Constructive Real Time Feedback for a Temporal Bone Simulator

Yun Zhou[1], James Bailey[1], Ioanna Ioannou[2], Sudanthi Wijewickrema[2], Gregor Kennedy[3], and Stephen O'Leary[2]

[1] Department of Computing and Information Systems, University of Melbourne
[2] Department of Otolaryngology, University of Melbourne
[3] Centre for the Study of Higher Education, University of Melbourne

Abstract. As demands on surgical training efficiency increase, there is a stronger need for computer assisted surgical training systems. The ability to provide automated performance feedback and assessment is a critical aspect of such systems. The development of feedback and assessment models will allow the use of surgical simulators as self-guided training systems that act like expert trainers and guide trainees towards improved performance. This paper presents an approach based on Random Forest models to analyse data recorded during surgery using a virtual reality temporal bone simulator and generate meaningful automated real-time performance feedback. The training dataset consisted of 27 temporal bone simulation runs composed of 16 expert runs provided by 7 different experts and 11 trainee runs provided by 6 trainees. We demonstrate how Random Forest models can be used to predict surgical expertise and deliver feedback that improves trainees' surgical technique. We illustrate the potential of the approach through a feasibility study.

Keywords: real time feedback, surgical simulation, random forest

1 Introduction

Over the past two decades, a variety of virtual reality simulations have been developed for surgical training purposes, using novel techniques such as 3D illusion, haptic feedback and augmented reality. These advanced high fidelity simulations offer many potential benefits for surgical training, but also raise new challenges [7]. One potential benefit is the ability to use surgical simulators as self-guided learning tools, thus reducing the burden of work on surgical trainers. However, to realise this benefit, simulators must possess the ability to provide timely meaningful feedback to trainees, in order to facilitate effective learning through deliberate practice [2].

In minimally invasive surgery (MIS), [3, 9] have applied data mining techniques to evaluate surgical processes or identify surgical gestures. This type of surgery typically requires surgeons to manipulate a set of tools in a prescribed way, and there are identifiable "gestures" associated with correct surgical technique. On the other hand, open surgery such as temporal bone surgery often

utilises a small instrument set of surgical drills and suction devices and there are many ways to achieve a correct outcome. As such, it is difficult to identify specific gestures that represent good surgical technique. Furthermore, the time frame of the analysis is typically longer than in MIS tasks, thus increasing the complexity of identifying underlying motion patterns. Thus, evaluating performance in open surgery simulators is a challenging task.

Most existing work [8, 6] on automated performance evaluation in open surgery simulators is limited to assessment of surgical outcomes. Work on the provision of online feedback is still in its infancy. One such work is [8], where users are provided with an evaluation console allowing review of their performance based on surgical motion metrics. Interactive feedback took the form of coloured voxels indicating whether the correct region was drilled. While this type of feedback provides some guidance to achieve the correct surgical outcome, it provides no assistance in improving surgical technique, which can be equally important.

We introduce a method based on Random Forests (RF) [1] to design and deliver online technique feedback within a temporal bone surgical simulator, which can be generalised to other types of open surgery. First, a RF model is built from drilled region data to predict expertise. During a simulator task, if this model predicts that a user is a trainee, a second model combining RF and nearest neighbour search is used to generate human understandable feedback where necessary. The RF model used to generate such feedback is based on surgical stroke data, such as stroke force and length. We evaluated the RF approach against a baseline and our experimental results suggest that RF is a robust technique suitable for expertise classification and feedback delivery during an ongoing temporal bone surgical simulation.

In summary, the paper makes the following contributions: 1) We present the first virtual surgery system which can provide automatic real time feedback to improve surgical technique. 2) The first use of random forest classifiers in virtual simulations as the basis for providing feedback to users.

The remainder of the paper is organised as follows. We first introduce the simulator dataset that was used to train the RF feedback models. We proceed to explain how the two RF models are used to assess expertise and provide feedback. Finally, we define two evaluation metrics to measure the quality of the feedback and present the results of our experiments.

2 Method

Simulation Metrics. Training data was collected using the University of Melbourne temporal bone simulator [5]. This simulator displays a 3D temporal bone model based on segmented micro-CT data and provides haptic feedback through a Sensable PHANToM Desktop haptic device. The simulator can be used to perform any temporal bone drilling task and it records two kinds of performance measures at a sample rate of approximately 15 Hz: outcome measures and technique measures. Outcome measures consist of a time series of drilled voxel positions. Technique measures include motion-based metrics, simulator parameters

and proximity data as shown in Table 1[4]. A k-cos [4] approach is used to segment drill trajectories into a series of surgical strokes for the calculation of motion-based metrics.

Table 1. Technique measures derived from the Temporal Bone Simulator

Motion-based	**Simulator parameters**	**Proximity**
stroke duration	drill burr size	distance to facial nerve
stroke distance	zoom level	distance to hearing bone
stroke speed		distance to membrane
stroke acceleration		distance to dura
stroke force		distance to sigmoid
stroke straightness		distance to tendon
stroke centroid distance		distance to round window
# bone drilled by stroke		

We collected 16 expert and 11 trainee temporal bone simulation runs. The data was provided by 7 different experts and 6 trainees. The training data was unevenly distributed due to limitations in the availability of trainees, but this does not affect the training of RF models significantly. Each simulator run consisted of three surgical tasks: cortical mastoidectomy, followed by posterior tympanotomy and cochleostomy. Cortical mastoidectomy is the preparatory step of many ear operations while posterior tympanotomy and cochleostomy are parts of cochlear implantation surgery.

Random Forest Based Feedback. The Random Forest [1] algorithm builds a strong classifier out of an ensemble of decision trees. A single decision tree (DT) is not a stable classifier, because a small change in the training set can significantly alter the tree structure. To overcome this drawback, RF creates a set of randomly selected subsets from the training data and each subset is used to build a DT. In each tree, nodes are also split using a random subset of all features (i.e. measures) in the data. Each tree classifies each data point (in our case, as expert or trainee), and RF uses the majority vote as the final prediction.

The first step of our approach was to predict the expertise of the surgeon. This step was vital to provide feedback appropriate to the surgeon's expertise. We built a RF model to carry out this task. From a surgical point of view, drilled bone regions (i.e. voxels) are highly related to surgeon expertise so they are appealing to use as features to train a classifier. However, the bone volume contained a large number of voxels (9090750 to be precise) and not every voxel is related to expertise. Thus, it was necessary to employ a feature selection step: mutual information was used to extract the top 10% of voxels based on their

[4] The stroke force metric refers to the force (in Newtons) being generated by the haptic device motors in response to user interaction.

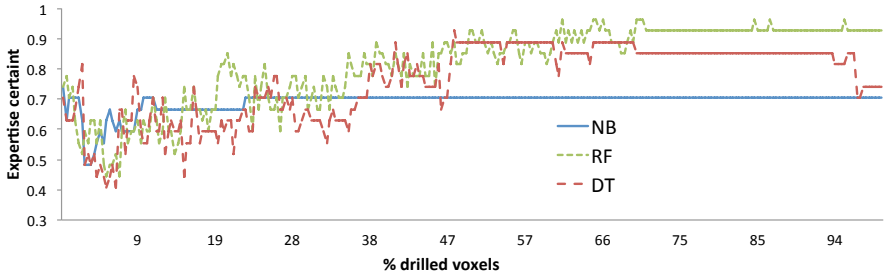

Fig. 1. Expertise classification certainty against surgical task progress. X-axis is the percentage of removed voxels. Y-axis is the expertise classification certainty.

capability to distinguish expertise. The RF tree was built using these voxels as features. We note that previous work by Sewell et al [8] used Naive Bayes(NB) to predict expertise. We chose RF over NB in our method, since the features (voxels) are not independent. If a voxel has been drilled, its neighbours are highly likely to have been drilled as well. Figure 1 illustrates expertise classification certainty using three methods. Prediction models were trained at multiple times during the surgical task, according to when a certain number of voxels were drilled. We expected that the accuracy of prediction models would increase as the number of drilled voxels increased. Figure 1 shows that we can predict the expertise of a user with increasing certainty, by using more information about the voxels drilled so far. By deploying a model that considers the locations of the first 37% (approximately) of voxels drilled, we can be around 80% certain of whether the user is an expert or trainee. In the end, RF misclassified only 1 out of 27 simulation runs. A single DT was generally better than NB, but it is unstable, since we see a large decrease in certainty, even near the end of the task. Overall, we can see that drilled voxel measures provided a very good prediction of expertise. Once we are 80% confident about the trainee's expertise, we can start delivering real time feedback to improve their performance. We note that while drilled voxels provided good expertise classification, they could not provide useful guidance on improving surgical technique. Therefore we used the surgical technique measures shown in Table 1 to create a second RF model for the purposes of feedback generation.

Human trainers often suggest ways to improve surgical technique as part of their feedback. We attempt to mimic this interaction by making suggestions to change one technique feature at a time (e.g. "use longer strokes"). Therefore we begin by selecting the feature on which to provide feedback. A naive way to do this is to select the feature possessing the highest association with expert technique. Work in [1] proposed a way to compute feature importance by randomly shuffling the values of each feature across the dataset. Feature importance is defined as $imp(f) = err(d_f) - err(d)$, where d_f denotes the data with shuffled values for feature f, d is the original dataset, and err denotes the classification

error. We computed $imp(f)$ for each feature and chose the feature of highest importance as our global feature. This feature was used as the baseline (naive) feedback in our feasibility study. This naive approach can provide basic feedback to improve one feature towards expertise, but other features may be just as important at different times during the surgical task. Therefore we propose a dynamic way to deliver feedback using a joint RF model and nearest neighbour approach, outlined in Algorithm 1. The algorithm begins when we identify a

```
   Input: si = new stroke, eg = expert stroke group, dist = distance function
   Output: fn
 1 es = nearest(si, eg, dist);
 2 F = feature vote array;
 3 for each tree in forest do
 4 |   l1=classify si; l2=classify es;
 5 |   if l1 is trainee and l2 is expert then
 6 |   |   f = feature id at which si and es go to different branches;
 7 |   |   if es[f] > si[f] then F[f+]++;
 8 |   |   else F[f−]++;
 9 |   end
10 end
11 fn = maxIndex(F);
```

Algorithm 1: Random Forest feedback algorithm

user as a trainee using the first RF model introduced above. Once the user has performed a stroke, the algorithm identifies the most similar expert stroke (from a historical database) using a nearest neighbour strategy. Instead of using Euclidean distance, we use the distance function derived from RF [1]. The expert stroke serves as a reference for delivering feedback. In order to choose the specific feedback feature, the user stroke and the reference stroke are classified by each tree in the RF feedback model. In a given tree, provided both strokes have been classified correctly, we compute the first feature on which the strokes are split into different branches and this feature receives one vote. Once a feature in a given tree is chosen, we calculate the degree of change (in terms of magnitude and direction) on that feature between the user stroke and the reference expert stroke. As we iterate through the forest, we store the votes for each feature in each direction in an array F. The size of F is twice the number of features, since we count votes for increase (f^+) and decrease (f^-) separately.

Figure 2 shows a running example. This forest contains four decision trees. The dashed cyan line indicates the path of an expert stroke while the solid magenta line is the path of a trainee stroke. Expert leaves are dashed cyan buckets while trainee leaves are magenta solid buckets. In the first tree, the trainee and expert were classified correctly and the split was at the zoom feature. Experts used a lower zoom value, so the vote for zoom$^-$ increased by one. In the second and fourth trees, the split feature suggested a decrease in force, so force$^-$

received two votes. In the third tree both strokes were classified incorrectly, so this tree was ignored. The final feedback chosen would be to "decrease force".

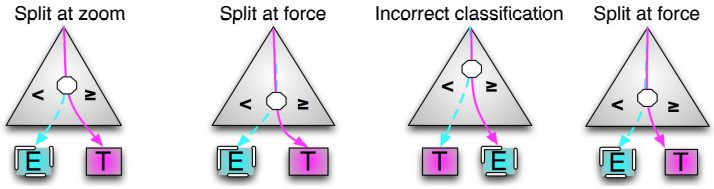

Fig. 2. Example of voting in a forest with four trees (E=expert, T=trainee)

3 Feasibility Study and Results

Comprehensively evaluating the quality of feedback is difficult and would require prospective methods, such as a randomised controlled trial. A controlled trial is beyond the scope of this paper, as the aim is to present the RF-based feedback method and evaluate its practical potential. To evaluate our methods, we conducted a feasibility study based on two metrics we designed to assess feedback quality.

The first metric is *recovery rate*. We generated a "synthetic trainee" as follows: 1) Randomly select a stroke made by an expert. 2) Randomly select a feature from the selected expert stroke. 3) Randomly change the value of this feature. 4) If the altered stroke is classified as trainee by the RF model, go to step 5, otherwise discard the stroke and go to step 2. 5) Label the altered stroke as a "synthetic trainee stroke". Then we input this synthetic trainee stroke to the RF feedback model, and if the suggested feedback corresponded to the modification we made, the number of correct suggestions was increased by one. We repeated this process 10 times on each expert stroke in the dataset and calculated the recovery rate as $\frac{\#\text{ correct suggestions}}{\#\text{ expert strokes}}$.

Table 2 presents an example of recovery rate computation using both the baseline and RF approaches. Suppose we created 5 synthetic trainee strokes by changing the features listed in the first column. We assume that force is the global feature used as the baseline, so the baseline will only make suggestions about force. The only correct suggestion by the baseline is for the first stroke, hence the recovery rate is 1/5. On the other hand, RF feedback makes 4 correct suggestions so the recovery rate is 4/5.

The second method to assess our RF model is to take each trainee stroke, apply the suggested feedback (i.e manipulate the stroke's characteristics according to the feedback), and then determine whether the altered stroke is classified as 'expert' by the RF model . We call this metric *promotion rate* and it is equal to $\frac{\#\text{ trees which classify stroke as expert in RF}}{\#\text{ total trees in RF}}$.

Table 2. Example of recovery rate computation

feature change	baseline	RF feedback
force$^+$	force$^+$	speed$^+$
speed$^-$	force$^-$	speed$^-$
zoom$^-$	force$^+$	zoom$^-$
zoom$^+$	force$^+$	zoom$^+$
stroke length$^+$	force$^-$	stroke length$^+$

To calculate the above measures for our data set, each simulator run was first divided into five stages or sub-tasks. These stages have different surgical goals and characteristics, so we created separate RF models for each stage. We set the number of trees in each RF to 500, which is large enough to tolerate the noise caused by variability in surgical performance. All experiments were conducted using a ten-fold cross validation scheme. In each fold we used 24 runs as the training set and the remaining 3 runs as the test set. For the recovery rate calculation, we changed the value of features by a random percentage ranging from 10% to 50%. For the promotion rate calculation we changed the value of the proposed feature by 10% to 200% in the suggested direction of the feedback. We used a range of value changes to reflect the real life situation, where a trainee is unlikely to be able to achieve the exact suggested correction.

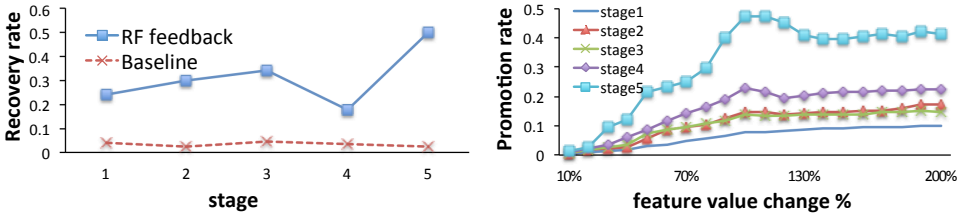

Fig. 3. Average recovery and promotion rate across 5 stages. Curve colour is significant

Figure 3 shows the results. Both rates are expected to increase across the stages, since later stages involved a more restricted surgical work area with less freedom of movement. As shown in Figure 3, the later the stage, the higher the rate for both metrics. This suggests that there were more easily identifiable differences between experts and trainees in later stages. The recovery rate for stage 4 is an outlier and requires further investigation. One possible explanation is that the trainees in our dataset were more skilled in this stage.

RF feedback achieved significantly higher recovery rates than the baseline (using 95% significance t-test). The peak point at stage 5 appears at 100% change, which is consistent with our expectations since it is the exact suggested

correction. RF feedback achieved 50% in stage 5 for both rates. This is a good result because it shows that stroke technique improved considerably by making just one feature correction. However, the promotion rate appears to taper off after 120% change, suggesting that change in a single feature has limited potential to improve overall technique. This is unlikely to be a serious problem, since our approach is not limited to providing feedback on only one feature. In a real simulated training situation, the model could provide a series of suggestions, gradually guiding the trainee towards expertise.

4 Discussion and Conclusion

We have presented a method to automatically deliver online constructive feedback on surgical technique within a temporal bone surgical simulation. This approach is generalisable to other types of open surgery simulation. Our evaluation showed that the RF based approach is effective at classifying expertise and outperformed the baseline in feedback quality measures. The measures of recovery rate and promotion rate demonstrated the feasibility of this approach. Further work including controlled trials is needed to evaluate the feedback system in situ. In addition, future work will also focus on automatic approaches to improve promotion rate. One possible direction might be to investigate correlations between metrics when generating and responding to feedback. In general, there remain intriguing open questions regarding automated feedback in simulation-based training, such as when to provide feedback and how to provide it (such as auditory, visual or haptic).

References

1. Breiman, L., Schapire, E.: Random forests. Mach Learn 45(1), 5–32 (2001)
2. Ericsson, K.A.: Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 79(10), S70–S81 (2004)
3. Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., Jannin, P.: Classification of surgical processes using dynamic time warping. J Biomed Inform 45(2), 255–264 (2012)
4. Hall, R., Rathod, H., Maiorca, M., Ioannou, I., Kazmierczak, E., O'Leary, S., Harris, P.: Towards haptic performance analysis using k-metrics. In: HAID. pp. 50–59 (2008)
5. Kennedy, G., Ioannou, I., Zhou, Y., Bailey, J., O'Leary, S.: Mining interactions in immersive learning environments for real-time student feedback. Australas J Educ Tec 29(2) (2013)
6. Kerwin, T., Wiet, G., Stredney, D., Shen, H.W.: Automatic scoring of virtual mastoidectomies using expert examples. IJCARS 7(1), 1–11 (2012)
7. Kneebone, R.: Simulation in surgical training: educational issues and practical implications. Med Educ 37(3), 267–277 (2003)
8. Sewell, C., Morris, D., Blevins, N., Dutta, S., Agrawal, S., Barbagli, F., Salisbury, K.: Providing metrics and performance feedback in a surgical simulator. Comput Aided Surg 13(2), 63–81 (2008)
9. Stylopoulos, N., Cotin, S., Maithel, S., Ottensmeyer, M., Jackson, P., Bardsley, R., Neumann, P., Rattner, D., Dawson, S.: Computer-enhanced laparoscopic training system (celts): bridging the gap. Surg Endosc 18(5), 782–789 (2004)