Computer Aided Surgery
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713723590

Providing metrics and performance feedback in a surgical simulator

Christopher Sewell [a]; Dan Morris [a]; Nikolas H. Blevins [b]; Sanjeev Dutta [c]; Sumit Agrawal [b]; Federico Barbagli [a]; Kenneth Salisbury [ac]
[a] Computer Science, Stanford, California
[b] Otolaryngology, Stanford, California
[c] Surgery, Stanford University, Stanford, California

Online Publication Date: 01 March 2008
To cite this Article: Sewell, Christopher, Morris, Dan, Blevins, Nikolas H., Dutta, Sanjeev, Agrawal, Sumit, Barbagli, Federico and Salisbury, Kenneth (2008) 'Providing metrics and performance feedback in a surgical simulator', Computer Aided Surgery, 13:2, 63 - 81
To link to this article: DOI: 10.1080/10929080801957712
URL: http://dx.doi.org/10.1080/10929080801957712

Computer Aided Surgery, March 2008; 13(2): 63–81

informa
healthcare

BIOMEDICAL PAPER

Providing metrics and performance feedback in a surgical simulator

CHRISTOPHER SEWELL[1], DAN MORRIS[1], NIKOLAS H. BLEVINS[2], SANJEEV DUTTA[3], SUMIT AGRAWAL[2], FEDERICO BARBAGLI[1], & KENNETH SALISBURY[1,3]

Departments of [1]Computer Science, [2]Otolaryngology, and [3]Surgery, Stanford University, Stanford, California

(Received 9 July 2007; accepted 11 December 2007)

Abstract

One of the most important advantages of computer simulators for surgical training is the opportunity they afford for independent learning. However, if the simulator does not provide useful instructional feedback to the user, this advantage is significantly blunted by the need for an instructor to supervise and tutor the trainee while using the simulator. Thus, the incorporation of relevant, intuitive metrics is essential to the development of efficient simulators. Equally as important is the presentation of such metrics to the user in such a way so as to provide constructive feedback that facilitates independent learning and improvement. This paper presents a number of novel metrics for the automated evaluation of surgical technique. The general approach was to take criteria that are intuitive to surgeons and develop ways to quantify them in a simulator. Although many of the concepts behind these metrics have wide application throughout surgery, they have been implemented specifically in the context of a simulation of mastoidectomy. First, the visuohaptic simulator itself is described, followed by the details of a wide variety of metrics designed to assess the user's performance. We present mechanisms for presenting visualizations and other feedback based on these metrics during a virtual procedure. We further describe a novel performance evaluation console that displays metric-based information during an automated debriefing session. Finally, the results of several user studies are reported, providing some preliminary validation of the simulator, the metrics, and the feedback mechanisms. Several machine learning algorithms, including Hidden Markov Models and a Naïve Bayes Classifier, are applied to our simulator data to automatically differentiate users' expertise levels.

Keywords: *Simulation, metrics, feedback, performance evaluation, mastoidectomy, validation*

Key link: *http://jks-folks.stanford.edu/bonesim*

Introduction

One of the most important advantages of computer simulators for surgical training is the opportunity they afford for independent learning. Unlike the anatomy lab or operating room, a simulator allows a student to practice at his/her convenience, regardless of the availability of cadavers or patients. Since all data about the environment and the user's actions may be recorded, a sufficiently realistic computer simulation would provide the opportunity to develop objective metrics that may more fairly evaluate a student's performance in a competency-based curriculum than subjective instructor ratings. However, if the simulator does not provide useful instructional feedback to the user, its educational value is significantly reduced, requiring an instructor to supervise and tutor the trainee while using the simulator. In fact, the continued need for instructor feedback with most existing simulators is often cited as a primary reason for the reluctance of many medical schools to fully embrace simulator technology [1]. Thus, the incorporation of relevant metrics is essential to the development of efficient simulators that provide convenience for trainees while minimizing the costs

Correspondence: Christopher Sewell, 1325 Mills St., Apt. 9, Menlo Park, California 94025, USA. E-mail: csewell@cs.stanford.edu

Parts of this research were presented at the 13th, 14th and 15th Medicine Meets Virtual Reality Conferences (MMVR13, MMVR14 and MMVR15) held in Long Beach, California, in January 2005, January 2006 and February 2007, respectively.

ISSN 1092–9088 print/ISSN 1097–0150 online © 2008 Informa UK Ltd.
DOI: 10.1080/10929080801957712

associated with instructors directly supervising training. Equally important is the presentation of such metrics to the user in such a way so as to provide constructive feedback that facilitates independent learning and improvement. The simulator should serve not only as a ''grader'' but also as an ''intelligent tutor.''

In this paper, we present a number of novel metrics and feedback mechanisms for the automated evaluation of surgical technique. They have been implemented in the context of a simulation of mastoidectomy, a surgical procedure that involves drilling away part of the temporal bone in order to gain access to the inner ear. While the examples given in this paper are based on our mastoidectomy simulator, many of the principles behind the metrics and their presentation to the user should generalize well to many other procedures and simulators.

The remainder of this paper is organized as follows. The next section provides a review of related work on performance evaluation in surgical simulation. This is followed by a brief description of our mastoidectomy simulator. The subsequent section describes a variety of metrics intended to evaluate a user's performance on the simulator, and is followed by a presentation of mechanisms for providing feedback to the user based on these metrics, both interactively in the simulator while performing the virtual procedure, and afterwards in an automated ''debriefing'' session using a novel performance evaluation console that highlights problem areas and details how to correct them. The penultimate section reports the results of several studies that were conducted to provide some preliminary validation of the simulator, the metrics, and the feedback mechanisms, making use of both user studies and machine learning algorithms. Finally, conclusions and future work are suggested.

Related work

The economics [2], efficiency [3], reliability [4], effectiveness, degree of responsibility, and ethics [5] of the traditional ''apprenticeship'' model of surgical training, in which assessment of proficiency is based on the subjective impressions of the surgical educators, have come into question in recent years, particularly for physicians in the early stages of training. To address these challenges, some surgical educators have moved toward enhancing, or perhaps replacing, the apprenticeship model with a competency-based curriculum [6]. OSATS (Objective Structured Assessment of Technical Skill, University of Toronto, Toronto, Canada) has gained popularity as a means to move towards objective assessment, and Reznick et al. demonstrated its construct validity [7]. Within a competency-based 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. Trainees may be required to meet a rigorous standard of proficiency before being allowed to enter the workforce. Thus, there has been an increasing interest in incorporating objective metrics into surgical simulators.

Several basic metrics, such as the number of collisions between a user's tool and simulated anatomy, task completion time, and efficiency of movement, are reported by several existing endoscopy simulators, including MIST-VR (Minimally Invasive Surgery Trainer – Virtual Reality, Mentice Medical Simulation, Gothenburg, Sweden), MISTELS (McGill Inanimate System for Training and Evaluation of Laparoscopic Skills, McGill University, Montreal, Canada), ES3 (Endoscopic Sinus Surgery Simulator, Lockheed Martin, Bethesda, MD), and the Upper GI Endoscopy Simulator (5DT Inc., Santa Clara, CA). Recently, the ETH-Zurich Hysteroscopy Simulator has incorporated metrics for percentage of surface area visualized, amount of distension fluid used, task completion time, number of wall collisions, and path length [8]. Several researchers have attempted to define a standard set of metrics for laparoscopic skill trainers. Cotin et al. detailed five critical kinematic parameters: time to completion, path length, motion smoothness, depth perception, and response orientation [9]. A few simulators for non-endoscopic procedures, including CathSim AccuTouch (Immersion Medical, Gaithersburg, MD) and the E-Pelvis (Stanford University, Stanford, CA), also provide some metrics.

Several researchers have attempted to classify users of simulators or instrumented surgical robots as ''expert'' or ''novice'' by applying machine learning algorithms to data recorded during their performance of certain procedures. Rosen et al. affixed force and torque sensors to instruments used in cholecystectomies and Nissen fundoplications on porcine models [10]. Based on the continuous data stream consisting of three-dimensional (3D) forces, rotational torques, and grip force, a vector quantization algorithm was used to group data into clusters, each with one of 14 pre-defined force/torque profiles indicative of a particular action state. Markov Models were developed independently for a group

of novices and a group of experts. A new Markov Model was developed for each user and compared to the novice model and to the expert model according to a simple measure of statistical similarity. Of 24 procedures performed by four subjects, 87.5% were correctly classified. Dosis et al. took data using ICSAD (Imperial College Surgical Assessment Device, Imperial College of Medicine, London, UK) during a laparoscopic suturing task, and directly from the da Vinci robotic surgery system (da Vinci Surgical System, Intuitive Surgical, Sunnyvale, CA) during robotic suturing and rope passing tasks. They defined 18 states based on characteristic XY-plane rotations, elevations, and grip states, and attempted to recover the hidden states of a Hidden Markov Model (HMM) using Baum-Welch optimization. The recovered states were compared to the sequence of states identified by expert surgeons watching video of the procedures. Results were good for the laparoscopic suturing task, but unsatisfactory for the robotic suturing and rope passing tasks [11]. Mackel et al. constructed expert and novice instances of a 32-state HMM, with each state corresponding to one of the 25 possible combinations of currently activated pressure sensors in the E-Pelvis, and used them to successfully classify 92% of 82 subjects [12]. Murphy et al. at Johns Hopkins collected detailed motion data from the da Vinci system and achieved successful classification rates of approximately 85% using HMMs and linear discriminant analysis [13]. They also decoded the HMMs to recognize individual motion states.

Nevertheless, all of these metrics give an incomplete assessment of user performance. Some, such as task completion time, have been explicitly demonstrated to be a poor measure of skill [14]. Most assume a simple global optimum value, such as a minimal number of wall collisions, a minimal path length, or a minimal completion time. They do not consider quantities (such as forces and velocities) whose ideal values may vary in relation to changes in conditions such as tool proximities to anatomic structures, and do not analyze expert performance to learn the nature of such dependencies. Many important elements of good surgical technique have not yet been explored, including proper exposure and identification of anatomic structures, maintenance of proper visibility of the surgical field, and proper drilling and suctioning technique.

Perhaps most importantly, while some work has explored the use of metrics for quantitative evaluation, there has thus far been little focus on the development of mechanisms for providing constructive feedback that may lead to improved performance based on such metrics. For example, determining that a Hidden Markov Model classifies a user's performance as ''novice'' or that the user's path efficiency score is a 75.6 may well have significant value for grading and certifying potential surgeons. However, it tells the user very little about how to improve his/her performance.

The learning theory community has demonstrated that allowing a person to review a video of his/her performance is insufficient in itself to facilitate learning, but that video feedback with cuing (directing attention to the most relevant aspects of the video) can be very valuable [15,16]. Several studies, including those by Feygin et al. [17], Yang et al. [18], Morris et al. [19], and Kahol et al. [20], have explored the use of real-time visual feedback (usually in the form of 1D or 2D graphs comparing the current value of some parameter such as force magnitude to a specified ideal value) and/or haptic feedback (usually directly applying the force to be learned to the user) for teaching generic gestures. This has recently begun to be applied to teaching surgical gestures [21]. The ETH-Zurich Hysteroscopy Simulator has recently added the ability to present the user with a report after completion of the procedure that highlights surface patches not visualized during the procedure in red, and displays a line tracing the path of the endoscope that is colored red when in collision with the uterine wall and green when a safe distance from the wall [8]. Silverstein et al. generated arc graphs that visualized average pressures applied by experts and by novices to each of the pressure sensors in the E-Pelvis [22]. The positions of the arcs corresponded to the physical locations of the corresponding sensors, and bands within the arcs used color brightness to show pressure variation over time, in order to facilitate understanding of the difference in the performance of the two groups.

The simulator

Mastoidectomy was chosen as a test-bed for our metrics development due to the suitability of the dynamic range of forces and the size of the surgical field to the capabilities of existing haptic hardware; the need for computer simulation to teach haptic, anatomic, and cognitive aspects of this surgery which cannot be taught by drilling cadaver bones (the current modality of pre-operative instruction for otology residents); the fairly high risk of morbidity due to the proximity of nerves and blood vessels; and the available opportunities for collaboration with surgeons and residents in this field.

Several other groups have also worked on the simulation of temporal bone surgery. Agus et al. [23] have developed an analytical model of bone erosion as a function of applied drilling force and rotational velocity, which they have verified with experimental data [24]. Pflesser et al. [25] and Petersik et al. [26] modeled a drilling instrument as a point cloud, and used a modified version of the Voxmap-Pointshell algorithm [27] to sample the surface of the drill and generate appropriate forces at each sampled point. This work has also been incorporated into a commercial simulator (Voxel-Man TempoSurg, Spiggle & Theis, GmbH, Overath, Germany). Bryan et al. have also developed a simulator for temporal bone surgery [28], and Stredney et al. integrated a simple tutor into the system, helping the user to identify relevant anatomy [29]. Each of these projects has incorporated haptic feedback into volumetric simulation environments that make use of CT and MR data and use volume-rendering techniques for graphical display.

We have presented the details of our mastoidectomy simulator in reference [30]. 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. The voxel representation may be generated either from surface meshes drawn by hand in a modeling program such as Maya, or from bitmap stacks produced in a program such as Amira from CT DICOM data.

A new triangular mesh is then generated over this voxel mesh, with each voxel a potential vertex. Initially, the isosurface consists of triangles joining all sets of three mutually adjacent voxels on the surface of the bone. When a voxel is removed, all triangles containing the vertex at that voxel are removed, and all of its neighbors are checked to see if any have now become surface voxels; if so, new triangles are created with that voxel and each pair of its neighbors (that are also neighbors of each other) on the surface. The partial transparency of the bone is modeled by shading voxels near underlying structures (using the structure's color or texture and with intensity inversely proportional to distance) when the ray from the current viewpoint through the voxel intersects the structure.

The primary component of the haptic feedback is computed by first discretizing the drill burr into a voxel grid (at a finer resolution than the bone grid). A preprocessing step computes an occupancy map for the tool's voxel array. At each interactive timestep, each of the volume samples in the burr is checked for intersection with the bone volume (a constant-time, integer-based operation using a hash table). A sample point that is found to lie inside a bone voxel generates a unit-length contribution to the overall haptic force vector that tends to push this sample point toward the tool center, which – with adequate stiffness – is always outside the bone volume. The overall force generated by our approach is thus oriented along a vector that is the sum of the ''contributions'' from individual volume sample points. The magnitude of this force increases with the number of sample points found to be immersed in the bone volume. Additional algorithms modify this force by accounting for other effects, including a multi-gain function in which the magnitude of haptic feedback is a nonlinear function of the number of immersed sample points (to increase stiffness while maintaining stability upon initial contact), removal rates that vary depending on drill ''latitude'' (since the drill spins faster around the equator than at the poles), vibrations that vary based on bone thickness (based on data recorded using accelerometers in the cadaver lab), and tangential forces (modeling variations depending on the direction of spin of the burr's flumes).

Realistic drill sounds, based on data recorded while drilling cadaver temporal bones, are produced, with frequencies giving cues to bone depth. Other features include particle simulations of bone dust, blood, and irrigation (each of which can be removed using a suction controlled by a second haptic device); shadows; detailed anatomical models of surrounding structures and of the inner ear; stereo graphics using a Cyberscope (Simsalabim Systems, Berkeley, CA) split-screen mirror system; bimanual collocated haptic devices (one usually used for the drill and the other for the suction); and an interface for switching between tools and maneuvering the view point. The simulator is networked, allowing a user at one computer to observe another user on a different computer, to feel the forces being applied in the simulator on the remote computer, or to collaboratively drill along with the other user. At the bottom of the screen is a neurophysiology monitor, providing realistic feedback regarding nerve response.

Metrics

This section briefly describes the metrics that were developed in order to evaluate user performance on our simulator. The general approach was to take criteria that are intuitive to surgeons and develop ways to quantify them. The metrics are numbered as

listed in Tables III and IV. The numerical ''threshold'' values given in the subsequent descriptions and used in the validation study are based on our informal adjustments using training data and feedback from surgeons, but are all easily modifiable within our performance evaluation console. The details of the implementations are discussed more fully in references [31–33].

Visibility

One of the most important ways in which risk is minimized in temporal bone surgery is by taking care to only remove bone that is within the line of sight, using a ''saucerizing'' drilling technique (removing bone so as to create a saucer-shaped cavity on the bone surface). This enables the surgeon to avoid vulnerable structures just below the bone surface, using subtle visual cues that indicate their locations. If instead some bone is removed by ''undercutting'' (drilling beneath a shelf of bone that obscures visibility), these cues may not be seen, increasing the risk of damaging critical underlying structures (Figure 1A). Thus, Metric 1 reports the percentage of voxels that were removed while maintaining proper visibility. To determine whether a voxel being removed is within the current field of view, a line is simply traced from the voxel to the viewpoint. Points at discrete intervals along this line are tested for intersection with any obstructing object: the voxel mesh (excluding voxels covered by the drill burr) by indexing into the voxel mesh hash table, soft-tissue anatomy surface meshes by traversing an axis-aligned bounding box hierarchy, and (optionally) any accumulated bone dust by directly indexing into the particle grid.

Drilling and suctioning technique

Good technique in performing a mastoidectomy also involves proper handling and coordination of the surgical instruments. During most of a procedure, the surgeon will hold a drill fitted with one of several burr types in one hand and a suction device in the other. An expert will tend to make smooth, purposeful movements with the drill, drill primarily with the side of the burr rather than the tip, and select burr types and sizes appropriate for use in different regions of the bone. The suction device is used for removing dust and fluid from the surgical field to maintain visibility of the bone surface, and may also provide irrigation to cool the bone surface as it interacts with the rapidly moving drill burr. An expert will tend to keep the field relatively free of debris and will keep the suction device close to the drill as bone is removed.

Metric 2 reports the percentage of voxels removed using a 6-mm drill burr when more than 75% of experts used a 3-mm 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 mean deviation from 90° of the angle between the generated force vector and the primary drill axis,

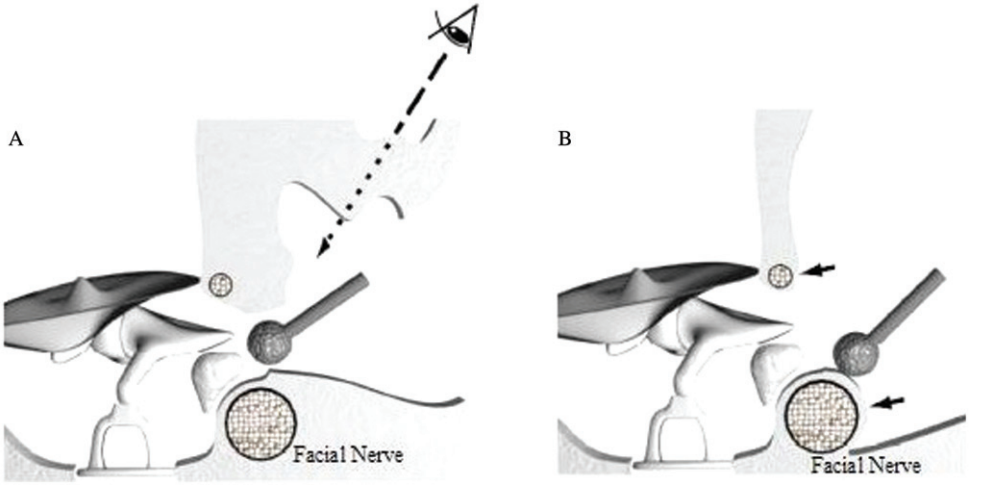

Figure 1. A. The surgeon is not maintaining proper visibility of the drilling region because it is obstructed by a shelf of bone, and will be unable to see that he/she is nearing the facial nerve. B. Correct exposure of the facial nerve has been achieved, as all but a thin layer of bone has been removed above it.

since drilling with the side of the burr is more efficient and reliable than using the tip. Metric 5 reports the percentage 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 6 reports the percentage of voxels removed while the surgical field was obscured by more than 300 bone dust particles, since this can reduce visibility.

Bone removal

One of the most obvious criteria for the evaluation of a mastoidectomy is whether correct decisions were made as to which regions of bone to remove and which to leave intact. A simple method for evaluating this is to have an instructing surgeon label the regions that should and should not be removed, or to automatically label the voxels drilled away by the instructor, and then compare the set of voxels removed by the trainee to this model. However, there is not necessarily a single correct technique; different experts may make somewhat different choices as to which bone to remove, and a given expert may vary somewhat between runs. In addition, not all regions are of equal importance; in some regions, it does not matter much exactly what is removed, while the choices may be much more critical in other areas, especially near nerves and other critical structures.

Thus, a Naïve Bayes classifier (with Laplace smoothing) was constructed to calculate the maximum likelihood estimates for the probabilities that each voxel is removed by an expert and by a novice, based on training data recording during virtual mastoidectomies performed by users of known skill level. Similar to how many spam classifiers use this algorithm based on the assumption that words in an e-mail are chosen for inclusion based on different distributions by spammers and non-spammers, this metric is based on the assumption that voxels are chosen from the bone voxel mesh ''dictionary'' for removal according to different distributions by experts and novices. This classifier can then be used to determine the probabilities that a given mastoidectomy was performed by an expert or by a novice. Since the bone mesh is so large, and so many voxels are unlikely to be very informative (i.e., they will almost always be removed or not be removed, regardless of the subject's expertise), we calculated the mutual information (equivalent to a Kullback-Leibler divergence) for each voxel and built the classifier using only the 1000 most informative voxels in order to optimize the speed and accuracy of the classifier. Metric 7 reports the mean of the expert removal probabilities for all voxels removed by the user. Metric 8 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.

Exposure

Another essential component of good surgical technique is achieving proper exposure of critical anatomic structures so that their shapes, which may vary somewhat among patients, can be confidently established and avoided. In the context of a mastoidectomy, achieving proper exposure involves drilling until only a thin layer of bone remains over vulnerable structures (such as the facial nerve, sigmoid sinus, and dura). When the layer of bone is sufficiently thin, the structure can be seen, due to the partial transparency of the bone. However, the bone must not be completely removed, as this would result in severe trauma to these vulnerable structures. Although it is usually not necessary to directly expose an entire structure, many structures, such as nerves and blood vessels, twist and turn in unpredictable ways, so it is imperative to expose enough of a structure (and in the right places) such that its entire shape (within the surgical field) can be confidently inferred (Figure 1B). Therefore, Metric 9 reports the percentage of the facial nerve that has been properly exposed, Metric 10 the percentage that has either been directly exposed or can be inferred from the directly exposed area, and Metric 11 the percentage that has been overexposed. Direct exposure and overexposure may be computed for each vertex of a surface mesh representing the structure, similar to how visibility is determined for removed bone voxels. If no bone voxels are intersected between the vertex and the viewpoint, the point is overexposed; if some voxels are intersected within a small threshold distance from the vertex but none beyond this distance, it is properly exposed. These calculations must be repeated whenever the viewpoint is moved, and generally only the upward-facing surface of the structure need be considered. Inferred exposure is calculated by taking each directly exposed vertex as a source and propagating a front from it (viewing the surface mesh as a graph) similar to Dijkstra's Algorithm, adding vertices to the inferred list until the distance (along the surface) from the source to the candidate vertex is too long or the curvature too great.

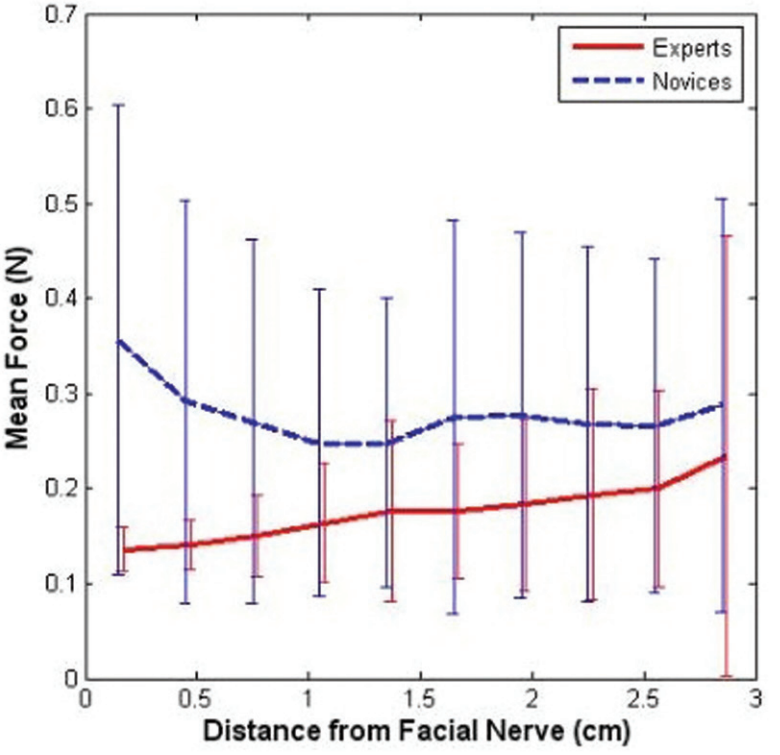

Figure 2. Force magnitudes applied by experts and novices as functions of distance from the facial nerve. Data points were sorted into bins of width 0.2 cm based on distance from the structure. The means of each bin are plotted, along with error bars showing standard deviations within the bins. The data was collected in the user study described in the *Validation* section. Experts applied smaller forces than the novices and decreased their forces as they approached the facial nerve. [Color version available online.]

Forces and velocities

A hallmark of safe drilling technique is applying appropriate forces and operating the drill at appropriate velocities. The acceptable range of forces and velocities is closely related to the drill's distance from vulnerable structures. As a good surgeon gets closer to certain vulnerable anatomic structures, such as the sigmoid sinus, the dura, the facial nerve, or the inner ear, he/she drills more carefully. This increased caution may be reflected in such parameters as decreased drill forces and decreased drill velocities. Changes in these quantities as a function of distance from these structures indicate that the surgeon recognizes the landmarks indicating that he/she is nearing a vulnerable structure, and is responding appropriately (Figure 2).

Metric 12 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 13 through 16 report the percentage of voxels within 1 cm of, respectively, the dura, sigmoid sinus, 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 17 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 18 through 21 report the percentage of voxels within 1 cm of, respectively, the dura, sigmoid sinus, facial nerve, and inner ear that were removed while moving the drill faster than 2 cm/s.

Feedback mechanisms

The utility of all the metrics described in the preceding sections is maximized if they are visualized for the user in a format that clearly highlights the user's strengths and weaknesses and draws attention to problem areas. Therefore, we have developed a number of mechanisms for providing informative feedback to the user based on these metrics. The metrics console described in this section is illustrated in Figure 3.

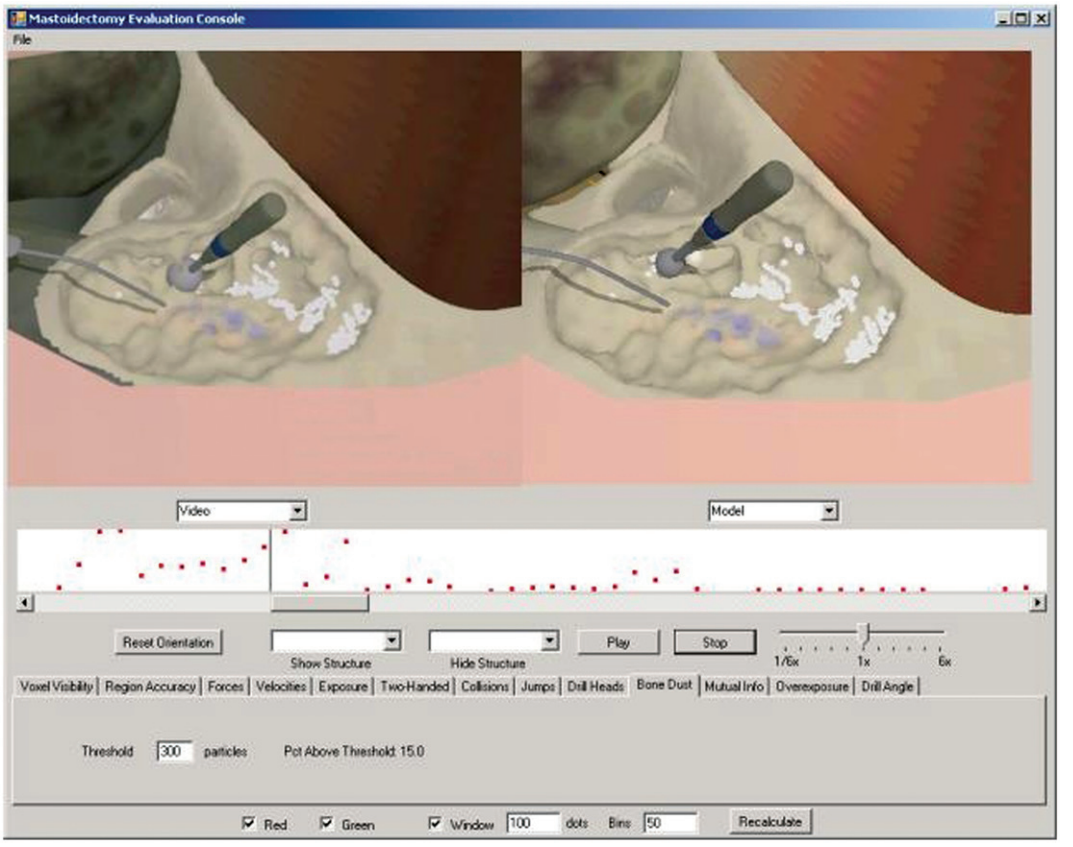

Figure 3. An overview of the metrics console. A video is replaying in the left panel in sync with a model reconstruction of the virtual environment in the right panel. The currently selected metric is amount of accumulated bone dust. An interval of relatively high accumulation is being viewed, as shown by the frame position scroll bar located below a peak in the graph of the time variation of the currently selected metric (red dots). [Color version available online.]

Data logging and video rendering

When a user performs a virtual mastoidectomy on our simulator, all of the data relevant to the user's actions is recorded to a file. Samples are taken at millisecond intervals and logged to disk using a blocked linked list data structure that prevents race conditions between the haptics thread that is producing data and a separate, higher latency disk I/O thread that consumes the data [34]. This data includes the current time, the position and orientation of both tools and the viewpoint, and indices indicating the identity of the currently selected tool in each hand. If a drill is currently selected, the force vector being applied to the drill, the velocity vector of the drill, the power setting of the drill, and the currently selected drill burr type are also recorded. If a suction is currently selected, its power setting is recorded. Additional samples are logged each time a voxel of bone is removed by the drill, recording, in addition to the previously described fields, the index of the removed voxel and its surface normal.

A data file can be loaded and replayed directly in the simulator. In this mode, ''user input'', specifically the positions and orientations of the tools and button states, is read from the file rather than from the haptic devices, while everything else, such as collision detection, computation of bone removal, and re-triangulation of the bone isosurface, is computed just as in the interactive simulation. During this replay, the images from the viewport may be written to an AVI movie file using public domain C++ AVI utilities by Lucian Wischik [35].

Reconstructing the model

After a virtual mastoidectomy has been performed and recorded on the simulator and rendered to video, it may be loaded into the metrics console for detailed analysis and visualization of metrics. When a simulation data file is selected for loading into the console, several different types of files are opened. In addition to the simulator data file and the associated AVI video file, the same models used to

Table I. Time taken by different tasks when opening the performance evaluation console. The times given are averages based on 32 data files collected in the user study described in the *Validation* section.

	Time		Time
Actual procedure	844	Exposure	8.34
Total startup time	38.5	Direct and overexposure	7.89
Load anatomy	15.2	Inferred	0.446
Read data file	6.01	Visibility	0.845
Compute metrics	17.2	All other metrics	0.0604
Mesh collisions	4.46	Building metric graphs	0.0211
Bone dust	2.79	Reset simulation	0.0772

represent the bone and surrounding anatomy during simulation are read in by the metrics console. This allows all of the anatomy present in the simulator to be reconstructed visually in the metrics console. Both the simulator and console make use of the CHAI haptic libraries [36].

As the procedure is reviewed (see *Replaying the procedure* below), the user can rotate, translate, and zoom his/her view of the reconstructed anatomy, as well as select any structure or combination of structures to ''cut away'', providing informative and customizable views that were not seen during the interactive procedure and are not available with only a recorded video. A button is also provided to easily reset the view to align with that used in the original procedure.

Pre-computing metrics

After a simulation data file has been loaded, the complete procedure is immediately replayed once internally while computing all of the metrics. No rendering is done during this process, so the speed is not limited by graphics I/O, and, since no output is provided to the user during this pre-computation step, it can be ''replayed'' in much less time than the actual procedure took. Furthermore, since all the metrics use only the voxel representation of the bone, the costly re-triangulations of the isosurface used for graphical rendering that are normally performed whenever a voxel of bone is removed need not be computed during this phase.

The time required for the pre-computation of metrics is typically 15 to 20 seconds on our dual-processor 3-GHz machine. The exposure metrics are the most costly, as shown in Table I. All results are stored, so the metrics do not need to be recalculated during the user's subsequent interactive review of the procedure, allowing for faster refresh rates.

Replaying the procedure

Once all of the data has been loaded and the metrics computed, the user may replay the recorded procedure in the console. There are two display panels adjacent to one another, and the user may select to watch either the video or reconstructed model in either. The latter type of view may be rotated, translated, and zoomed using the mouse and may show either the unaltered models or be enhanced with metric-specific visualizations as described in the next section. Both views play in sync with one another. In addition to real-time playback, features such as fast-forwarding, rewinding, slow motion, and scrubbing are available using a frame-selection scroll bar below the display panels and a playback-rate slider control.

The AVI video file, generated as described in the *Data logging and video rendering* section above, is shown using a QuickTime ActiveX control embedded in the Microsoft.net form. This control allows for scrubbing (i.e., immediate frame updates as the frame-selection scroll bar is dragged), which facilitates fine control over playback rate and quick overview of the procedure.

The limiting factor in updating the model reconstruction during scrubbing or fast-forwarding is the time required to update the isosurface as bone is removed. However, our program takes advantage of the fact that this triangulation is a ''state function'' in that it depends only on which voxels are currently active and not on the order of removal of previously active voxels. Therefore, voxels are ''queued for removal'' as they are drilled away, but the costly re-triangulation of the mesh is only performed once per frame update, regardless of how many voxels were removed since the previous frame. This prevents many triangles involving voxels that are neighbors of a removed voxel, but are themselves later removed, from ever being generated. When fast-forwarding, re-triangulating only once each frame eliminates even more intermediate

triangulations. Furthermore, normals need only be recomputed for surface voxels in the vicinity of removed bone once per frame.

Each recorded data packet contains the actual elapsed time during the procedure at which it was recorded. When playing at normal speed, during each iteration of the rendering loop, all data packets since the last render are processed (without updating the isosurface) up until the first packet with time greater than the current replay time. The current replay time is maintained by storing the simulation time when play was most recently resumed and adding this to the elapsed time since then using a high-precision timer. (This elapsed time may be scaled to allow variable playback rates.) At the end of this iteration, any necessary updates are then made to the isosurface. Rewinding is accomplished by first resetting the triangulation and all other states (such as bone dust accumulations) to their initial states (which are saved after the initial triangulation and other set-up is completed when loading the files at the beginning of the session) and then fast-forwarding to the desired frame.

This process is generally too slow to keep up with arbitrary scrubbing, so, when the user starts to scrub, the model reconstruction is frozen and only the video updates with the scroll bar. Once the scroll bar is released and normal play resumed, the model reconstruction fast-forwards or rewinds to the new frame and begins playing in sync with the video once again. Resetting (as required for rewinding) takes roughly one-tenth of a second, while the time required for fast-forwarding depends on how much bone is removed and may take up to several seconds. As an example, on our dual-processor 3-GHz machine, fast-forwarding to the end of a full procedure of 526 seconds, involving the removal of 18,408 voxels, takes approximately 4.5 seconds. Of this time, approximately 1.0 second is spent generating the new isosurface triangulation, 0.4 seconds computing new normals, 1.5 seconds updating the bone dust simulation, 1.0 second reshading the bone based on the new viewpoint, and 0.4 seconds processing all the recorded data. We experimented with saving triangulation ''key-frames'', consisting of triangulation data at regular time intervals to reduce the amount of re-triangulation needed when rewinding or fast-forwarding, but found this to be too memory-intensive to be worthwhile on our system.

Visualizing the metrics

One of the most important features of the metrics console that allows it to be used as an active aid to improving performance rather than simply as an assessment of skill level is its ability to help the user localize and identify exactly when and where problems occurred. By making it easy to find trouble spots – times and places where performance was rated poorly – the console provides an efficient mechanism for the user to watch what he/she did wrong and to see how to improve.

When the metrics are initially computed (see *Pre-computing metrics* above), in addition to determining the final overall scores for each metric, scores are also computed and recorded over each of N time intervals, where N is a user-specified number of bins. These sub-scores are plotted on a graph with the x-axis corresponding to time and the y-axis to the value of the sub-scores. The graph is positioned just below the replay scroll bar, with the time scale of the x-axis corresponding to that of the scroll bar. Therefore, the user can quickly scrub to a time of particular interest, such as an interval with a particularly low or high score, and watch the video and/or explore the reconstructed model at that time.

Many of the metrics compute whether a certain condition held at the time of removal of each drilled voxel, and report the percent of such voxel removals for which the condition held. Thus, in addition to breaking down the scores for these metrics based on time intervals, the scores can also be localized in space by maintaining with each voxel whether the condition held when it was removed.

This information may then be presented to the user in the form of colored dots appearing as bone is drilled away (Figure 4). In addition to the video and reconstructed model options for the two display panels, the user may select a ''colored voxel'' view that is equivalent to the reconstructed model view except that, instead of displaying the bone isosurface, dots appear whenever a voxel of bone is removed. All of the other anatomy is present (unless additional cut-away view options are selected), so the proximity of these voxels to key structures (such as nerves and blood vessels) may be easily seen. Typically, the user would view a video in one panel and the colored voxel view in the other panel, watching the dots appear in the latter view as the drill advances in the former view. The dots are colored according to performance at the time that voxel was removed, according to the currently selected metric. In most cases, one color (such as green) is shown for voxels removed with correct technique and another color (such as red) for those drilled improperly. For example, for the visibility metric, voxels removed while maintaining proper visibility may be shown in green while those removed while visibility was blocked by a shelf of bone may be shown in red. Some metrics, such as

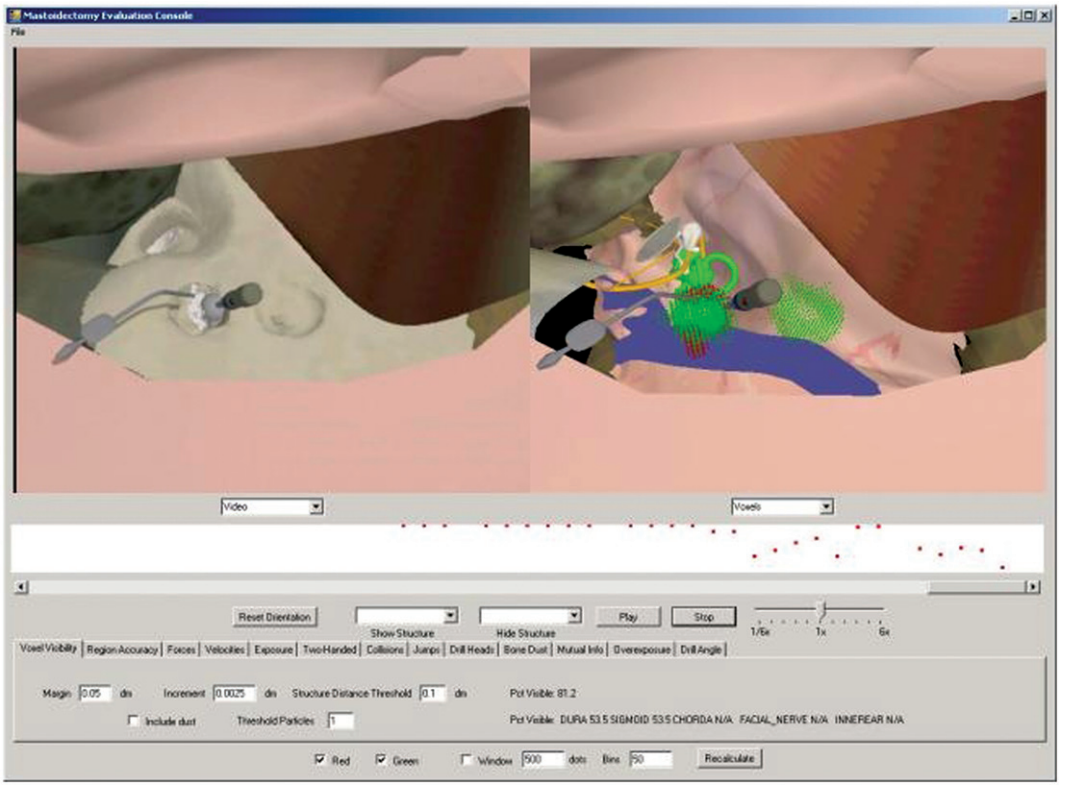

Figure 4. An overview of the console, with the visibility metric selected. It is clear that the user used proper saucerizing technique while making the hole on the right but is now undercutting while drilling the hole on the left. Red dots in the visually annotated reconstructed model view show where the mistake occurred, and the metric-versus-time graph below the display panels shows when it occurred. [Color version available online.]

the expert removal probabilities, have a continuous rather than binary value per voxel, which may be expressed by shading the dots with varying color intensities.

An additional option allows the user to set a ''window size'' W while replaying the simulation in order to display only the W most recently removed voxels, rather than all voxels removed up to that point, so that the voxels of current interest are not obscured by previously removed ones.

Additional visualizations are available to elucidate information about collisions and structure exposure and overexposure. Whenever a collision between a drill head and a critical soft-tissue structure occurs during playback, a sphere appears at the intersection point with radius proportional to impact force. Structures may be shaded in one color for regions that have met the direct exposure criteria, in another color for regions that may be inferred from the directly exposed regions, and in yet another color for regions that have been overexposed. Just as for the other metrics, the percentage exposure (direct, inferred, or over) of critical structures (e.g., facial nerve, sigmoid sinus, chorda tympani) and the number of collisions may also be plotted as a function of time on the graph above the scroll bar to allow for rapid location of problem areas.

At the bottom of the screen is a tab control that allows the user to select individual metrics for evaluation. Within each tab page there are a number of edit boxes and other controls that allow the user to customize the parameters of all of the metrics. For example, the user may specify how thick a layer of bone may remain over a structure for it still to be considered exposed, how far apart the suction and drill may be before it is considered poor technique, or how large a force magnitude is considered safe. Some metric parameters may also be loaded from files, such as the expert removal probabilities for voxels and maximum allowable force magnitudes at any number of different distances from any of the structures (in effect defining a step function between distance and maximum allowable magnitude for each specified structure). The final metric values for the complete procedure are also displayed on the tab pages of their respective metrics. In addition to these overall scores, many metrics also report sub-scores for voxels within a threshold distance (also modifiable on the tab page) of each of several key structures.

If the user changes one or more parameters after having loaded the simulation data file, he/she may press a ''Recompute'' button to reset the console and pre-compute all the metrics again using the new parameters.

Interactive feedback while performing the procedure

In addition to their use in the performance evaluation console, many of these visualizations are also available in the simulator itself while performing the virtual mastoidectomy (Figure 5). Edit boxes, check boxes, and drop-down menus in the simulator's user interface provide the ability to display the colored voxels according to several customizable metrics (such as visibility, forces, and removal region) while drilling.

As the user performs a virtual mastoidectomy, interactive feedback may be provided in the form of colored dots. Bone that has been removed while maintaining proper technique according to the currently selected metric is colored green, while improperly removed bone is shown in red. The user may specify how many of the most recently removed bone voxels to show, set the parameters of the evaluation metrics, toggle rendering of the non-removed bone on and off, set the current metric, and toggle the colored dots on and off. A bar is also available at the edge of the screen to show the relative proportion of correctly and incorrectly removed bone.

Validation

It is essential that a simulator itself, the metrics used in it, and the mechanisms for providing feedback have all been validated in order to ensure that the time spent using it is beneficial and that evaluations provided by the ''virtual instructor'' match those that the real instructor would provide were he/she present. Therefore, in this section, we present the results of several studies that attempt to establish some preliminary validation for each of these aspects of our system.

Validating the simulator

In our first study, 15 right-handed participants were asked to perform a mastoidectomy in our simulator. Participants included four experienced surgeons,

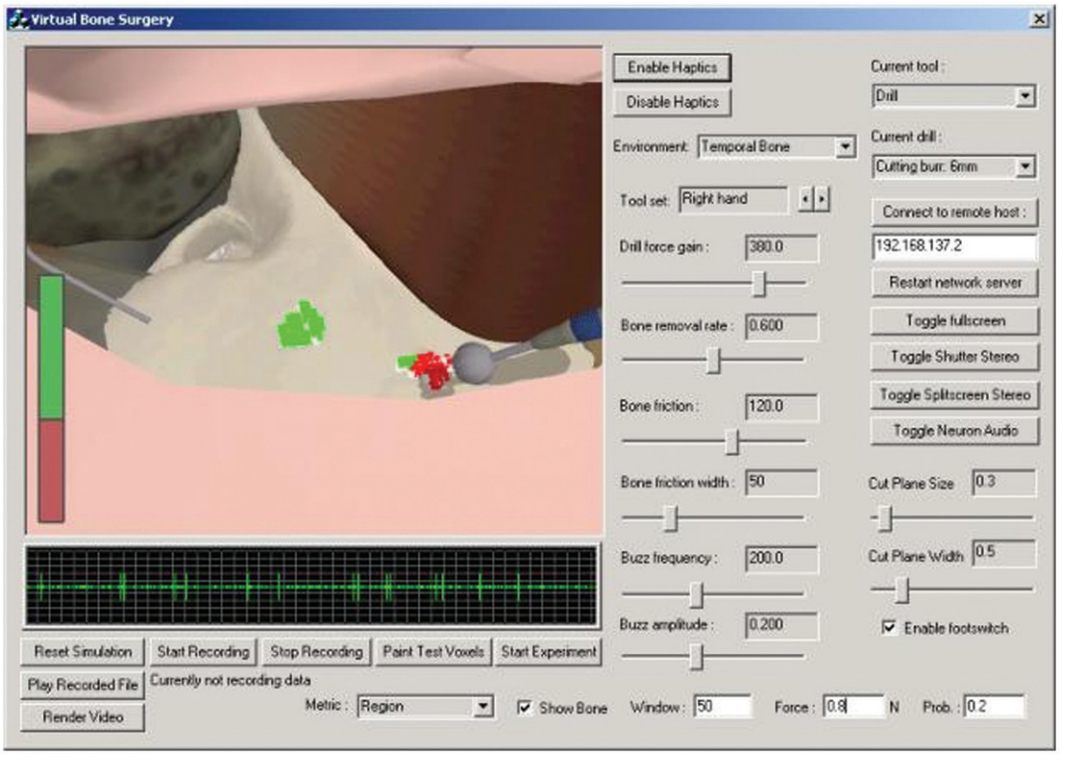

Figure 5. An example of interactive feedback in the simulator, with the bone removal metric selected. Bone removed by the user at the far right that was removed by less than 20% of the experts (as parameterized in the ''Prob'' text box in the bottom right corner) is shown in red, while bone in the center that was removed by a higher percentage of the experts is shown in green. The bar at left shows the relative proportion of correctly (green) and incorrectly (red) removed bone. [Color version available online.]

four residents in head and neck surgery with surgical experience, and seven novices with no surgical experience. They were presented with a tutorial of the simulator and were given 15 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. Participants were asked to perform the same procedure twice.

Each participant's hand movements, haptic forces, and surgical interactions were logged to disk, and then later rendered to video. Videos were assigned a global score on a scale of 1 to 5 by two experienced ear surgery instructors; the instructors were not aware of which videos came from which subjects and viewed them in randomized order.

Results from instructor-assigned scores. The mean of the global scores received by the participants with prior surgical experience was found to be significantly different (4.1 to 2.7, $p < 0.0001$ using one-tailed t-test) from the mean of the global scores received by the novices, thus establishing discriminative validity of our simulator. The scores assigned by the two instructors were well correlated ($r = 0.718$, $p < 0.0001$), demonstrating inter-rater reliability.

Results from classification algorithms. The field of machine learning has developed a wide array of algorithms used for pattern recognition in fields such as speech recognition and handwriting recognition. Such algorithms may also be used to recognize the ''pattern'' of expert behavior, as distinct from novice behavior, in a surgical simulator, which can provide further evidence of the discriminative validity of the simulator.

A Hidden Markov Model (HMM) is one such algorithm. A sequence of observed values is viewed as having been generated by some process that can transition amongst a set of states, each state having its own probability distribution for generating output values. It makes the simplifying assumption, called the Markov assumption, that the output value probability distribution depends only on the current state and not on any of the previous states or output values, and that the state transition probabilities at any point also depend only on the current state.

The model's design is defined by the number of states, and it is parameterized by the transition probabilities between each pair of states and the output distribution for each state. If the output values are continuous, and if each state may be assumed to output values according to a Gaussian distribution, then a mean μ and standard deviation σ may be associated with each state. If there are N multiple, synchronized streams of output data, each observed output is an N-dimensional vector. Thus, a multivariate Gaussian model may be used, each state being characterized by an N-dimensional vector μ, with each element giving the mean for one of the streams, and an $N \times N$ covariance matrix Σ.

Using the data from the previously described user study, the eight highest-scoring procedures were deemed ''expert'' and the eight lowest-scoring as ''novice''. All of these highest-scoring procedures were in fact performed by participants with surgical experience in mastoidectomy, and the lowest-scoring by those without such experience. Using leave-one-out cross-validation, one data set was held out, and separate novice and expert multivariate Gaussian-output HMMs were trained using the Baum-Welch Algorithm [37] (making use of Kevin Murphy's Matlab HMM toolbox [38]). The probabilities of the held-out data set with respect to each of the two HMMs were then calculated using the Forward Algorithm [39], and the data set was classified with the model yielding the higher probability. This was repeated 16 times for each HMM architecture and feature set, holding out one of the data sets each time. The entire procedure was repeated for 10 trials, as the learning process can sometimes be sensitive to the random values used for the initial parameter guesses.

HMM architectures with between 3 and 10 states were tested. Various combinations of raw data

Table II. Average proportion correctly classified using HMMs with different numbers of states and different sets of features.

	3	4	5	6	7	8
Force	0.681	0.700	0.713	0.744	0.763	0.775
Force, position, distance	0.750	0.769	0.750	0.806	0.756	0.731
Position	0.750	0.538	0.631	0.619	0.631	0.656
Force, position, distance, suction position	0.788	0.744	0.769	0.825	**0.850**	0.800
Force, position, distance, suction-drill distance	0.738	0.788	0.819	0.769	0.819	0.813

features, consisting of force magnitudes, velocity magnitudes, drill positions, suction positions, distances from the facial nerve, and drill-suction distances, were examined. ''Positions'' were compressed to one dimension: distances from the origin. Data were sampled 10 times per second, with a 25-millisecond sliding window smoothing the original data sampled at 1-millisecond intervals. This resulted in data sets of on the order of several thousand observations each.

The feature set consisting of force magnitudes, drill positions, drill distances from the facial nerve, and suction positions proved most successful in our experiments. Using a seven-state HMM, an average of 6.8 of the 8 experts (85%) and 6.8 of the 8 novices (85%) were correctly classified. Sensitivity to initial parameter guesses was low (standard deviation: 0.4 for experts, 0.6 for novices). The seven-state HMM architecture worked best for this data. Results for additional reasonably successful feature sets are given in Table II.

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

Metric	r	p
Drilling technique		
Pct bone visible at removal (1)	0.728	<0.001
Pct removed with burr too large (2)	−0.405	0.022
Jump frequency (3)	−0.226	0.213
Mean drill angle (4)	−0.405	0.022
Suctioning technique		
Pct excessive inter-tool distance (5)	−0.469	0.007
Pct excessive dust (6)	−0.365	0.040
Bone removal		
Mean removal probability (7)	0.337	0.059
Pct improbable (non)removals (8)	−0.794	<0.001
Facial nerve exposure		
Pct directly exposed (9)	0.469	0.007
Pct direct or indirect exposed (10)	0.519	0.002
Pct overexposed (11)	−0.536	0.002
Drill forces		
Pct excessive force (12)	−0.355	0.046
Near dura (13)	−0.471	0.007
Near sigmoid (14)	−0.420	0.017
Near facial nerve (15)	−0.563	<0.001
Near inner ear (16)	−0.468	0.007
Drill Velocities		
Pct excessive velocity (17)	−0.131	0.474
Near dura (18)	−0.152	0.407
Near sigmoid (19)	−0.143	0.434
Near facial nerve (20)	−0.387	0.029
Near inner ear (21)	−0.339	0.058

Validating the metrics

By computing the values of the metrics for each of the virtual mastoidectomies performed in the user study described in the previous section, we attempted to validate the metrics.

Results from instructor-assigned scores. The correlations of each of the metrics with the average of the two global scores assigned by the instructors are shown in Table III. Most of the metrics (1, 2, 4–6, 8–16, 20) had a statistically significant correlation ($p < 0.05$) with assigned global scores. The strongest correlations were for Metrics 1 (visibility) and 8 (bone removal region). While the avoidance of applying excessively large drill forces (12–16) did have a significant (though somewhat weak) correlation with performance, there was little such correlation for overall drill velocities (17), or for velocities when near the dura or sigmoid (18, 19), but there was somewhat of a correlation for velocities when near the facial nerve and inner ear (20, 21). 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.

The number of voxels for which the user's choice to remove or not to remove differed from that of a large majority of the experts (8) correlated much better with instructor scores than the mean expert probabilities of removed voxels (7), perhaps because

Table IV. 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 (5)	−0.737	<0.001	−0.469	0.007
Pct excessive dust (6)	−0.681	<0.001	−0.365	0.040
Mean removal probability (7)	0.323	0.072	0.337	0.059
Pct improbable choices of removal regions (8)	−0.736	<0.001	−0.794	<0.001
Pct of facial nerve directly exposed (9)	0.400	0.023	0.469	0.007
Pct of facial nerve directly or indirectly exposed (10)	0.411	0.019	0.519	0.002
Pct of facial nerve overexposed (11)	−0.500	0.004	−0.536	0.002

only the former considers non-removed voxels, and the experts may have been careful to remove all of the required region, or because experts may have different styles and vary somewhat from the ''average'' expert while still rarely making any extremely unorthodox decisions. However, the results for the bone removal metrics must be viewed with caution because there was overlap in the training data used to define expert removal probabilities and the test data. Thus, a more careful analysis of this metric using leave-one-out cross-validation was performed, which also showed a strong correlation ($r = 0.74$, $p < 0.00001$).

Direct exposure (9) correlated reasonably well with instructor scores, and adding the computed inferred regions to the directly exposed regions (10) further strengthened the correlation, from $r = 0.469$ to $r = 0.519$. Overexposure (11) had an even stronger (negative) correlation ($r = -0.536$).

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. Several metrics correlated much more strongly with these specific sub-scores than with the global scores, as shown in Table IV. The frequency of drill jumps (3) was found to be significantly (though weakly) correlated with the instructor's assessment of making ''purposeful, confident motions'', while the distance between instruments while drilling (5) and the percent of time drilling with excessive bone dust in the surgical field (6) strongly correlated with the specific 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 most scores or relatively low on most scores.

Results from classification algorithms. In addition to using raw data as features, HMMs were also developed using metric values as the features in order to evaluate the discriminative validity of the metrics. Time-varying streams were acquired by recording, as each voxel was removed, whether it was drilled while using proper technique according to several different metrics. These binary data streams were then smoothed using a sliding window of width 25. Leave-out-one cross-validation was then performed using the sixteen expert and novice data sets with Hidden Markov Models, as in the *Results from classification algorithms* sub-section of the *Validating the simulator* section above. Correct classification rates of 87.5% were consistently obtained using models of one, three, five, seven, and nine states in repeated trials with random initializations, using distance to facial nerve and metric scores for visibility, removal region, force magnitude, suction to drill distance, jump frequency, burr choice, and dust accumulation as the features. The relative ease with which high classification rates could be achieved using a variety of models and initializations offers some evidence of the metrics' validity for differentiating skill level.

A logistic regression classifier was also trained with leave-out-one cross-validation independently for each of 20 metrics. Results, showing the proportion of experts and novices correctly classified when using each metric, are presented in Table V. The visibility metric was successful in all cases, suggesting that this is a good indicator of expertise. Metrics testing subjects' abilities to remove the correct bone and to keep their instruments close together had correct classification rates of 87.5%. Application of excessive forces in general was not discriminative (50%), but application of excessive forces near the facial nerve was highly discriminative (87.5%), underscoring the experts' recognition of areas in which extra caution is necessary.

Validating the feedback mechanisms

In order to validate our performance evaluation console as an instructional tool that can lead to

Table V. Percentage of experts and novices correctly classified with leave-one-out cross-validation using a logistic regression classifier with individual metrics as features.

	Proportion correct		
Metric	**Expert**	**Novice**	**Overall**
Pct bone visible at removal	1.000	1.000	1.000
Improbable nonremovals	0.875	0.875	0.875
Pct excessive inter-tool distance	0.875	0.875	0.875
Pct exc forces near facial nerve	0.875	0.875	0.875
Pct removed with burr too large	0.875	0.750	0.813
Pct overexposed facial nerve	1.000	0.625	0.813
Mean removal probability	0.750	0.750	0.750
Improbable removals	0.750	0.750	0.750
Pct directly exposed facial nerve	0.875	0.625	0.750
Average drill angle deviation from 90°	0.875	0.625	0.750
Pct excessive velocities near facial nerve	0.875	0.500	0.688
Pct excessive dust	0.750	0.625	0.688
Pct excessive velocities	0.750	0.500	0.625
Number of collisions	0.875	0.375	0.625
Jump frequency	0.750	0.500	0.625
Pct removed with burr too small	0.875	0.375	0.625
Pct indirectly exposed facial nerve	0.750	0.500	0.625
Pct overexposed chorda	0.875	0.375	0.625
Pct overexposed sigmoid	1.000	0.125	0.563
Pct excessive forces	0.375	0.625	0.500

improved performance and to begin to explore how it could be used most effectively, we conducted a user study in which ten subjects were asked to perform four trials each in which they were to properly expose a wishbone-shaped tubular structure (mimicking a nerve or blood vessel) embedded in a cube of bone in our simulator. The orientation, thickness, and twisting of the structure varied randomly between trials. All the participants were shown a video of how to properly perform the virtual procedure and given a description of the metrics with which their performance would be evaluated: voxel visibility, structure direct exposure and overexposure, and drill-tube collisions. The interface for the simulator was simplified from our general mastoidectomy simulator in several ways; for example, there were no menus (the right hand was always the drill and the left always the camera) and no bone dust. Also, inferred exposure was not considered.

All participants were notified with a beep when they caused an injury by contacting the tube with the drill, but four participants (the control group) received no additional direct feedback. The others (the feedback group) were instructed on the use of the performance evaluation console and allowed to independently review their performance on each trial according to the metrics before performing their next trial, and were also provided with interactive feedback with regards to voxel visibility, except during the final trial (Figure 6). Within the feedback group, half received the augmented feedback during and after each trial, while the other half received this feedback only during and after every other trial.

Overall, the feedback group learned to maintain proper visibility more effectively than the control group, with the difference between their percentage of properly removed voxels on the final trial (during which no participants received augmented feedback) being significant at the 92% confidence level ($p = 0.074$). Within the feedback group, those who received feedback with every other trial tended to do somewhat better than those who received feedback with every trial (although this difference was not statistically significant). This is consistent with

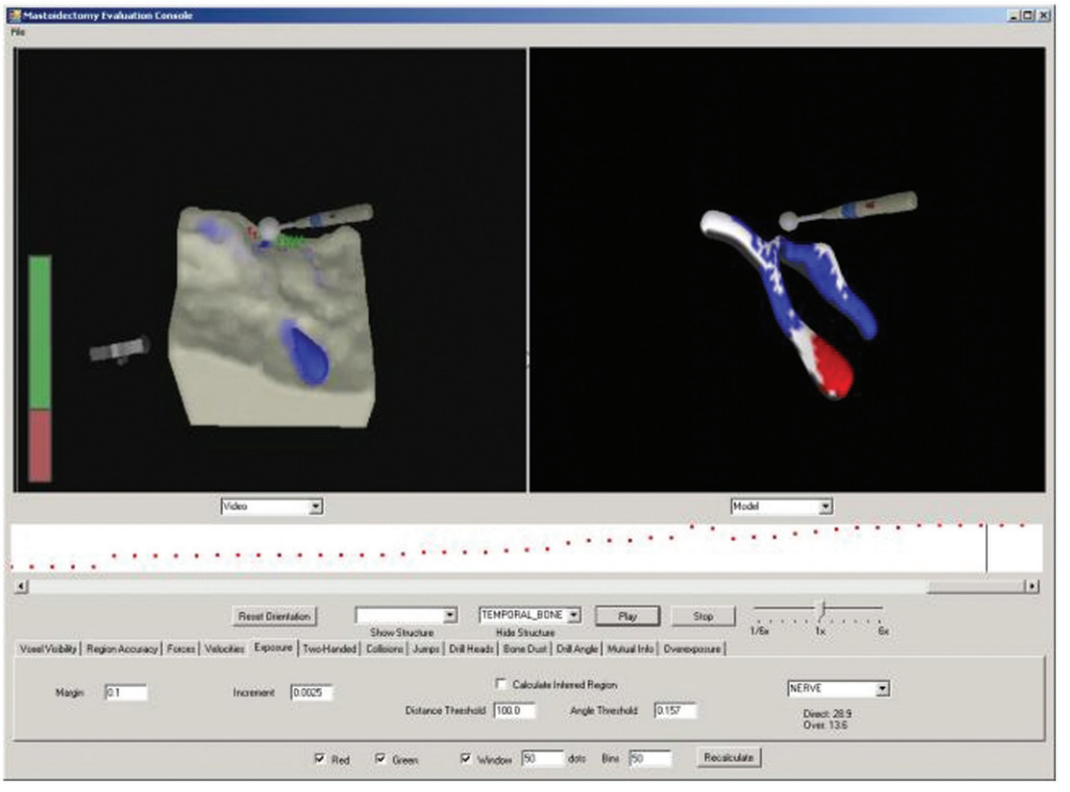

Figure 6. Reviewing a procedure in the performance evaluation console study. Members of the Feedback Group received interactive feedback about their maintenance of visibility with red (poor) and green (good) dots as bone was drilled away and a meter showing the percentage of red and green voxels over the last 50 removals, as shown in the video replaying in the left panel. After completing a trial, they were able to review their performance with respect to visibility, exposure/overexposure, and collisions with the tube. Shown at right is a visualization of achieved exposure, with properly exposed regions in white, overexposed regions in red, and unexposed regions in blue. [Color version available online.]

learning theory results suggested in such papers as reference [40], because people may become dependent on the feedback if it is always available (as was actually noted by one participant in the group that received feedback with every trial). The difference in the means between the alternating feedback subgroup and the control group was significant at the 94% confidence level ($p = 0.058$). The feedback group did not receive much more information about the other metrics than the control group, since all were given beeps when colliding with the nerve and there was significant inherent feedback with regards to exposure, so there were no significant differences with regards to these aspects.

Conclusion

In this paper, we have described our mastoidectomy simulator, proposed a wide variety of algorithms for assessing performance on this simulator and providing constructive feedback to the user, and reported on several user studies attempting to establish some level of validation for this work. By considering a large number of metrics and feedback mechanisms such as these, simulators may soon be able to serve as a virtual instructor that may be an adequate substitute for the real instructor throughout much of the learning process, greatly reducing the time demands on instructors and increasing learning opportunities for trainees. In addition to making time spent on the simulator more educational, users have also noted, not surprisingly, that the inclusion of metrics and visual feedback makes the simulator experience more fun, competitive, and intense, which is likely to result in more time being spent learning on the simulator. With the increasing emphasis on incorporating patient-specific data into simulators, the value of such ''intelligent tutors'' may be actualized not only for young surgical residents but also for experienced surgeons in the context of patient-specific rehearsal for upcoming procedures.

All of the metrics presented in the preceding sections have been developed and implemented in the context of a specific procedure: mastoidectomy. However, nearly all are based on principles that are common throughout the surgical profession. Maintaining proper visibility of the surgical field, sufficiently exposing critical anatomic structures, applying appropriate forces and velocities as vulnerable structures are approached, removing the optimal volume of bone or tissue, and exercising efficient, safe, and well-coordinated control of surgical instruments are essential components of good technique in a wide variety of procedures. For example, properly exposing the recurrent laryngeal nerve during thyroid surgery and the cystic duct during gall bladder removal are essential skills for the general surgeon; choosing the correct tissue for removal is vital in bile duct excision for biliary atresia and choledochal cyst; and application of appropriate forces and skillful manipulation and coordination of instruments are paramount in all laparoscopic procedures. Nevertheless, one of the guiding principles of this work was to use domain-specific knowledge to mimic the evaluation criteria used by expert ear surgeons, so, just as human otolaryngologists are better tutors for residents in ear surgery than orthopedic surgeons (and vice versa), it may be unrealistic to expect automated evaluation methods to generalize too broadly without additional domain-specific knowledge.

Much remains to be done with regards to the validation of our simulator and our metrics. We have not shown that training on our simulator or receiving our feedback visualizations improves performance in the operating room, nor that metric scores obtained while performing a virtual procedure are good predictors of corresponding aspects of performance during a real procedure. Furthermore, while we have shown that use of our performance evaluation console can facilitate learning at least within the virtual environment, there is a wide array of questions about such factors as when to provide such feedback (interactively or in a post-procedure debriefing session?), how to provide it (which visualizations are most useful?), and how often to provide it (during or after each trial, or with some schedule of decreasing frequency?). The answers to many of these questions may be different for different metrics.

Acknowledgments

Support for this research was provided by NIH LM07295.

References

1. Luperfoy S. Intelligent tutoring technology: Accelerating change in medical instruction. Plenary session at Medicine Meets Virtual Reality 14 (MMVR14), Long Beach, CA, January 2006.
2. Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. Am J Surg 1999;177:28–32.
3. Anastakis DJ, Wanzel KR, Brown M, Herold J, McIllroy J, Hamstra S, Ali J, Hutchison C, Murnaghan J, Reznick R, Regehr G. Evaluating the effectiveness of a 2-year curriculum in a surgical skills center. Am J Surg 2003;185:378–385.

4. Paisley AM, Baldwin PJ, Paterson-Brown S. Accuracy of medical staff assessment of trainees' operative performance. Med Teach 2005;27:634–638.
5. Gates EA. New surgical procedures: Can our patients benefit while we learn? Am J Obstet Gynecol 1997;176:1293–1298; discussion 98-99.
6. Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: Where to begin? Surgery 2004;135:6–20.
7. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative ''bench station'' examination. Am J Surg 1996;172:226–230.
8. Tuchschmid S, Bajka M, Bachofen D, Szekely G, Harders M. Objective surgical performance assessment for virtual hysteroscopy. In: Westwood JD, Haluck RS, Hoffman HM, Mogel GT, Phillips R, Robb RA, Vosburgh KG, editors. Proceedings of Medicine Meets Virtual Reality 15 (MMVR15), Long Beach, CA, February 2007. Studies in Health Technology and Informatics 125. Amsterdam: IOS Press; 2007. pp 473–478.
9. Cotin S, Stylopoulos N, Ottensmeyer M, Neumann P, Rattner D, Dawson S. Metrics for laparoscopic skills trainers: The weakest link! In: Dohi T, Kikinis R, editors. Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2002), Tokyo, Japan, September 2002. Lecture Notes in Computer Science 2488. Berlin: Springer; 2002. pp 35-43.
10. Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng 2001;48(5):579–591.
11. Dosis A, Bello F, Gillies D, Undre S, Aggarwal R, Darzi A. Laparoscopic task recognition using Hidden Markov Models. In: Proceedings of Medicine Meets Virtual Reality 13 (MMVR13), Long Beach, CA, January 2005. Studies in Health Technology and Informatics 111. Amsterdam: IOS Press; 2005. pp 115–122.
12. 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. In: Proceedings of Medicine Meets Virtual Reality 14 (MMVR14), Long Beach, CA, January 2006. Studies in Health Technology and Informatics 119. Amsterdam: IOS Press; 2006. pp 355–360.
13. Murphy T. Towards objective surgical skill evaluation with Hidden Markov Model-based motion recognition. M.S. thesis, Department of Mechanical Engineering, The Johns Hopkins University, August 2004.
14. Ritter EM, McClusky DA, Gallagher AG, Smith CK. Real-time objective assessment of knot quality with a portable tensiometer is superior to execution time for assessment of laparoscopic knot-tying performance. Surgical Innovation 2005;12(3):233–237.
15. Rothstein AL, Arnold RK. Bridging the gap: Application of research on videotape feedback and bowling. Motor Skills: Theory Into Practice 1976;1:35–62.
16. Kernodle MW, Carlton LG. Information feedback and the learning of multiple-degree-of-freedom activities. J Motor Behavior 1992;24:187–196.
17. Feygin D, Keehner M, Tendick F. Haptic guidance: Experimental evaluation of a haptic training method for a perceptual motor skill. In: Proceedings of the 10th IEEE Haptics Symposium, Orlando, FL, March 2002.
18. Yang U, Kim GJ. Implementation and evaluation of Just Follow Me: An immersive, VR-based, motion-training system. Presence: Teleoperators and Virtual Environments 2002;11(3):304–323.
19. Morris D, Tan HZ, Barbagli F, Chang T, Salisbury K. Haptic feedback enhances force skill learning. In: Proceedings of the IEEE World Haptics Conference, Tsukuba, Japan, March 2007. pp 21–26.
20. Kahol K, Tripathi P, Panchanathan S. Documenting motion sequences: Development of a personalized annotation system. IEEE Multimedia Magazine 2007 (in press).
21. Rissanen MJ, Kuroda Y, Nakao M, Kuroda T, Nagase K, Yoshihara H. A novel approach for training of surgical procedures based on visualization and annotation of behavioural parameters in simulators. In: Westwood JD, Haluck RS, Hoffman HM, Mogel GT, Phillips R, Robb RA, Vosburgh KG, editors. Proceedings of Medicine Meets Virtual Reality 15 (MMVR15), Long Beach, CA, February 2007. Studies in Health Technology and Informatics 125. Amsterdam: IOS Press; 2007. pp 388–393.
22. Silverstein J, Selkov G, Salud L, Pugh C. Developing performance criteria for the E-Pelvis Simulator using visual analysis. In: Westwood JD, Haluck RS, Hoffman HM, Mogel GT, Phillips R, Robb RA, Vosburgh KG, editors. Proceedings of Medicine Meets Virtual Reality 15 (MMVR15), Long Beach, CA, February 2007. Studies in Health Technology and Informatics 125. Amsterdam: IOS Press; 2007. pp 436–438.
23. Agus M, Giachetti A, Gobbetti E, Zanetti G, Zorcolo A. A multiprocessor decoupled system for the simulation of temporal bone surgery. Computing and Visualization in Science 2002;5(1):35–43.
24. Agus M, Brelstaff GJ, Giachetti A, Gobbetti E, Zanetti G, Zorcolo A, Picasso B, Franceschini SS. Physics-based burr haptic simulation: Tuning and evaluation. In: Proceedings of the 12th IEEE Haptics Symposium, Chicago, IL, March 2004. pp 128–135.
25. Pflesser B, Petersik A, Tiede U, Höhne KH, Leuwer R. Volume cutting for virtual petrous bone surgery. Comput Aided Surg 2002;7(2):74–83.
26. Petersik A, Pflesser B, Tiede U, Höhn KH, Leuwer R. Haptic volume interaction with anatomic models at sub-voxel resolution. In: Proceedings of IEEE Virtual Reality (VR 2002), Orlando, FL, March 2002.
27. Renz M, Preusche C, Potke M, Kriegel HP, Hirzinger G. Stable haptic interaction with virtual environments using an adapted voxmap-pointshell algorithm. In: Proceedings of Eurohaptics, Birmingham, England, July 2001. pp 149–154.
28. Bryan J, Stredney D, Wiet G, Sessanna D. Virtual temporal bone dissection: A case study. In: Ertl T, Joy KI, Varshney A, editors. Proceedings of IEEE Visualization 2001, San Diego, CA, October 2001. pp 497–500.
29. Stredney D, Bryan J, Sessanna D, Weit G. Integration of an intelligent tutor in interactive medical simulation. Presentation at Ohio Learning Network Windows on the Future 2003, Columbus, OH, March 2003.
30. Morris D, Sewell C, Barbagli F, Blevins N, Girod S, Salisbury K. Visuohaptic simulation of bone surgery for training and evaluation. IEEE Trans Comput Graph Appl 2006;26(6):48–57.
31. Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K. Quantifying risky behavior in surgical simulation. In: Proceedings of Medicine Meets Virtual Reality 13 (MMVR13), Long Beach, CA, January 2005. Studies in Health Technology and Informatics 111. Amsterdam: IOS Press; 2005. pp 451–457.
32. Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K. Achieving proper exposure in surgical simulation. In: Proceedings of Medicine Meets Virtual Reality 14

(MMVR14), Long Beach, CA, January 2006. Studies in Health Technology and Informatics 119. Amsterdam: IOS Press; 2006. pp 497–502.
33. Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K. Evaluating drilling and suctioning technique in a mastoidectomy simulator. In: Westwood JD, Haluck RS, Hoffman HM, Mogel GT, Phillips R, Robb RA, Vosburgh KG, editors. Proceedings of Medicine Meets Virtual Reality 15 (MMVR15), Long Beach, CA, February 2007. Studies in Health Technology and Informatics 125. Amsterdam: IOS Press; 2007. pp 427–432.
34. Morris D. Algorithms and data structures for haptic rendering: Curve constraints, distance maps, and data logging. Stanford University Department of Computer Science Technical Report 2006-06.
35. http://www.wischik.com/lu/programmer/avi_utils.html
36. http://www.chai3d.org
37. Baum LE, Egon JA. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull Amer Meteorol Soc 1967;73:360–363.
38. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Statist 1970;41(1):164–171.
39. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html
40. Winstein CJ, Schmidt RA. Reduced frequency of knowledge of results enhances motor skill learning. J Exp Psych: Learning, Memory and Cognition 1990;16:677–691.