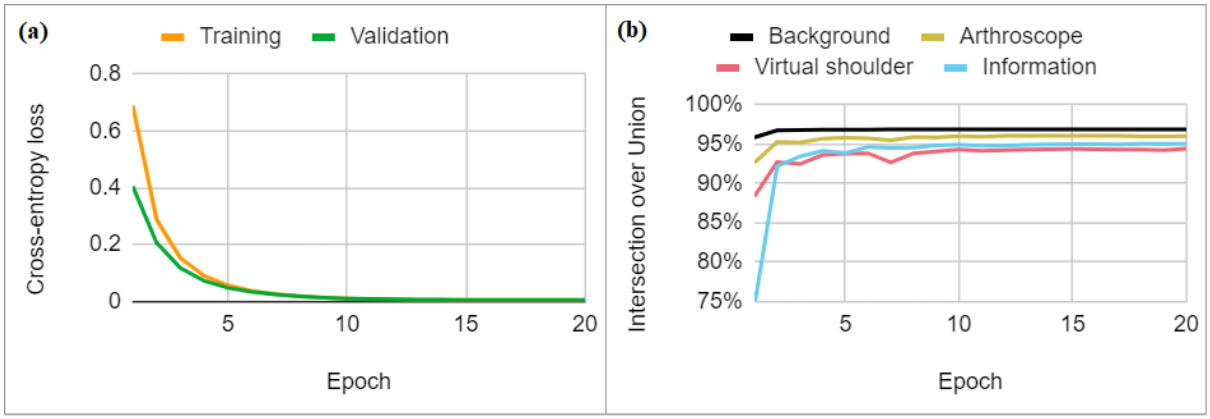

Figure 4: (a) Cross-entropy loss (N=3,732) and the validation (N=1,596) datasets and (b) Intersection over Union (%) for each class of the validation dataset during the training.

On the test dataset, the IoU averaged 96.59%, with 99.68% for the Background, 97.46% for the Arthroscope area, 93.72% for the Virtual shoulder area and 95.50% for the Information area. The segmentation performance was similar for the three simulation centers.

3.2. Gaze distribution in the environment

The number of gazes excluded from the analysis varied from one session to another. On average, the confidence threshold removed 32.90% ± 30.15% of the gaze, the matching threshold removed 17.54% ± 17.50% of the gaze, and the gaze outside the world camera FoV represented 4.09% ± 5.76% of the gaze. Note that the gaze selection (section 2.3.1) removed more than 30% of the data for 15 out of 23 sessions (Figure 5).

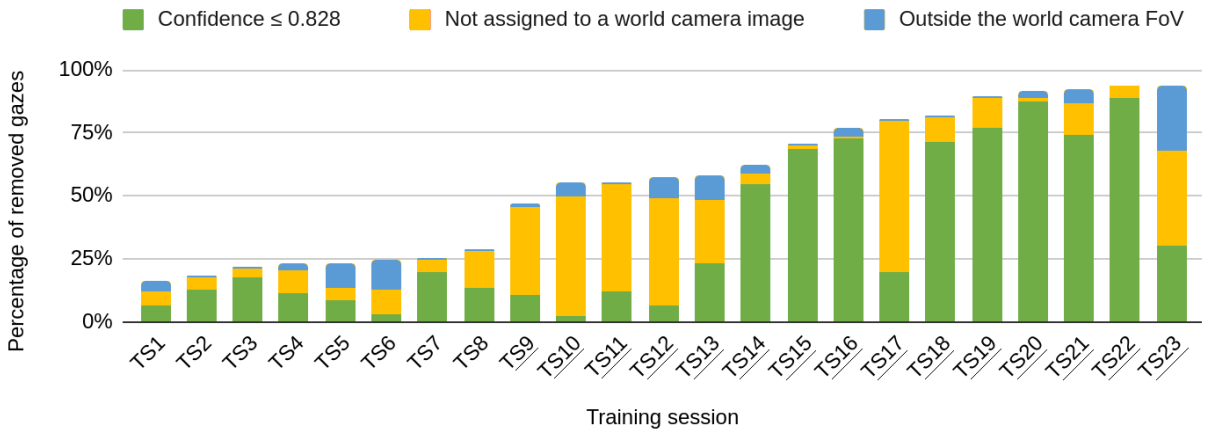

Figure 5: Gaze data loss. The underlined sessions are excluded for further analysis (more than 30% of data loss).

The gaze distribution within the AoIs varied among sessions (Figure 6). For instance, the %gaze in the Arthroscope area ranged from 16% up to 95%, with a mean of 78% and a standard deviation of 25%.

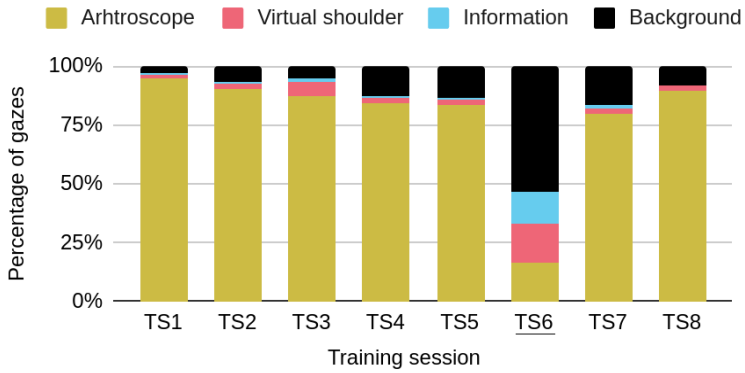

Figure 6: Percentage of gazes in the simulator-specific AoIs (after gaze selection, see Section 2.3.1) during the cuff repair exercise. The underlined session was excluded for further analysis (more than 50% of gaze on the Background area suggesting a calibration defect).

3.3. Link between gaze position and simulator metrics

There was no significant correlation before session selection. When we only focused on the 7 selected sessions (see Section 2.4.2), correlation was found between the %gaze in the Virtual shoulder and the Camera path length score, the Drilling attempts at incorrect location score and the Procedure score (Table 3). Note that the Camera path length score and Drilling attempts at incorrect location score are computed according to, respectively, the camera path length (cm) and the number of drilling attempts at incorrect location, compared to a target value. The Procedure score is based on 3 scores: the number of drilling attempts at incorrect location score and the maximal deviation from the optimal angle score for both anchor holes.

Table 3: Significant correlations between the percentage of gazes on the Virtual shoulder area and the simulator scores. Training sessions were selected according to (a) Data removed by gaze selection < 30% and (b) Percentage of gazes on the Background area < 50% to overcome calibration defect.
The values correspond to Spearman correlation coefficients and p-values. A slash means no significant correlation.

	Without training session selection	**With representative gaze data and satisfying calibration**
Number of training sessions	23	7
Camera path length score	/	r=0.837, p=0.019
Drilling attempts at incorrect location score	/	r=0.800, p=0.031
Procedure score	/	r=0.800, p=0.031

Furthermore, the higher the %gaze in the Arthroscope area, the higher the percentage of instruments in scope view (Pearson correlation test, N=7, r=0.808, p=0.028) for selected sessions.

4. Discussion

Segmentation of simulator-specific areas

The neural network segmented the three AoIs and the background with an IoU above 94% for each. Interestingly, its performance was similar for three different simulation centers, suggesting its robustness to the background environment.

Some images were complex to annotate due to (i) the lack of fixed delineation between the three screen areas, and (ii) the blurring of some images due to motion. In those cases, the network might be more objective than the annotators to predict the different classes. Although the network performance is sufficient as it is, its performance and learning speed could be refined through a fine-tuning step testing several parameters such as the learning-rate and the optimizer.

Link between gaze position and simulator metrics

Some of the metrics returned by the simulator were correlated with the %gaze in the Arthroscope and with the Virtual shoulder areas. For instance, the more the trainees looked at the Virtual shoulder, the lesser they attempted to drill an incorrect location. Since the Virtual shoulder area is an anatomical visual aid, it helps the trainees to properly position their instruments relative to the shoulder joint. That the procedure score was higher when the %gaze on the Virtual shoulder was high tends to confirm our assumption that visual assistance carries the risk of skewing the skills transfer to a real environment. Since our dataset concerned the initial (first or second) trainees' training session, we can assume that the visual aid is relevant at this point, but should be removed over the training program.

Similar results were found during laparoscopic cholecystectomy. By dividing the environment into several AoIs such as the screen (endoscope feedback) and the instruments, they found that the more time spent fixing on the more relevant AoI (i.e. the screen), the higher the OSATS [22], indicating better technical skills. Another study showed that more time spent looking at the display led to a better objective prediction of trainees' technical skills [23]. This emphasizes the role of gaze behavior in surgical technical skills with better performance linked to a specific usage of the surgeon's gaze. Our findings align with those results highlighting the importance of gaze behavior in surgical training [17 - 20].

Limitations

This work presents several limitations. First, we did not record nor quantify distractions during the simulation. The simulation protocol and the simulation setup were designed to avoid distraction: the experiments took place in a closed room and only the trainee and the supervising surgeon were in the room during the simulation. Although the supervising surgeon limited its interactions with the trainee as much as possible, we could have added a simple camera to record the interactions, qualify the distractions and remove the corresponding gazes.

There was a relatively high loss of raw video data due to camera freezes. For this proof of concept study, the ET device was chosen for its low price and its eyeglass compatibility and should be upgraded to avoid data loss. A loss of data was also found at the ET level, where the confidence threshold alone removed more than 70% of the gazes for 6 sessions. Although our 0.828 threshold might be too high, it is noteworthy that those 6 trainees wore eyeglasses. Despite the ET device's eyeglass compatibility, it made a glare appear on the trainee's eye, preventing proper pupil detection and resulting in lower confidence. In future acquisitions, special attention should be paid to the light sources.

The %gaze estimation in each AoI presents some limitations. First, this distribution depends on the segmentation network performance. The exploration of predicted segmentation masks shows that the inaccuracies are mainly at the edge of the areas. As the segmented areas are larger than the region of interest for the surgeon (e.g. the central circle for the Arthroscope area), we can assume that the number of impacted gazes is small.

Another limitation is the calibration phase performed at the beginning of each session, allowing the estimation of the gaze position according to pupil position. Indeed, we assumed that a trainee whose %gaze in the Background area is higher than 50% (without calibration defect) could not complete the exercise. Since the cuff repair exercise was generally performed after 20 minutes of recording, the ET device's micro-displacements might have gradually led to a decalibration. One solution would be to validate the calibration before each exercise, or perform a post-hoc calibration by tracking a predetermined object.

Regarding the gaze distribution in the environment, the meaning of the measurement itself could be a limitation. In our case, we considered the %gaze in an area and not gaze fixations. Indeed, the gazes crossing an area, to go from point A to point B, were counted in the percentage. However, we believe that over an exercise of about 15 minutes (~160,500 gazes on average), the amount of time spent crossing a zone without actually fixing on it is negligible.

The major limitation of our work is the number of individuals considered to perform the correlations between gaze position and simulator metrics (N=7). Indeed, although we performed 25 training sessions, we ended up excluding 18 of them to keep only those with good data quality. However, since all correlations make sense from a surgical point of view, this study provides a proof of concept regarding visual behavior of surgical trainees during simulation-based training.

5. Conclusion

Using a neural network, we were able to segment several AoIs specific to the VR simulator. The gaze distribution between those areas differed among trainees. After fitting out data based on their representativeness (considering data loss and calibration defect), we highlighted some correlations between the gaze position during the simulation and the metrics and scores returned by the simulator. Even though these correlations must be qualified due to the small number of individuals considered, it provides a proof of concept regarding visual behavior of surgical trainees during simulation-based training. With a calibration adjustment, this experimental

setup could be used to characterize the use of specific visual cues (such as arrows and colors) provided by the simulator with more accuracy. This would mean identifying to what extent the trainees use these cues, and how their visual behavior evolves without them and/or during the learning process. An analysis of the pupillometry could also provide an additional insight about the participant's workload and add to the potential impact and applicability of this study. Ultimately visual behavior could be used to quantitatively assess surgeons' learning curve and expertise while training on VR simulators, in a way that complements existing metrics.

6. Bibliography

[1] Badash, I., Burtt, K., Solorzano, C. A., & Carey, J. N. (2016). Innovations in surgery simulation: a review of past, current and future techniques. *Annals of translational medicine*, *4*(23).

[2] De Visser, H., Watson, M. O., Salvado, O., & Passenger, J. D. (2011). Progress in virtual reality simulators for surgical training and certification. *Medical journal of Australia*, *194*, S38-S40.

[3] Arjomandi Rad, A., Vardanyan, R., Thavarajasingam, S. G., Zubarevich, A., Van den Eynde, J., Sá, M. P. B., ... & Weymann, A. (2022). Extended, virtual and augmented reality in thoracic surgery: a systematic review. *Interactive CardioVascular and Thoracic Surgery*, *34*(2), 201-211.

[4] Canalichio, K. L., Berrondo, C., & Lendvay, T. S. (2020). Simulation training in urology: state of the art and future directions. *Advances in Medical Education and Practice*, *11*, 391.

[5] Hasan, L. K., Haratian, A., Kim, M., Bolia, I. K., Weber, A. E., & Petrigliano, F. A. (2021). Virtual reality in orthopedic surgery training. *Advances in Medical Education and Practice*, *12*, 1295.

[6] Walbron, P., Common, H., Thomazeau, H., Hosseini, K., Peduzzi, L., Bulaid, Y., & Sirveaux, F. (2020). Virtual reality simulator improves the acquisition of basic arthroscopy skills in first-year orthopedic surgery residents. *Orthopaedics & Traumatology: Surgery & Research*, *106*(4), 717-724.

[7] Satava, R. M. (2008). Historical review of surgical simulation—a personal perspective. *World journal of surgery*, *32*(2), 141-148.

[8] Brunckhorst, O., Khan, M. S., Dasgupta, P., & Ahmed, K. (2015). Effective non-technical skills are imperative to robot-assisted surgery. *BJU international*, *116*(6), 842-844.

[9] Anderson, O., Davis, R., Hanna, G. B., & Vincent, C. A. (2013). Surgical adverse events: a systematic review. *The American Journal of Surgery*, *206*(2), 253-262.

[10] Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Medical education*, *40*(11), 1098-1104.

[11] Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., & Vincent, C. A. (2008). Reliability of a revised NOTECHS scale for use in surgical teams. The American Journal of Surgery, 196(2), 184-190.

[12] Undre, S., Sevdalis, N., Healey, A. N., Darzi, A., & Vincent, C. A. (2007). Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World journal of surgery, 31(7), 1373-1381.

[13] Casy, T., Tronchot, A., Thomazeau, H., Morandi, X., Jannin, P., & Huaulmé, A. (2022). "Stand-up straight!": human pose estimation to evaluate postural skills during orthopedic surgery simulations. *International Journal of Computer Assisted Radiology and Surgery*, 1-10.

[14] Tolvanen, O., Elomaa, A. P., Itkonen, M., Vrzakova, H., Bednarik, R., & Huotarinen, A. (2022). Eye-Tracking Indicators of Workload in Surgery: A Systematic Review. *Journal of Investigative Surgery*, *35*(6), 1340-1349.

[15] Gil, A. M., Birdi, S., Kishibe, T., & Grantcharov, T. P. (2022). Eye Tracking Use in Surgical Research: A Systematic Review. *Journal of Surgical Research*, *279*, 774-787.

[16] Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical teacher*, *40*(1), 62-69.

[17] Wilson, M. R., Vine, S. J., Bright, E., Masters, R. S., Defriend, D., & McGrath, J. S. (2011). Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: a randomized, controlled study. *Surgical endoscopy*, *25*(12), 3731-3739.

[18] Fox, S. E., & Faulkner-Jones, B. E. (2017). Eye-tracking in the study of visual expertise: methodology and approaches in medicine. *Frontline Learning Research*, *5*(3), 29-40.

[19] Law, B., Atkins, M. S., Lomax, A. J., & Wilson, J. G. (2003). Eye trackers in a virtual laparoscopic training environment. In *Medicine Meets Virtual Reality 11* (pp. 184-186). IOS Press.

[20] Cai, B., Xu, N., Duan, S., Yi, J., Bay, B. H., Shen, F., ... & Chen, C. (2022). Eye tracking metrics of orthopedic surgeons with different competency levels who practice simulation-based hip arthroscopic procedures. Heliyon, 8(12), e12335.

[21] Dilley, J., Singh, H., Pratt, P., Omar, I., Darzi, A., & Mayer, E. (2020). Visual behaviour in robotic surgery—Demonstrating the validity of the simulated environment. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *16*(2), e2075.

[22] Evans-Harvey, K., Erridge, S., Karamchandani, U., Abdalla, S., Beatty, J. W., Darzi, A., ... & Sodergren, M. H. (2020). Comparison of surgeon gaze behaviour against objective skill assessment in laparoscopic cholecystectomy-a prospective cohort study. *International Journal of Surgery*, *82*, 149-155.

[23] Snaineh, S. T. A., & Seales, B. (2015). Minimally invasive surgery skills assessment using multiple synchronized sensors. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 314-319). IEEE.

[24] Burkhart, S. S., & Lo, I. K. (2006). Arthroscopic rotator cuff repair. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, *14*(6), 333-346.

[25] Kassner, M., Patera, W. & Bulling, A. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 2014. p. 1151-1160.

[26] Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, *23*(7), 903-921.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In P*roceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).