Providing Automated Real-Time Technical Feedback for Virtual Reality Based Surgical Training: Is the Simpler the Better?

Sudanthi Wijewickrema[1], Xingjun Ma[1], Patorn Piromchai[1,2], Robert Briggs[1], James Bailey[1], Gregor Kennedy[1], and Stephen O'Leary[1]

[1] The University of Melbourne, Australia,
[2] Khon Kaen University, Khon Kaen, Thailand

Abstract. In surgery, where mistakes have the potential for dire consequences, proper training plays a crucial role. Surgical training has traditionally relied upon experienced surgeons mentoring trainees through cadaveric dissection and operating theatre practice. However, with the growing demand for more surgeons and more efficient training programs, it has become necessary to employ supplementary forms of training such as virtual reality simulation. However, the use of such simulations as autonomous training platforms is limited by the extent to which they can provide automated performance feedback. Recent work has focused on overcoming this issue by developing algorithms to provide feedback that emulates the advice of human experts. These algorithms can mainly be categorized into rule-based and machine learning based methods, and they have typically been validated through user studies against controls that received no feedback. To our knowledge, no investigations into the performance of the two types of feedback generation methods in comparison to each other have so far been conducted. To this end, we introduce a rule-based method of providing technical feedback in virtual reality simulation-based temporal bone surgery, implement a machine learning based method that has been proven to outperform other similar methods, and compare their performance in teaching surgical skills in practice through a user study. We show that simpler rule-based methods can be equally or more effective in teaching surgical skills when compared to more complex methods of feedback generation.

Keywords: Virtual Reality Surgery, Automated Real-Time Feedback, Simulation-Based Surgical Education

1 Introduction

As surgeons' skill is often the determining factor between life and death during critical operations, it is important that they are properly trained. Traditionally, surgical education has revolved around the apprenticeship model where experienced surgeons would mentor trainees during cadaveric dissection and later in the operating theatre. However, in the current climate, the sole use of this model

of surgical training is not practical due to issues such as scarcity of cadavers and demand for better trained surgeons in less time.

With the advent of haptic technology, it has become possible to incorporate the sense of touch into virtual reality (VR) applications, thus leading to its wide spread use as a supplemental training technique in surgical education. VR offers a risk-free, repeatable, and accessible platform where surgeons can practice operations. It can be used to develop a standardized curriculum with the possibility of including rare pathologies that are difficult to come by in practice. As such, it is being employed in a wide range of surgical fields.

Although practice on a VR simulator has been proven to improve surgical skills [1, 2], the sole availability of a simulator is not always adequate to offer a meaningful educational experience [3]. As such, understanding how surgical skills are acquired and designing a curriculum based on VR simulators accordingly is essential. One important aspect when designing an effective surgical curriculum is the provision of performance feedback [3, 4]. Feedback is essential for effective skill acquisition, and must be both timely and contextually relevant [5, 6]. Performance feedback in surgical training has typically been provided by human experts. However, if time-poor expert surgeons are solely relied upon to provide feedback during practice, the use of VR as an independent surgical training platform will be limited. Thus, in recent years, research has been conducted on how performance feedback can be automated in VR surgical simulation.

Both real-time and summative feedback are important for skill acquisition [7] and should considered when developing automated feedback systems that emulate human experts. Although there exist numerous examples of automated feedback systems that provide summative feedback at the end of a procedure [8–10], provision of real-time feedback in VR simulation is relatively rare.

Feedback is closely related to assessment [7]. Assessment is typically based on a set of performance goals, and thus, feedback should be provided so that it assists in achieving these goals. In surgery, there are standardized assessment scales that test different aspects of surgical skill [11, 12]. These skills include knowledge of anatomy, use of correct instruments and settings such as magnification, knowledge of procedure (procedural skills), and technical/motor skills.

Performance feedback on the first three types of skills/knowledge is relatively straightforward to provide. For example, trainees can be assisted in locating anatomical structures or landmarks using verbal warnings [13]. Feedback on environmental settings can be provided by comparing against pre-defined value ranges [14]. Procedural guidance has typically been provided visually based on an expert procedure [15–17]. For example, Rhienmora et al. [18] presented a ghost drill that a trainee had to follow in a dental surgery simulation. Step-by-step guidance on how to perform temporal bone surgery was presented through visual cues such as highlighted overlays and instructions in Anderson et al. [15], Wijewickrema et al. [13], and Copson et al. [19].

In contrast, the quality of technical skills in surgery is harder to define and therefore, there has been no consensus on how best to provide feedback on tech-

nical skills. Work on real-time technical feedback can be broadly categorized into two types: rule-based and machine learning based methods.

Rule-based feedback methods typically work with fixed operation rules based on performance metrics identified by experts in the field. Operations that violate these rules are considered poor technique. The range of acceptable values for these performance metrics are either determined by experts or calculated through the analysis of pre-collected expert data. For example Fried et al.[20] introduced measures for evaluating performance in virtual endoscopic sinus surgery (such as violation of tissue, violation of instrument tolerances, force patterns etc.) and developed a database of expert performances, that they used to identify performance that deviates from these value ranges. In temporal bone surgery simulation, Sewell et al. [9] provided feedback on technique based on individual metrics selected by the user such as visibility, force, and region of bone removed.

Machine learning has also been applied to generate technical feedback in VR based training. Feedback generated via machine learning techniques is generally considered to be more flexible than that generated by checking fixed rules [21]. These methods learn characteristics of expert and novice behaviour using pre-collected data. The learned characteristics are then used to identify novice behaviour in real-time and generate feedback to improve technique.

Zhou et al. [22] introduced one such method that extracted expert and novice patterns by training a pattern mining algorithm on expert and novice data. Zhou et al. [23] further demonstrated that random forest prediction models can also be used to generate technical feedback. Cui et al. [24] showed how optimal performance feedback can be generated by transforming the random forest feedback problem to an integer programming problem. To overcome the issue of high processing time of this method, Ma et al. [21] introduced a near-optimal method that was more suitable for real-time feedback generation. Ma et al. [25] also introduced a less memory intensive, yet accurate and efficient method of providing feedback in temporal bone surgery simulation. Here, the adversarial concept [26] was used on a pre-trained neural network to generate minimal changes in behaviour (suggested feedback) that changes novice technique to expert technique.

An important consideration in providing technical feedback is the number of metric changes suggested by the feedback. As it is difficult to change more than one feature at a time in practice, some methods limit the feedback generation to the 'best' feature change [22, 23, 21]. However, there may be instances where changing one metric may not be sufficient to move from novice to expert technique, rendering these methods inaccurate in such instances. Other methods generate the optimal feedback irrespective of the number of metric changes suggested in the feedback [24]. Ma et al. [25] finds a balance between these two extremes, by introducing a penalty that ensures the best performance change (feedback) is returned with the least number of metrics involved in that change.

The feedback provided by these methods have been proven to improve surgical skill acquisition in comparison to the case where no feedback is provided [13]. However, to our knowledge, no investigations have yet been conducted to ascertain whether machine learning based methods of feedback generation are better

in practice than their simpler counterparts: rule-based methods. In this paper, we undertake such an investigation with respect to the provision of technical feedback in VR temporal bone surgery simulation. To this end, we introduce a novel rule-based method for technical feedback generation in an existing VR simulator, implement a validated machine learning based generation technique for the same, integrate them with other types of automated feedback (procedural guidance, proximity warnings, and environmental settings feedback) and evaluate their effectiveness with respect to practical surgical skill acquisition.

2 Simulation Environment

The VR platform used in this research is a temporal bone surgery simulator (see Fig. 1). The virtual temporal bone comprises anatomical structures such as the dura, sigmoid sinus, and facial nerve. A virtual drill reflects the movements of a haptic device that provides tactile feedback to the user. The impression of depth is achieved through NVIDIA 3D vision technology. A MIDI controller is used as a convenient input device to change environment variables such as magnification level and burr size. Using the VR simulator, surgeons can perform middle and inner ear operations to remove diseased tissue and improve hearing.

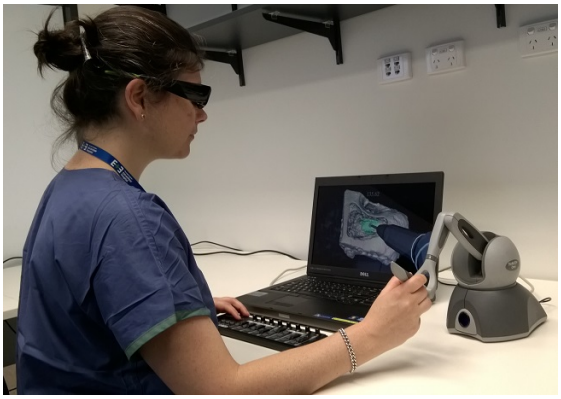

Fig. 1. The University of Melbourne VR temporal bone surgery simulator.

Our VR simulator provides different types of real-time feedback. Procedural guidance is presented using the method of Wijewickrema et al. [13]. Each step of a surgery is highlighted on the temporal bone (see Fig. 1) and the next step is only provided once the current step is completed. Following existing work [27], verbal proximity warnings are provided when drilling within a specified distance of an anatomical structure. Verbal feedback on environment settings such as magnification level and burr size is provided by comparing the current setting with preset ranges of acceptable values [14].

Technical skills, although usually related to instrument handling, may have different meanings in different application contexts. Our focus is on open surgeries such as temporal bone surgery, where technical skill is reliant on how drill 'strokes' are fashioned. A stroke represents a continuous motion of the drill with no abrupt changes in direction. To segment strokes from a surgical trajectory, we use the method introduced by Hall et al. [28]. The quality of a stroke can be determined by analyzing the characteristics, or stroke metrics, that define that stroke (stroke length, duration, speed, acceleration, force, straightness) [21, 25].

Surgical technique varies according to the region being drilled. For example, long strokes and high force can be used when drilling in an open area, but more caution is warranted when drilling near an anatomical structure. Therefore, it is important to identify regions and define the quality of surgical technique for each region separately. We identify the regions of a temporal bone using the method discussed in Wijewickrema et al. [14]. Fig. 2 illustrates these regions.

As a whole, the guidance/feedback provided by the system pertains to how a task is conducted and as such, can be placed in the task level of the feedback framework introduced by Hattie and Timperley [7].

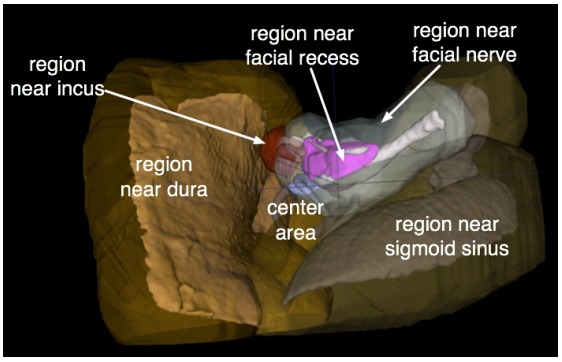

Fig. 2. Regions where surgical technique can be considered uniform.

3 Implementation of Surgical Technique Feedback

3.1 Rule-Based Technical Feedback

To develop our rule-based feedback generation method (RBFB), we first identified a set of rules that have to be satisfied (and their importance) for technique to be deemed acceptable for each region. This was done in consultation with two expert ear surgeons, based on the technical performance metrics (or stroke metrics) discussed in Ma et al. [21, 25]. Table 1 illustrates the rules (in order of priority) along with the suggested action or feedback when a rule is violated.

Table 1. Rules defined for surgical regions in decreasing order of priority.

Region	Priority	Rule	Feedback
Center area	1	Force too low	Use more force
	2	Strokes are too short	Use longer strokes
	3	Drilling too slow	Drill faster
Dura Sigmoid sinus Facial nerve	1	Force too high	Use less force
	2	Drilling too fast	Drill slower
	3	Strokes are too long	Use shorter strokes
	4	Strokes are too short	Use longer strokes
	5	Drilling too slow	Drill faster
	6	Force too low	Use more force
Incus	1	Strokes are too long	Use shorter strokes
	2	Drilling too fast	Drill slower
	3	Force too high	Use less force
	4	Strokes are too short	Use longer strokes
	5	Force too low	Use more force
	6	Drilling too slow	Drill faster
Facial recess	1	Force too high	Use less force
	2	Drilling too fast	Drill slower
	3	Strokes are too long	Use shorter strokes
	4	Force too low	Use more force
	5	Strokes are too short	Use longer strokes
	6	Drilling too slow	Drill faster

In real-time, during a VR surgery, strokes are segmented from the surgical trajectory, and performance metrics calculated for each stroke. Depending on the region that is being drilled, the individual stroke metrics are compared against pre-defined ranges determined from expert data (discussed next) according to the order of importance of the rules. For example, in the center area, the current force is first compared with the minimum recommended force value for the area. If the force is less than the minimum, feedback to increase force is provided. If the force is greater than the minimum, the next rule is checked (that is, if the stroke length is less than the pre-defined minimum).

The acceptable ranges for each stroke metric was calculated offline using 16 surgeries performed by 7 experts. Each expert trajectory was segmented into strokes and divided according to which region they belonged to. Then, for each region, and for each stroke metric, the acceptable range x_{range} was calculated as $x_{range} = [x_{mean} - 2 * x_{std}, x_{mean} + 2 * x_{std}]$, where, x_{mean} and x_{std} are the mean and standard deviation of the pre-collected expert values of a stroke metric for a given region respectively. Note that the full range of expert data was not used as the acceptable range because of the possible existence of outliers.

3.2 Machine Learning Based Technical Feedback

We chose the neural network feedback generation method (NNFB) of Ma et al. [25] over other machine learning based methods due to a few reasons. First, it has been evaluated on pre-collected simulator data in comparison with other existing methods and found to be efficient and accurate. Second, it has an in-built penalty term related to the number of metrics on which to provide feedback, and

as such, returns the best feedback that has the least number of metric changes. Third, its memory footprint is low (especially when compared to random forest based methods), thus making it ideal for implementation in a VR application with substantial memory requirements. Fig. 3 shows how NNFB can be used to provide real-time technical feedback in VR surgical simulation.

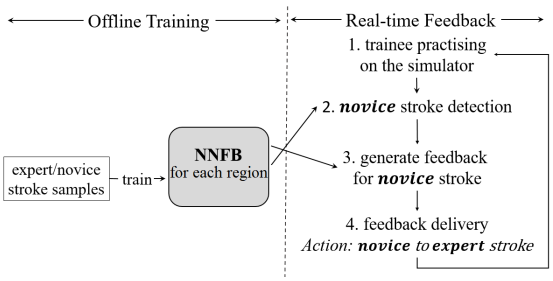

Fig. 3. NNFB feedback generation method.

For the offline training of the neural network classifier, we used a dataset of 16 surgeries recorded by 7 experts (same data used in the RFBF range calculation above) and 34 surgeries from 18 novices. The surgical performances were segmented into strokes with all strokes in expert and novice performances considered to be expert and novice strokes respectively. The strokes were separated according to the region they were performed in and stroke metrics were calculated for each of these. Expert and novice strokes of each region were used to train neural networks with one hidden layer. The number of hidden neurons for each region was chosen using cross validation [25].

In real-time, strokes are segmented from the surgical trajectory, and the neural network classifier for the relevant region is used to identify whether it is an expert or novice stroke. In the case of a novice stroke, the change in metrics that modifies it to an expert stroke that minimizes the distance between the original novice stroke and the resulting expert stroke subject to a constraint (the number of metrics that need to be changed) is calculated [25].

The resulting feedback consists of a change in one stroke metric (for example, 'decrease force') if this can be achieved at a minimum cost. If not, it will comprise changes in multiple features (for example, 'decrease force' and 'decrease speed'). However, as higher numbers of metrics changes are penalized, the number of metric changes in the suggested feedback remains low. This ensures that the feedback is useful in practice where changing multiple aspects of performance at once can be difficult. Performance results of NNFB in comparison with other machine learning based methods can be found in Ma et al. [25].

Although we assume that all expert strokes are indeed of high quality, this may not always be the case. Therefore, to ensure accuracy, exceptional behaviour (or outliers) in data has to be removed prior to training the classifiers in NNFB. We adopted a commonly used outlier removal method: isolation forests [29].

Unlike simpler methods (for example, that used in Section 3.1), isolation forests consider the interaction between the stroke metrics in multi-dimensional space. For example, a high force value may be detected as an outlier when considered by itself. However, when combined with a low speed value, a high force value may be acceptable. By considering all stroke metrics at the same time in relation to each other, outliers can be detected more accurately. Approximately 10% of our original dataset was removed as outliers using an isolation forest. Outlier removal was done separately for experts and novices and for each region and the resulting strokes were used to train the classifiers.

4 Presentation of Feedback

As mentioned above, surgical skills are multi-faceted and feedback on these have to be presented as a whole. Thus, we investigated how different types of feedback (procedural guidance, technical feedback, proximity warnings, and environment settings feedback) can be presented together based on existing work [30, 31] and consultation with surgeons, education psychologists, and computer scientists. Presentation strategies thus implemented in the feedback system are as follows. Note that the same settings were used in presenting both RBFB and NNFB.

Presentation medium: The presentation medium of feedback is important, specifically when there are multiple types of feedback being presented. The effectiveness of a feedback presentation medium depends on factors such as task complexity and skill level of the trainee [32]. Thus, it is not possible to define global rules relating to the effectiveness of a presentation medium. Studies have however, shown that multimodal feedback (such as visual and auditory) can be used effectively in teaching skills [32, 33]. In view of this, we combined the use of verbal auditory and visual feedback following previous work in the field [13, 14, 27]. Technical feedback (along with environmental settings feedback and proximity warnings) was provided in the form of verbal auditory instructions [14, 27]. Pre-recorded audio clips were saved for each type of feedback for each region, to be played when the relevant type of feedback is generated. For example, if a trainee is using too much force around the facial nerve, the presented feedback would be 'use less force near the facial nerve'. Procedural guidance was provided through a visual step-by-step process [13].

Priority of different types of feedback: In our system, different types of feedback are generated in parallel by different algorithms. As such, it is important to determine the order of priority in case two or more types of feedback are generated at the same time. In consultation with expert surgeons, considering the level of skill of the novices being taught, it was decided that proximity warnings were to take first priority as they warn trainees when nearing anatomical structures so that damage could be minimized. The second most important was environment settings feedback, as use of magnification and burr size are critical to a successful surgery. Technical feedback was the last in order of priority. Therefore, for example, if a proximity warning and an instance of technical feedback

(say 'reduce force') are generated at the same time, the proximity warning will be played by the system. As procedural guidance was being provided through a different medium (visual), there were no restrictions on its presentation.

Confidence of presented feedback: The temporal placement of concurrent (real-time) feedback can be immediate or delayed [34]. The choice of temporal placement depends on the task at hand and skill level of the trainee. We used immediate presentation for proximity warnings and environmental settings feedback as they are considered to be critical and immediate action should be taken to avoid damage. Also, these forms of feedback are more straightforward to define than technical feedback and therefore, deviations from the norm can easily be determined with a high level of confidence. In contrast, as surgical technique cannot be as clearly defined, it is prudent to delay the presentation of technical feedback until a high level of confidence in its accuracy is reached. This ensures that feedback is provided only when a certain aspect of surgical technique is consistently poor. We implemented this through a form of buffering. Technical feedback (a suggestion to increase or decrease a certain stroke metric) was added to a buffer every time the system generated one. A feedback item in the buffer could consist of one metric change (RBFB or NNFB) or several metric changes (NNFB). The latest generated feedback was checked against those already in the buffer, and considered to be a candidate for presentation only if it had a significant presence in the buffer. In our application, the buffer size was 10 and the confidence threshold for a feedback to be eligible for presentation was 60%.

Region crossing: As we implemented separate feedback generation models for different regions to account for variations in surgical technique, we also employed strategies to maintain the confidence of feedback when regions are crossed. First, we cleared the feedback buffer when a region is crossed so that the integrity of the region-based feedback is maintained. Second, to account for situations where the boundaries of regions are being drilled, we kept track of the regions of the latest strokes and used this to ensure that the feedback provided is consistent. To this end, we saved the regions of the 100 latest strokes in a region buffer. The generated feedback in a region was considered for presentation only if the percentage of that region in the region buffer was more than a preset threshold. Lower threshold values were used for regions of higher importance so that more feedback can be provided. The threshold values were: center area - 30%, dura and sigmoid sinus - 20%, and incus, facial nerve, and facial recess - 10%.

Frequency of presentation: The guidance hypothesis [35] predicts that the guiding properties of concurrent (or real-time) feedback are beneficial for learning motor skills (such as technical skills in surgery) when used to reduce error, but detrimental when relied upon. It also suggests that a reduced frequency of feedback may facilitate learning. Reducing the frequency of feedback also ensures that trainees do not face cognitive overload. In view of this, we implemented strategies to only present a subset of feedback that pass the conditions discussed above and are deemed eligible for presentation. First, we suppressed any feedback that is generated at the start of a surgery while trainees familiarize themselves

with the procedure. This initial time period was set to 30 seconds for technical and environment settings feedback and 10s for proximity warnings. Second, we suppressed the presentation of feedback generated immediately after another feedback was presented. The period of suppression for technical and environment settings feedback was 7s while for proximity warnings, it was set to 30s. Third, once a feedback has been presented, if the same feedback was generated within a given interval from its last presentation, it was not presented. This time interval was set to 15s for all three types of auditory verbal feedback.

5 Experimental Results

We conducted a randomized trial of medical students to evaluate the effectiveness of the above two feedback methods in improving surgical technique. Ethics for this study was obtained from the Human Ethics Sub-Committee of the University of Anonymous (Ethic ID: 1647227). 26 students participated in this study. In previous studies, significant effects were seen with as little as 20 participants (10 per group) [13], thus ensuring that our study was sufficiently powered. Data from one participant was removed from the study due to an error in data collection. Participants were randomized into two groups: rule-based feedback (RBFB) and neural network based feedback (NNFB) using a block randomization technique.

After the consent procedure, all participants were shown a video tutorial on how to perform a simple ear surgery (i.e. cortical mastoidectomy). Then, they were given 5 minutes to familiarize themselves with the simulator. Next, they were asked to perform the same procedure without any guidance (pre-test), so that their initial skill level could be gauged. Then, they received two training sessions with automated guidance: technical feedback, procedural guidance, proximity warnings, and environment settings feedback as discussed in Section 2. Depending on which group the participant was in, technical feedback was provided using the two techniques RBFB or NNFB. After the completion of the two training sessions, the participants performed a mastoidectomy procedure without any guidance (post-test). Finally, they filled out a questionnaire on their experience of the technical feedback. Fig. 4 illustrates the study design.

Non-parametric tests were used in the statistical analysis of the results due to the non-normal nature of the data and small sample sizes. Kruskal-Wallis and Mann-Whitney U tests were used to compare between groups and to compare pre and post behaviour within groups respectively. The effect sizes were calculated as $r = \sqrt{\frac{\chi^2}{N}}$ and $r = \frac{Z}{\sqrt{N}}$ for the two tests respectively, where χ^2 and Z are the outputs of the tests and N is the total sample size [36].

Comparison of performance: As mentioned in Section 1, competency in surgery is typically evaluated against validated objective assessment scales that consider all aspects of surgical skill (procedural, technical etc.). Here, we used a scale developed specifically for cortical mastoidectomy [12]. It was shown to have high construct validity (accuracy in determining skill level) and inter-rater reliability (agreement between different assessors) and comprised two parts: checklist

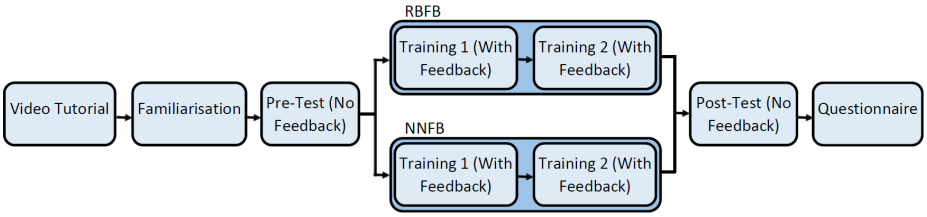

Fig. 4. Design of the user study. Note that all conditions other than the source of the technical feedback were the same for the two groups in the training sessions.

and global scores. Participant performance for the pre- and post-tests was determined by an expert surgeon via anonymized videos using this scale.

The checklist score (CS) of the cortical mastoidectomy assessment scale is based on the procedural steps emphasized in surgical course curricula and dissection manuals, with a total score of 110. The global score (GS) of the assessment tool measures technical as well as procedural competence, and has a total score of 50. The results based on both scores for within group (pre-post analysis) and between group (for performance improvement) are shown in Table 2.

Table 2. Results of the performance analysis.

Within groups	**Pre: Median (IQR)**	**Post: Median (IQR)**	**p**	**r**
CS - RBFB	29 (25 → 34)	46 (39.5 → 48)	**0.0001**	-0.7633
CS - NNFB	29 (26 → 36)	41 (32 → 44.25)	**0.0040**	-0.5757
GS - RBFB	14.5 (10.5 → 22)	30 (27 → 39)	**0.0003**	-0.7239
GS - NNFB	16 (10.75 → 18.25)	28 (21.75 → 30)	**0.0013**	-0.6426
Between groups	**RBFB: Median (IQR)**	**NNFB: Median (IQR)**	**p**	**r**
Improvement in CS	16.5 (9 → 20.5)	10 (0.25 → 15.75)	0.0906	0.3385
Improvement in GS	17 (13.5 → 21.5)	11 (2.75 → 13.75)	**0.0409**	0.4089

Assessment of feedback accuracy: The accuracy of the feedback generation methods was determined by an expert surgeon through the analysis of anonymized videos and calculated as $ACC = \frac{TF-FP-WC}{TF+FN} \times 100\%$, where, FP are false positives (feedback provided when stroke technique is acceptable), WC is wrong content (technique is accurately detected as poor, but the content of the feedback is inaccurate), FN are false negatives (feedback not provided when stroke technique is unacceptable) and TF is total feedback provided [27]. Table 3 shows the results of this analysis.

Perception of participants: A questionnaire based on a 5-point Likert scale was used to gather participants' impressions of the technical feedback. Table 4 shows the results of this qualitative analysis.

Table 3. Results of the expert analysis of accuracy.

	RBFB: Median (IQR)	NNFB: Median (IQR)	p	r
False positives (FP)	2 (0 → 2.75)	0 (0 → 1)	**0.0137**	0.3522
Wrong content (WC)	0 (0 → 0)	0 (0 → 1)	0.1637	0.1990
False negatives (FN)	0 (0 → 1)	0 (0 → 1)	0.3218	0.1416
Total feedback (TF)	17 (13 → 21.75)	15.5 (10 → 19)	0.2134	0.1778
Accuracy (ACC) %	84 (77.98 → 93.21)	87.5 (76.92 → 94.12)	0.7251	0.0502

Table 4. Results of the user perception analysis.

	RBFB: Median (IQR)	NNFB: Median (IQR)	p	r
Q1: Usefulness	4 (4 → 4.5)	5 (4 → 5)	0.1495	0.2882
Q2: Clarity	4 (4 → 5)	4 (4 → 5)	0.1495	0.0244
Q3: Accuracy	4 (4 → 5)	4 (4 → 5)	0.6292	0.0966
Q4: Timeliness	4 (4 → 4.5)	4 (4 → 5)	0.3919	0.1712
Q5: Too much feedback	2 (2 → 3.5)	2 (2 → 3)	0.8131	0.0473
Q6: Too little feedback	2 (2 → 3.5)	3 (2 → 3)	0.6623	0.0874

6 Conclusion

The results of the performance analysis showed that both groups showed significant improvement with respect to the checklist and global scores of the assessment scale from pre- to post-tests. This indicates that the different types of feedback/guidance, along with the presentation strategies used, worked in concert to successfully improve different aspects of surgical skill.

A between-group comparison of the improvement in checklist scores showed no significant difference. This is not surprising as both groups received the same procedural guidance which is the sole purview of the checklist score. However, a significant difference was observed in the improvement in the global scores. This indicates that the technical aspects of the assessment scale was indeed affected by the different feedback generation methods leading us to the conclusion that the simpler RBFB method may be more effective in improving technical skills.

The accuracy of the two types of feedback generation methods, as assessed by an expert surgeon was not significantly different. It was seen that more false positives were observed in the RFBF method. These observations do not however explain the better performance seen in the RBFB group with respect to the global score of the assessment scale. Further analysis, for example, a breakdown of the feedback content provided by the two methods, may be required to uncover the reasons behind the difference in performance. Similarly, the analysis of user perception showed that both groups had equally positive experiences of the feedback, indicating that the system was seen to be usable and useful.

From these observations, it can be concluded that simpler rule-based methods of providing technical feedback in VR temporal bone surgery simulation can be equally or more effective than those that are based on methods that consider more complex interactions between performance metrics. It can also be inferred that a system that provides different types of automated feedback/guidance can be used to effectively train different aspects of surgical skill.

References

1. Francis, H.W., Malik, M.U., Diaz Voss Varela, D.A., Barffour, M.A., Chien, W.W., Carey, J.P., Niparko, J.K., Bhatti, N.I.: Technical skills improve after practice on virtual-reality temporal bone simulator. Laryngoscope **122**(6) (2012) 1385–1391
2. Crochet, P., Aggarwal, R., Dubb, S.S., Ziprin, P., Rajaretnam, N., Grantcharov, T., Ericsson, K.A., Darzi, A.: Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical skills. Annals of surgery **253**(6) (2011) 1216–1222
3. Stefanidis, D.: Optimal acquisition and assessment of proficiency on simulators in surgery. Surgical Clinics of North America **90**(3) (2010) 475–489
4. Stefanidis, D., Heniford, B.T.: The formula for a successful laparoscopic skills curriculum. Archives of Surgery **144**(1) (2009) 77–82
5. Ericsson, K.A.: Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Academic medicine **79**(10) (2004) S70–S81
6. McGaghie, W.C., Issenberg, S.B., Petrusa, E.R., Scalese, R.J.: A critical review of simulation-based medical education research: 2003–2009. Medical education **44**(1) (2010) 50–63
7. Hattie, J., Timperley, H.: The power of feedback. Review of educational research **77**(1) (2007) 81–112
8. Mackel, T., Rosen, J., Pugh, C.: Data mining of the e-pelvis simulator database: a quest for a generalized algorithm for objectively assessing medical skill. Studies in Health Technology and Informatics **119** (2006) 355–360
9. Sewell, C., Morris, D., Blevins, N.H., Dutta, S., Agrawal, S., Barbagli, F., Salisbury, K.: Providing metrics and performance feedback in a surgical simulator. Computer Aided Surgery **13**(2) (2008) 63–81
10. Kerwin, T., Wiet, G., Stredney, D., Shen, H.W.: Automatic scoring of virtual mastoidectomies using expert examples. International journal of computer assisted radiology and surgery **7**(1) (2012) 1–11
11. Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British journal of surgery **84**(2) (1997) 273–278
12. Laeeq, K., Bhatti, N.I., Carey, J.P., Della Santina, C.C., Limb, C.J., Niparko, J.K., Minor, L.B., Francis, H.W.: Pilot testing of an assessment tool for competency in mastoidectomy. Laryngoscope **119**(12) (2009) 2402–2410
13. Wijewickrema, S., Zhou, Y., Bailey, J., Kennedy, G., O'Leary, S.: Provision of automated step-by-step procedural guidance in virtual reality surgery simulation. In: VRST. (2016) 69–72
14. Wijewickrema, S., Ioannou, I., Zhou, Y., Piromchai, P., Bailey, J., Kennedy, G., OLeary, S.: Region-specific automated feedback in temporal bone surgery simulation. In: CBMS. (2015) 310–315
15. Andersen, S.A.W., Foghsgaard, S., Konge, L., Cayé-Thomasen, P., Sørensen, M.S.: The effect of self-directed virtual reality simulation on dissection training performance in mastoidectomy. Laryngoscope **126**(8) (2016) 1883–1888
16. Crossan, A., Brewster, S., Reid, S., Mellor, D.: Multimodal feedback cues to aid veterinary training simulations. In: Haptic Human-Computer Interaction. (2000) 45–49
17. Lamata, P., Gomez, E.J., Bello, F., Kneebone, R.L., Aggarwal, R., Lamata, F.: Conceptual framework for laparoscopic vr simulators. IEEE computer graphics and applications **26**(6) (2006) 69–79

18. Rhienmora, P., Haddawy, P., Suebnukarn, S., Dailey, M.N.: Intelligent dental training simulator with objective skill assessment and feedback. Artificial intelligence in medicine **52**(2) (2011) 115–121
19. Copson, B., Wijewickrema, S., Zhou, Y., Piromchai, P., Briggs, R., Bailey, J., Kennedy, G., O'Leary, S.: Supporting skill acquisition in cochlear implant surgery through virtual reality simulation. Cochlear Implants International **18**(2) (2017) 89–96
20. Fried, M.P., Satava, R., Weghorst, S., Gallagher, A., Sasaki, C., Ross, D., Sinanan, M., Uribe, J., Zeltsan, M., Arora, H., et al.: Identifying and reducing errors with surgical simulation. Quality and safety in health care **13**(suppl 1) (2004) i19–i26
21. Ma, X., Wijewickrema, S., Zhou, Y., Zhou, S., OLeary, S., Bailey, J.: Providing effective real-time feedback in simulation-based surgical training. In: MICCAI. (2017) 566–574
22. Zhou, Y., Bailey, J., Ioannou, I., Wijewickrema, S., O'Leary, S., Kennedy, G.: Pattern-based real-time feedback for a temporal bone simulator. In: VRST. (2013) 7–16
23. Zhou, Y., Bailey, J., Ioannou, I., Wijewickrema, S., Kennedy, G., OLeary, S.: Constructive real time feedback for a temporal bone simulator. In: MICCAI. (2013) 315–322
24. Cui, Z., Chen, W., He, Y., Chen, Y.: Optimal action extraction for random forests and boosted trees. In: KDD (2015). 179–188
25. Ma, X., Bailey, J., Wijewickrema, S., Zhou, S., Mhammedi, Z., Zhou, Y., OLeary, S.: Adversarial generation of real-time feedback with neural networks for simulation-based training. In: IJCAI. (2017) 3763–3769
26. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR. (2015)
27. Wijewickrema, S., Ioannou, I., Zhou, Y., Promchai, P., Bailey, J., Kennedy, G., OLeary, S.: A temporal bone surgery simulator with real-time feedback for surgical training. In: NextMed/MMVR21. (2014) 462–468
28. Hall, R., Rathod, H., Maiorca, M., Ioannou, I., Kazmierczak, E., OLeary, S., Harris, P.: Towards haptic performance analysis using k-metrics. In: HAID. (2008) 50–59
29. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: ICDM. (2008) 413–422
30. Baddeley, A.: Working memory: looking back and looking forward. Nature Reviews. Neuroscience **4**(10) (2003) 829
31. Oviatt, S.: Human-centered design meets cognitive load theory: designing interfaces that help people think. In: ACMMM. (2006) 871–880
32. Sigrist, R., Rauter, G., Riener, R., Wolf, P.: Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. Psychonomic bulletin & review **20**(1) (2013) 21–53
33. Burke, J.L., Prewett, M.S., Gray, A.A., Yang, L., Stilson, F.R., Coovert, M.D., Elliot, L.R., Redden, E.: Comparing the effects of visual-auditory and visual-tactile feedback on user performance: a meta-analysis. In: ICMI. (2006) 108–117
34. Schmidt, R.A., Lee, T.D., et al.: Motor control and learning: A behavioral emphasis. Volume 4. (2005)
35. Schmidt, R.A., Young, D.E., Swinnen, S., Shapiro, D.C.: Summary knowledge of results for skill acquisition: support for the guidance hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition **15**(2) (1989) 352
36. Rosenthal, R., DiMatteo, M.R.: Meta-analysis: Recent developments in quantitative methods for literature reviews. Annual review of psychology **52**(1) (2001) 59–82