Validating Metrics for a Mastoidectomy Simulator

Christopher SEWELL [a], Dan MORRIS [a], Nikolas H. BLEVINS [b], Sumit AGRAWAL [b], Sanjeev DUTTA [c], Federico BARBAGLI [a], Kenneth SALISBURY [a]
[a] *Department of Computer Science, Stanford University*
[b] *Department of Otolaryngology, Stanford University*
[c] *Department of Surgery, Stanford University*

Abstract. One of the primary barriers to the acceptance of surgical simulators is that most simulators still require a significant amount of an instructing surgeon's time to evaluate and provide feedback to the students using them. Thus, an important area of research in this field is the development of metrics that can enable a simulator to be an essentially self-contained teaching tool, capable of identifying and explaining the user's weaknesses. However, it is essential that these metrics be validated in able to ensure that the evaluations provided by the "virtual instructor" match those that the real instructor would provide were he/she present. We have previously proposed a number of algorithms for providing automated feedback in the context of a mastoidectomy simulator. In this paper, we present the results of a user study in which we attempted to establish construct validity (with inter-rater reliability) for our simulator itself and to validate our metrics. Fifteen subjects (8 experts, 7 novices) were asked to perform two virtual mastoidectomies. Each virtual procedure was recorded, and two experienced instructing surgeons assigned global scores that were correlated with subjects' experience levels. We then validated our metrics by correlating the scores generated by our algorithms with the instructors' global ratings, as well as with metric-specific sub-scores assigned by one of the instructors.

Keywords. Surgical simulation, automatic performance evaluation, metrics, temporal bone, tutoring, mastoidectomy

Introduction

The existing "apprenticeship" model of surgical training relies on real-life patient encounters as the substrate for learning. Inherent to this opportunistic approach is the assumption that enough patient encounters will take place within the set period of time (the "residency") to effect proficiency. Assessment of this cognitive and technical proficiency is based on the subjective impressions of the surgical educators, and is often erroneous [1].

The economics [2], efficiency [3], effectiveness, degree of responsibility and ethics [4] of this traditional approach have come into question in recent years, particularly for physicians in the early stages of training. With recent data on unacceptable rates of medical error nationwide to fuel this criticism, medical interest groups and governing bodies have called for accountability. Finally, educators are concerned over the validity of a training system that relies on chance patient encounters to fulfill learning objectives.

To address these challenges some surgical educators have moved toward enhancing, or perhaps replacing, the apprenticeship model with a competency-based curriculum [5]. Within such a system, proficiency is determined by successive mastery of skills as opposed to a prescribed length of training. Mastery is assessed not only by the subjective assessment of the surgeons that are responsible for training, but also by objective and standardized assessment tools. Furthermore, opportunities are put in place for repetitive practice of the necessary cognitive and technical skills in a non-threatening environment where errors are opportunities for learning rather than precursors to adverse outcomes. Finally, trainees are required to meet a rigorous standard of proficiency before being allowed to enter the workforce.

Recognizing that the current system of training cannot accommodate the above criteria, surgical educators have turned to simulation as a novel approach to instruction. Simulators are devices, often technologically intensive as in the case of virtual reality, that provide an ideal platform for repetitive practice, a key component to building expertise [6]. Tutorials that are developed through established methods of expert knowledge extraction (e.g. cognitive task analysis) can be programmed into simulators to teach the learner the preferred way of performing a procedure that they are then required to replicate. The simulators can be programmed to gradate difficulty of tasks to suit the level of the learner. The learner can "test" a variety of approaches to solving the same problem, promoting reflection and analysis of alternative strategies.

A key aspect of training by repetitive practice is constructive feedback. Without it, trainees are not able to improve upon their performance, and may reinforce substandard techniques. However, real-time or offline expert assessment of trainee performance on simulated tasks can be time consuming and costly. As such, it is crucial that virtual reality simulators automatically generate valid and reliable performance metrics that can be used by the trainee to gauge their progress. This study attempts to establish construct validity and inter-rater reliability for performance metrics generated by a novel mastoidectomy virtual reality simulator.

1. Simulator and Metrics

In close collaboration with an otolaryngologist, we have developed a visuohaptic mastoidectomy simulator [7]. In the simulator, a hybrid data structure is maintained that allows computation of appropriate drill forces using rapid collision-detection in a spatially-discretized volumetric voxel representation while graphically rendering a smooth triangular mesh that is modified in real-time as the voxels are drilled away. Other features include realistic drill sounds, bone dust (which can be removed using a suction controlled by a second haptic device), shadows, detailed anatomical models of surrounding structures and the inner ear, stereo graphics, a tool selection menu, networking for haptic mentoring, and a simulated neurophysiology monitor.

In order to take advantage of the opportunities for automated evaluation and intelligent tutoring made possible by a computer simulator's ability to record and analyze all of a user's actions, we have been particularly interested in developing performance metrics for our simulator. During a run of the simulator, all of the data is logged. A video can then be rendered, and the data can be loaded into a console that provides extensive calculation and visualization of all metrics (see Figure 1). The details of the implementations for each of these metrics are discussed in other papers [8][9][10], and are numbered here as listed in Table 1. The numerical "threshold"

values given in the subsequent descriptions and used in this study are based on our informal adjustments using training data and feedback from surgeons, but are all easily modifiable in the console.

Metric 1 reports the percent of voxels that were removed while maintaining proper visibility, since it is important to keep the drilled bone within the line of sight so as to be able to notice visual cues and avoid underlying vulnerable structures. Metric 2 reports the percent of voxels removed using a 6mm drill burr when more than 75% of experts used a 3mm burr for that voxel, since using a large burr is dangerous near certain structures (while using a small burr in safe areas can prolong the procedure). Metric 3 reports the frequency of drill "jumps": the number of removed voxels per thousand that were more than 1 cm away from the previously removed voxel, since smooth, continuous drill strokes reflect expertise and confidence. Metric 4 reports the percent of voxels removed with the drill and suction more than 2 cm apart, since the suction should be kept near the drill to remove obscuring dust and provide irrigation. Metric 5 reports the percent of voxels removed while the surgical field was obscured by more than 300 bone dust particles, since this can reduce visibility. Each voxel of the bone is associated with a probability that an expert removes it, learned from expert training data, since removing all and only the correct bone is essential for a complete yet safe procedure. Metric 6 reports the mean of this probability for all voxels removed by the user. Metric 7 reports the sum of the number of voxels with expert probability over 0.8 not removed by the user and the number of voxels with expert probability under 0.2 that were removed by the user. Metric 8 reports the percentage of voxels removed while applying a drill force magnitude above 0.2 N (using a sliding-window average over 20 milliseconds), as pushing too hard could result in popping through bone and harming underlying structures. Metrics 9 through 12 report the percentage of voxels within 1 cm of, respectively, the dura, sigmoid, facial nerve, and inner ear, that were removed while applying a drill force above 0.2 N, since it is especially critical to be careful around these. Metric 13 reports the percentage of voxels removed while moving the drill faster than 2 cm/s (using a sliding-window average over 20 milliseconds), since moving too quickly can result in a loss of control. Metrics 14 through 17 report the percentage of voxels within 1 cm of, respectively, the dura, sigmoid, facial nerve, and inner ear, that were removed while moving the drill faster than 2 cm/s. Metric 18 reports the percentage of the facial nerve that has been properly exposed: the bone sufficiently thinned over it so that it can be seen, located, and safely avoided. Metric 19 reports the percentage of the facial nerve that has either been directly exposed or can be inferred from the directly exposed area. Metric 20 reports the percentage of the facial that has been overexposed: too much bone has been removed, allowing it to be contacted and harmed.

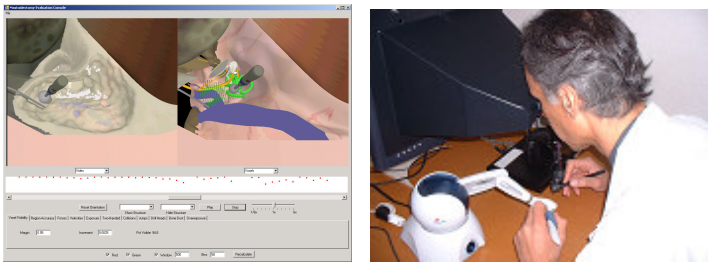

Figure 1. At left, a run of our simulator being replayed in the metrics console. At right, our simulator set-up.

2. Experimental Design

Fifteen right-handed participants were asked to perform a mastoidectomy (removal of a portion of the temporal bone and exposure of relevant anatomy) in our simulator. Participants included four experienced surgeons, four residents in head and neck surgery with surgical experience, and seven novices with no surgical experience.

Participants were presented with a tutorial of the simulator and were given fifteen minutes to practice using the haptic devices and the simulator's user interface. Participants were then presented with an instructional video describing the target procedure, and were given access – before and during the procedure – to still images indicating the desired appearance of the bone model at various stages in the procedure (Figure 2, left). Participants were asked to perform the same procedure twice.

Each participant's hand movements, haptic forces, and surgical interactions were logged to disk, then later rendered to video. Videos were assigned a global score on a scale of 1 to 5 by two experienced head and neck surgery instructors; the instructors were not aware of which videos came from which subjects and viewed them in randomized order. In addition, one of the instructors also assigned sub-scores (also on a scale of 1 to 5) to each of the videos according to several specific criteria directly related to individual metrics included in the simulator.

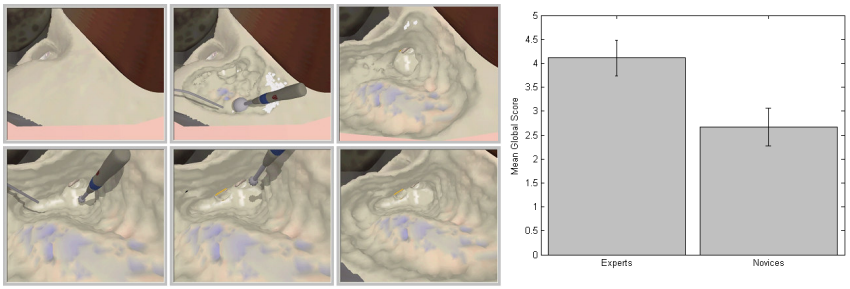

Figure 2. At left, still images presented to experimental participants, indicating the stages of the procedure. At right, expert and novice mean global scores, with 95% confidence interval error bars.

3. Results

The mean of the global scores received by the participants with prior surgical experience was found to be significantly different ($p<0.0001$ using one-tailed t-test) from the mean of the global scores received by the novices, whether considering the scores assigned by either of the instructors separately or considering the average of the two scores for each participant, thus establishing construct validity of our simulator (Figure 2, right). The scores assigned by the two instructors were well correlated ($r=0.718$, $p<0.0001$), demonstrating inter-rater reliability (Figure 3, left). The correlations of each of the metrics with the average of the two global scores assigned by the instructors are shown in Table 1. Table 2 presents the correlations of metrics for which one of the instructors assigned metric-specific sub-scores. A plot of instructor rating versus simulator score is shown for Metric 1 (visibility) in Figure 3, right side.

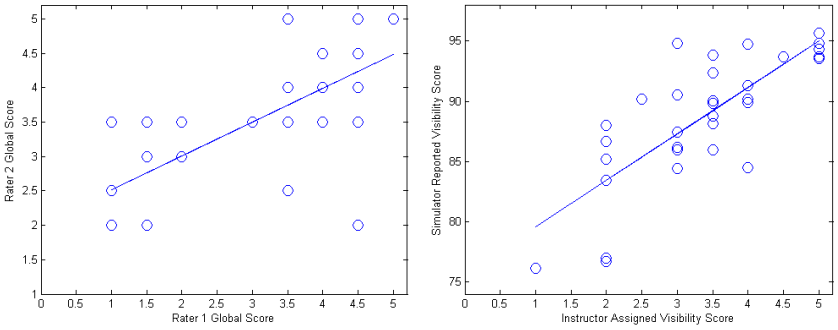

Figure 3. At left, correlation of the two instructors' scores. At right, correlation between instructor and computer assigned visibility scores (Metric 1).

Table 1. Correlations of metrics with the average of the global scores assigned by two instructors.

Metric	r	p	Metric	r	p
Drilling Technique			**Drill Forces**		
Pct Bone Visible at Removal (1)	0.728	<0.001	Pct Excessive Forces (8)	-0.355	0.046
Pct Removed with Burr Too Large (2)	-0.405	0.022	Near Dura (9)	-0.471	0.007
Jump Frequency (3)	-0.226	0.213	Near Sigmoid (10)	-0.420	0.017
Suctioning Technique			Near Facial Nerve (11)	-0.563	<0.001
Pct Excessive Inter-Tool Distance (4)	-0.469	0.007	Near Inner Ear (12)	-0.468	0.007
Pct Excessive Dust (5)	-0.365	0.040	**Drill Velocities**		
Bone Removal			Pct Excessive Vel. (13)	-0.131	0.474
Mean Removal Probability (6)	0.337	0.059	Near Dura (14)	-0.152	0.407
Pct Improbable (Non)Removals (7)	-0.794	<0.001	Near Sigmoid (15)	-0.143	0.434
Facial Nerve Exposure			Near Facial Nerve (16)	-0.387	0.029
Pct Directly Exposed (18)	0.469	0.007	Near Inner Ear (17)	-0.339	0.058
Pct Direct or Indirect Exposed (19)	0.519	0.002			
Pct Overexposed (20)	-0.536	0.002			

4. Discussion

Most of the metrics (1,2,4,5,7,8-12,16,18-20) correlated strongly ($p<0.05$) with assigned global scores. While the avoidance of applying excessively large drill forces (8-12) did strongly correlate with performance, there was little such correlation for overall drill velocities (13), or for velocities when near the dura or sigmoid (14, 15), but there was a moderate correlation for velocities when near the facial nerve and inner ear (16, 17). This may reflect a tendency for skilled participants to always avoid applying large forces while still working quickly and confidently in relatively safe areas and exercising extreme caution in the particularly dangerous regions near the facial nerve and inner ear.

Table 2. Correlations of metrics with global scores and with metric-specific sub-scores.

	Specific Score		Global Score	
Metric	**r**	**p**	**r**	**p**
Pct Bone Visible at Removal (1)	0.777	<0.001	0.728	<0.001
Jump Frequency (3)	-0.355	0.046	-0.226	0.213
Pct Excessive Distance Between Tools (4)	-0.737	<0.001	-0.469	0.007
Pct Excessive Dust (5)	-0.681	<0.001	-0.365	0.040
Mean Removal Probability (6)	0.323	0.072	0.337	0.059
Pct Improbable Choices of Removal Regions (7)	-0.736	<0.001	-0.794	<0.001
Pct of Facial Nerve Directly Exposed (18)	0.400	0.023	0.469	0.007
Pct of F.N. Directly or Indirectly Exposed (19)	0.411	0.019	0.519	0.002
Pct of Facial Nerve Overexposed (20)	-0.500	0.004	-0.536	0.002

Several metrics correlated much more strongly with specific sub-scores than with the global scores. The frequency of drill jumps (3) was found to be closely related to the instructor's assessment of making "purposeful, confident motions", while the distance between instruments while drilling (4) and the percent of time drilling with excessive bone dust in the surgical field (5) closely correlated with the assessment of "two-handed and suctioning technique". However, for most metrics, the correlations with sub-scores were fairly similar to their correlations with the global scores, probably due to the tendency of most participants to score either relatively high on nearly all scores or relatively low on nearly all scores.

References and Acknowledgements

Support was provided by NIH LM07295.

[1] Paisley AM, Baldwin PJ and Paterson-Brown S: Accuracy of medical staff assessment of trainees' operative performance. Med Teach 27:634-8, 2005.
[2] Bridges M and Diamond DL: The financial impact of teaching surgical residents in the operating room. Am J Surg 177:28-32, 1999.
[3] Anastakis DJ, Wanzel KR, Brown MH, et al: Evaluating the effectiveness of a 2-year curriculum in a surgical skills center. Am J Surg 185:378-85, 2003.
[4] Gates EA: New surgical procedures: can our patients benefit while we learn? Am J Obstet Gynecol 176:1293-8; discussion 98-9, 1997.
[5] Sidhu RS, Grober ED, et al: Assessing competency in surgery: where to begin? Surgery 135:6-20, 2004
[6] Ericsson K: The acquisition of expert performance: An introduction to some of the issues, in Ericsson KA (eds): The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games. Mahwah, N.J., Erlbaum, 1996, pp 1 – 50.
[7] Morris D, Sewell C, Barbagli F, Blevins NH, Girod S, Salisbury K: Visuohaptic simulation of bone surgery for training and evaluation. To appear in IEEE Tran. on Comp. Graph. & App, Nov 2006.
[8] Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K: Quantifying risky behavior in surgical simulation. Medicine Meets Virtual Reality, Long Beach, CA, January 2005, IOS Press, pp. 451-457.
[9] Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K: Achieving proper exposure in surgical simulation. Medicine Meets Virtual Reality, Long Beach, CA, January 2006, IOS Press, 497-502.
[10] Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K: Evaluating drilling and suctioning technique in a mastoidectomy simulator. To appear in Medicine Meets Virtual Reality, February 2007.