Toward Intelligent Tutorial Feedback in Surgical Simulation: Robust Outcome Scoring for Endodontic Surgery

Myat Su Yin, Peter Haddawy
Faculty of ICT,
Mahidol Unviersity
Nakhon Pathom, Thailand
myat.su@student.mahidol.ac.th,
peter.had@mahidol.ac.th

Siriwan Suebnukarn
Faculty of Dentistry,
Thammasat University
Pathum Thani,
Thailand
ssiriwan@tu.ac.th

Phattanapon Rhienmora
School of Science and
Technology,
Bangkok University
Klong Toey, Thailand
phattanapon.r@bu.ac.th

ABSTRACT

Numerous VR simulators have been developed as a means of addressing limitations of the traditional apprenticeship approach to dental surgical skill training. Most existing simulators support intra- and extra-coronal procedures such as carries removal. In this paper we address the problem of automated outcome assessment for endodontic surgery. Outcome assessment is an essential component of any system that provides formative feedback, which requires assessing the outcome, relating it to the procedure, and communicating in a language natural to dental students. This paper takes a first step toward automated generation of such comprehensive feedback. Our system automatically computes reference templates based on tooth anatomy, which provides flexibility to adjust parameters such as tolerance and to create new templates on demand. Detailed scores are transformed into the standard scoring language used by dental schools. Preliminary evaluation of our system on fifteen outcome samples with three expert endodontists shows a high degree of agreement with expert scores.

ACM Classification Keywords

H.5.2. **[Information Interfaces and Presentation (e.g. HCI)]**: User Interfaces (D.2.2, H.1.2, I.3.6); I.6.3 **[Simulation and Modeling (G.3)]**: Applications; K.3.1 **[Computers and Education]**: Computers Uses in Education.

Author Keywords

Surgical simulation; virtual reality; automated outcome assessment; formative assessment; intelligent tutorings.

INTRODUCTION

Acquisition of fine motor skill is essential for dental students. The mainstream approach to dental skill training combines didactic lectures with a surgical master apprenticeship model. Dental schools have increasingly been seeking ways to address a number of limitations of this approach to training, including subjectivity of evaluation, scarcity of available experts, and lack of standardization. Over the past decade, a variety of computer-based simulation systems have been developed as a way to address these limitations and dental schools have begun to incorporate simulators into their curricula. Among the various types of simulators, Virtual Reality (VR) simulators are becoming popular as they have the ability to record kinematic data on how a user performs each step of a task, with numerous dental VR systems developed academically and commercially [2, 3, 5, 6, 7, 8, 10, 11, 14, 15, 16, 17]. However, the full promise of such systems has yet to be realized due to the lack of sufficient support for formative feedback. Without such a mechanism, evaluation still demands dedicated time of experts in scarce supply.

Effective formative feedback requires assessment of outcome, procedure, and the relation of procedure to outcome, coupled with the ability to communicate the assessment in a language natural to dental students. In this paper we take a first step toward such a comprehensive approach to automated formative feedback by presenting the first algorithm for outcome scoring in the challenging area of endodontic surgery. The approach provides scores to the 3D voxel structure commonly used in VR dental simulators at a sufficient level of detail to allow correlation with procedure kinematic variables collected by such simulators. The fine-grained voxel level scores provide the precise error information that can be difficult to quantify in irregular 3D objects such as teeth with complex internal anatomy. To effectively communicate outcome score results, detailed level scores are translated into the language of the coarser level standard scoring system used by dental schools. The algorithm has been implemented for the procedure of access opening to the root canals. Agreement between system scores and those of expert endodontists is evaluated on fifteen outcome samples with a range error types and severity. Results show a high degree of agreement between system scores and those of experts, while at the same time highlighting the variability in the subjective expert judgements.

DENTAL SIMULATOR AND DATA COLLECTION

In this study, we employed the VR dental simulator developed by Rhienmora et al. [9]. The simulator operates on a standard PC with nVidia GeForce 9600GT graphics card and two PHANToM Omni haptic devices in a dual configuration,

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.
 ISBN 978-1-4503-4137-0/16/03...$15.00
DOI: http://dx.doi.org/10.1145/2856767.2856810

Figure 1. Simulator interface

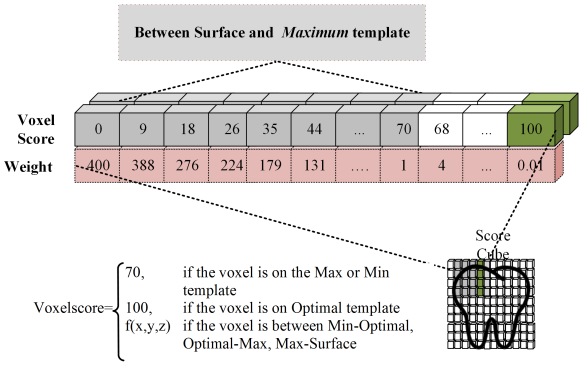


Figure 2. Portion of score cube between surface and Max

representing a hand-piece and a dental mouth mirror (Figure 1). The simulation software was developed in C++, OpenGL, and OpenHaptics SDK (HDAPI). The tooth model was acquired using three-dimensional micro-CT (RmCT, Rigaku Co., Tokyo, Japan) with a resolution of 50 x 50 x 50 μm, tube voltage of 90 KV, and tube current of 150 μA. Tomographic images were obtained using comprehensive dental imaging software (i-VIEW, Morita Co., Tokyo, Japan). Three-dimensional reconstruction was performed using 600 of these two-dimensional images processed by volume rendering. The tooth is stored in the form of a three dimensional grid of voxels representing the density of the structure at each point with a value between 0 and 255.

We have selected access cavity preparation of the root canal treatment procedure to demonstrate and evaluate our approach. This procedure was chosen because it exclusively involves drilling, which is supported by the simulator, and the desired outcome is challenging to score because it is a complex function of the internal tooth anatomy. In this preparation phase, the endodontist drills a small access hole through the surface of the tooth crown to gain access to the pulp chamber and root canals for treatment. The ideal result of access opening preparation is to create an unobstructed passageway to the pulp space and the apical portion of the root canals. The ideal shape of the opening is a function of the tooth shape, tooth size, and the number and location of the root canals. The number and location of the root canals can differ in the same tooth (e.g. mandibular left second molar) across different patients.

METHOD

To provide formative tutoring feedback automated outcome scoring (AOS) must have the ability to identify the type, location and severity of errors and be robust enough to account for a variety of possible outcomes. Since the optimal access opening route is an unobstructed passageway to the pulp space, we first locate the pulp chamber from the training tooth and project vertically from the base of the pulp chamber to capture the morphological information (shape, size and location) of the pulp. Using the projection, the areas on the training tooth are virtually removed to create the optimal outcome template. To permit a clinically acceptable amount of variation in outcome, Max (maximally acceptable) and Min (minimally acceptable) templates are defined by expanding and compressing the optimal template, respectively. Given a tooth and a procedure, a wide range of outcomes is possible. Our approach to outcome scoring is to evaluate the voxels in the tooth volume and label them with scores with respect to reference templates. In AOS, a new tooth volume is created and the voxels tagged with scores according to their locations. We call this a score cube.

As shown in Figure 2, the score cube is a 3D volume of the same size as the training tooth. To fill in the cells of the score cube, we first define a voxel scoring function. In practice, endodontists evaluate each access opening preparation on a scale between 0 and 100, with 0 representing perforation, 1-69 representing unacceptable, and 70-100 representing clinically acceptable. Following this scoring scheme, the score of 100 is assigned to the optimal template area and the score of 70 is assigned to Max and Min templates. The values of the voxels between Min-Optimal, Optimal-Max, Max-Surface are filled using linear interpolation.

Initially, we computed the overall outcome score as the average of the scores of voxels on the drilled area surface. But in a comparison with scoring by endodontists, we found that our scores computed in this way did not correspond well with the expert scores. In subsequent interviews we found that experts assign a significantly higher weight to more severe errors than to minor errors, such that a wall with a small area that is significantly over drilled (close to perforation) is given a much lower score than a wall with numerous small amounts of over drilling, even if the average amount of over drilling in the second case is greater than in the first case. Thus, the linear scheme was adjusted with a non-linear five-parameter logistic weight function [1] to apply higher weight to more severe errors at the voxel level. The parameters were estimated by curve fitting with least squares method prior to the experiments (Figure. 3).

To effectively communicate the assessment results, feedback should be made in a language easily comprehensible to students. Endodontic surgeons evaluate and communicate about errors in terms of scores for the four axial walls (Lingual,

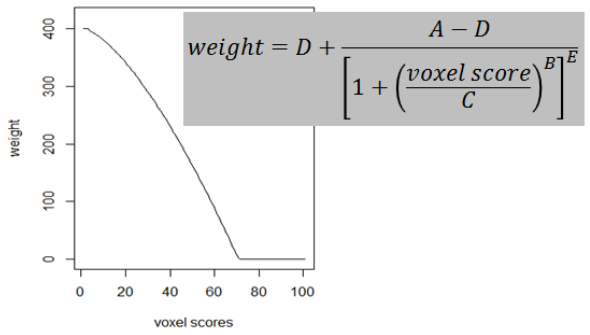


Figure 3. Nonlinear weight function

Buccal, Mesial, Distal), the pulpal floor, and an overall score. AOS thus translates the weighted voxel-level scores into these six scores. This is done by first extracting the surface contour from the drilled outcome area. The contour of the wall is then mapped onto the score cube and the voxels on the surface are assigned the weighted score points accordingly. The wall score is computed as the average weighted scores of the drilled area surface on the tooth slices in the wall region. The overall outcome score is determined in the same manner.

RELATED WORK

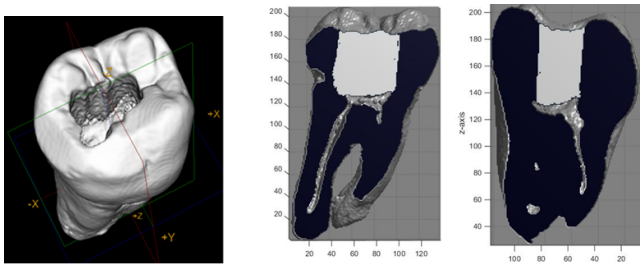
The Dentsim [6] and EPED [3] simulators provide training in intracoronal and extracoronal restoration by using plastic teeth and tracking kinematic data of the instruments using sensors. Training cases include cavity preparation in which the main task is to remove carries while preserving healthy tooth structure. The systems provide evaluation with respect to reference templates in terms of floor depth, outline shape, outline centralization, hand-piece positioning, wall angles, retention, floor smoothness, and wall smoothness. Because plastic teeth do not model internal tooth anatomy, the simulators do not cover endodontic surgery. A number of VR dental simulators have also been built for teaching intracoronal and extracoronal restoration. The common approach to outcome assessment is to analyze the amount tooth mass removed from the virtual tooth. The HapTel [16] and VOXEL-Man [17] simulators provide the percentage of caries removed, the percentage of healthy tissue removed, and injuries (e.g. pulp exposed). Both simulators support only cavity preparation. The MOOG Simodont [8] simulator supports cavity, crown, and bridge preparation but information on outcome assessment is not provided.

The Kobra VR simulator from Forsslund Systems [15] provides training in wisdom tooth extraction. Training cases cover bone removal, separation of crown and roots and positioning of elevator. The analysis report is provided on how much the operator has carved into risk areas by measuring the removed amount of bone, enamel, dentin and pulp on the virtual tooth. None of the above VR simulators supports outcome assessment for endodontic surgery.

Much work on automated outcome scoring has been done in the context of otology surgical skill training simulators. In evaluating mastoidectomy performance, Sewell et al. [12] used a Naïve Bayes classifier based on estimates of the probability that each voxel is removed by an expert and a novice to provide skill level classification of users. Their work was later extended by integrating various process-based features and incorporating metrics for exposure to critical anatomic structure into scoring [13]. Similarly, Kerwin et al. [4] present an approach to automated scoring of virtual mastoidectomy performance on a voxel level. They create a fully partitioned segmented dataset by defining surgically important regions on an iconic temporal bone. Then using earth mover distance, parts of an expert-drilled bone are compared with a student-drilled bone. A decision tree is created using the features derived from these comparisons to determine scores of surgical performance.

EXPERIMENTAL EVALUATION

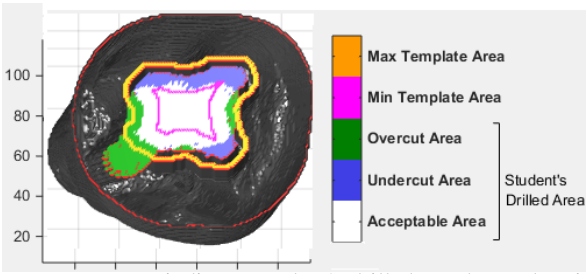
We sought to evaluate the degree of agreement between AOS and human expert scores over a range of outcomes, including varying numbers of errors, types of errors, and severities. Fifteen outcome samples using the mandibular left second molar were prepared by an experienced endodontist who is familiar with the use of the VR system. During data collection, the expert deliberately committed a range of errors on the training tooth to reflect the types of errors committed by the students during the access opening procedure. The set of outcomes contained optimal results and those with errors including perforation of the walls, floor, and both, as well as various combinations of more minor over drilling and under drilling errors. Three endodontists (R1, R2, R3) who had professional training and experience in root canal treatment participated as raters in the experiment. The raters were selected on the basis of expertise levels which varies from one year to more than ten years of experience. The raters received verbal instructions to score the four axial walls (Lingual, Buccal, Mesial, Distal) and the pulp floor using the standard scoring scheme to which they were accustomed. Human raters normally score outcomes using the external view of the tooth, making it difficult to see some errors. Thus any differences in rating between AOS and the human raters could be due to limitations of perception or to differences in judgement. In order to separate these two factors and determine the extent of the influence of perception on scoring, we ran three sets of experiments. In experiment I the raters were provided with a 360 degree external view (Figure 4 (a)) of the drilled tooth



(a) Experiment I **(b) Experiment II**

Figure 4. Example views the experiments I and II

just as they would have in a clinical setting with a real patient. In experiment II the level of information provided to the raters was increased by additionally providing mid cross-sectional views of the drilling area (Figure 4 (b)). This provided raters with visual information on the depth, size and shape of the drilling at the center of the pulp. In experiment III the ideal drilling area based on the internal anatomy of the tooth along with the acceptable drilling areas were provided as visual guidelines for all axial walls and the pulp chamber floor from both lateral and top views (Figure 5) separately.



Note: *Overcut area* indicates student's drilled area beyond optimal template; *Undercut area* indicates student's drilling needs further extension to get the optimal result; *Acceptable area* represents student's drilling within the optimal drilling area.

Figure 5. Min, Max and Optimal templates overlaid on student's drilling area from the top view

In each experiment, raters gave a score on the four axial walls, the pulp floor, and the overall outcome score was computed as the average of the five component scores. Inter-rater agreement for each experiment was determined using the concordance correlation. Concordance correlation tests whether two raters assign the same scores.

RESULTS AND DISCUSSION

Table 1 shows the range of scores, the mean and the median for the three raters and AOS. While the minimum for all is zero, the maximum for R2, R3, and AOS are quite close, while that for R1 is significantly higher (94). The mean and median values for AOS, R2, and R3 are again quite close but those for R1 are lower. To examine the degree of subjectivity in scoring of outcomes, we evaluated the agreement among the three experts.

	AOS	R1	R2	R3
Minimum	0	0	0	0
Maximum	85	94	87	84
Mean	65.6	59.9	66.4	66.6
Median	73.5	64	74	77

Table 1. Overall outcome score ranges in experiment I

Table 2 (columns 1-3) shows the concordance correlation among the three pairings of the raters. Raters R2 and R3 are generally well correlated across the experiments, while rater R1 is mostly poorly correlated with the other two. Weaker correlation between raters R2 and R3 is observed for their pulp floor scores in experiment II. Post experiment interviews indicated that this was likely due to difficulties in perception. They may have relied too heavily on the cross sectional view which, because it was taken mid-volume, revealed only half of the drilling area, thus raters needed to estimate the depth the drilling in the rest of the pulp floor. Since rater R1 disagreed with the consensus score, he was excluded from the analysis of agreement with AOS.

Experiments		R1-R2	R1-R3	R2-R3	AOS-R2	AOS R3
Overall Outcome	I	0.7	0.68	0.99	0.99	0.99
	II	0.36	0.33	0.97	0.99	0.97
	III	0.74	0.62	0.95	0.98	0.95
Axial Walls	I	0.55	0.52	0.8	0.69	0.67
	II	0.6	0.46	0.74	0.75	0.81
	III	0.42	0.27	0.71	0.7	0.9
Floor	I	**0.03**	**0.09**	0.69	0.82	0.86
	II	0.28	0.44	0.43	0.86	**0.28**
	III	**0.13**	0.23	0.88	0.81	0.82

Note: Correlation values are significant for 95% confidence interval except when indicated with bold font face

Table 2. Concordance correlation coefficients between human raters and AOS

Table 2 (columns 4, 5) shows the concordance correlation between AOS scores and those of the two raters R2 and R3. Both types of correlation are very strong for the overall outcome across all three experiments. But we see differences among the experiments moving to the component scores. All AOS scores are substantially correlated (above 0.7) except for the axial walls in experiment I and the floor for AOS-R3 in experiment II. Examining the variation in correlation values between AOS and the two raters from experiment I through experiment III shows no clear pattern of increase in response to increasing amounts of information provided. This suggests that the disagreement in scores is likely due to disagreement in judgement between AOS and the human raters rather than perceptual effects. This is no surprise given the lack of agreement among the human experts.

CONCLUSIONS

Our experimental results show disagreement among human expert scores, reflecting the subjective nature of human outcome scoring. Even in cases where additional perceptual information was provided, the correlations did not uniformly increase, which indicates the disagreement in scores is likely due to disagreement in judgement. Almost all our experiments showed a high degree of correlation between AOS and the two human raters whose scores were themselves well correlated. Given the lack of perfect correlation among the human experts, we would not expect to be able to do much better than this.

Our next steps will include a more thorough evaluation of the algorithm with a larger group of experts, as well as extending the work by identifying the portions of the procedure responsible for each error and combining this with procedure analysis [10] to provide formative feedback.

REFERENCES

1. M Baud. 1993. Data analysis, Mathematical modeling. In *Methods of Immunological Analysis, Vol .1: Fundamentals*, Rene Masseyeff, Winfried Albert, and Norman A. Staines (Eds.). VCH Publishers, Inc., New Yourk, 656–671.
2. Gilad Ben Gal, Ervin I Weiss, Naomi Gafni, and Amitai Ziv. 2011. Preliminary assessment of faculty and student perception of a haptic virtual reality simulator for training dental manual dexterity. *Journal of dental education* 75, 4 (2011), 496–504.
3. CDS-100-EPED Inc. 2015. EPED. (2015). Retrieved October 2, 2015 from http://www.eped.com.tw.
4. Thomas Kerwin, Gregory Wiet, Don Stredney, and Han-Wei Shen. 2011. Automatic scoring of virtual mastoidectomies using expert examples. *International Journal of Computer Assisted Radiology and Surgery* 7, 1 (May 2011), 1–11. DOI: http://dx.doi.org/10.1007/s11548-011-0566-4
5. Min Li and Yun Hui Liu. 2007. Dynamic modeling and experimental validation for interactive endodontic simulation. *IEEE Transactions on Robotics* 23, 3 (2007), 443–458. DOI: http://dx.doi.org/10.1109/TRO.2007.895062
6. DentSim Ltd. 2014. DentSim Labs - For perfect dental preps. (2014). Retrieved September 4, 2014 from http://www.dentsimlab.com/.
7. Min Li and Yun-Hui Liu. 2004. A virtual endodontics testbed for training root canal skills. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*. IEEE, 841–846 Vol.1. DOI: http://dx.doi.org/10.1109/ROBOT.2004.1307254
8. Moog Inc. 2015. MOOG Simodont Dental Trainer. (2015). Retrieved September 7, 2014 from http://www.moog.com.
9. Phattanapon Rhienmora, Kugamoorthy Gajananan, Peter Haddawy, Matthew N. Dailey, and Siriwan Suebnukarn. 2010. Augmented reality haptics system for dental surgical skills training. In *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology - VRST '10*, Vol. 1. ACM Press, New York, NY, 97–98. DOI:http://dx.doi.org/10.1145/1889863.1889883
10. Phattanapon Rhienmora, Peter Haddawy, Siriwan Suebnukarn, and Matthew N. Dailey. 2011. Intelligent dental training simulator with objective skill assessment and feedback. *Artificial Intelligence in Medicine* 52, 2 (Jun 2011), 115–121. DOI: http://dx.doi.org/10.1016/j.artmed.2011.04.003
11. V. A. Sandoval, R. A. Dale, W. D. Hendricson, and J. B. Alexander. 1987. A comparison of four simulation and instructional methods for endodontic review. *Journal of dental education* 51, 9 (Sep 1987), 532–8. http://www.ncbi.nlm.nih.gov/pubmed/2442230
12. Christopher Sewell. 2007. *Automatic Performance Evaluation in Surgical Simulation*. Ph.D. Dissertation. Stanford University.
13. Christopher Sewell, Dan Morris, Nikolas Blevins, Sanjeev Dutta, Sumit Agrawal, Federico Barbagli, and Kenneth Salisbury. 2008. Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery* 13, 2 (Mar 2008), 63–81. DOI: http://dx.doi.org/10.1080/10929080801957712
14. Siriwan Suebnukarn, R Hataidechadusadee, N Suwannasri, N Suprasert, P Rhienmora, and P Haddawy. 2011. Access cavity preparation training using haptic virtual reality and microcomputed tomography tooth models. *International endodontic journal* 44, 11 (Nov 2011), 983–9. DOI: http://dx.doi.org/10.1111/j.1365-2591.2011.01899.x
15. Forsslund Systems. 2014. Forsslund Systems. (2014). Retrieved September 9, 2014 from http://www.forsslundsystems.com/.
16. Brian Tse, William Harwin, Alastair Barrow, Barry Quinn, Jonathan San Diego, and Margaret Cox. 2010. Design and Development of a Haptic Dental Training System - hapTEL. In *Eurohaptics*, Vol. 1. 101–108. DOI: http://dx.doi.org/10.1007/978-3-642-14075-4
17. Voxel-Man. 2014. VOXEL-MAN Dental - Virtual Reality Dental Training Simulator. (2014). Retrieved September 7, 2014 from http://www.voxel-man.com/.