Evaluating Drilling and Suctioning Technique in a Mastoidectomy Simulator

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

Abstract. This paper presents several new metrics related to bone removal and suctioning technique in the context of a mastoidectomy simulator. The expertise with which decisions as to which regions of bone to remove and which to leave intact is evaluated by building a Naïve Bayes classifier using training data from known experts and novices. Since the bone voxel mesh is very large, and many voxels are always either removed or not removed regardless of expertise, the mutual information was calculated for each voxel and only the most informative voxels used for the classifier. Leave-out-one cross validation showed a high correlation of calculated expert probabilities with scores assigned by instructors. Additional metrics described in this paper include those for assessing smoothness of drill strokes, proper drill burr selection, sufficiency of suctioning, two-handed tool coordination, and application of appropriate force and velocity magnitudes as functions of distance from critical structures.

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

Introduction

In order to move towards the goal of enabling simulators to serve as intelligent, mostly-autonomous virtual instructors of surgical skill, we have previously proposed [1, 2] a number of metrics intended to capture some of the most important aspects of good technique that a real instructor tries to teach his/her residents in the field of temporal bone surgery, using our simulator [3]. In this paper we present several new metrics related to bone removal and suctioning technique.

Most existing surgical simulators, especially laparoscopic skill trainers, have attempted to incorporate a small number of simple metrics [4]. Most assume a simple global optimum value, such as minimize wall collisions, maximize path efficiency, or minimize completion time, and do not attempt to learn from runs of the simulators by experts or novices. Several have used learning algorithms such as Markov Models [5] or neural nets [6] to evaluate surgical performance.

1. Naïve Bayes Classifier for Removed Bone Voxels

One of the most obvious criteria for the evaluation of a mastoidectomy, a procedure in which part of the temporal bone is drilled away in order to access the inner ear, is

whether correct decisions were made as to which regions of bone to remove and which to leave intact. A simple method is to have an instructing surgeon label which regions should and should not be removed, or to automatically label the voxels (used as the underlying representation of the bone volume in our simulator) 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, similar to how many e-mail spam classification algorithms assume that words from a dictionary are chosen for an e-mail message according to separate distributions by spammers and non-spammers [7], we have made an assumption that voxels from the full voxel mesh are chosen for removal according to separate distributions for experts and novices. We have implemented a Naïve Bayes classifier that calculates the maximum likelihood estimates for the probabilities that each voxel is removed by an expert and by a novice, and uses these to determine the probabilities that a given mastoidectomy was performed by an expert or by a novice.

If $y \in \{0,1\}$ are the class labels (0 = novice, 1 = expert), there are n voxels in the temporal bone model, and $\mathbf{x} \in \{0,1\}^n$ is an n-dimensional vector encoding whether each of the n voxels was (1) or was not (0) removed in a particular simulator run, then, making the Naïve Bayes assumption that the x_i's are conditionally independent given y, the probability of a particular set of choices for removal or non-removal for each voxel, given the class which generated it, can be written

$$p(\mathbf{x} \mid y) = \prod_{i=1}^{n} p(x_i \mid y)$$

The model is parameterized by $\phi_{i|y=k} = p(x_i=1 \mid y=k)$, the probabilities for each voxel i that it is removed in a run performed by a member of class y=k, and $\phi_y = p(y=1)$, the prior probability that a run was performed by an expert (y=1). We assume no prior probabilities, so we set $\phi_y = 1/2$. Given a set of m training examples, the maximum likelihood estimates for the other parameters, using Laplace smoothing (adding one phantom example that removes every voxel and one that removes none in each class, so as to avoid zero probabilities for any voxel), and denoting $1\{s\}$ as the function that returns 1 if s is true and 0 if s is false, are simply the fractions of examples in each class in which the voxels were removed:

$$\phi_{i|y=k} = \frac{\sum_{j=1}^{m} 1\left\{\left(x_i^{(j)} = 1\right) \wedge \left(y^{(j)} = k\right)\right\} + 1}{\sum_{j=1}^{m} 1\left\{y^{(j)} = k\right\} + 2}$$

Then, given a new set $\mathbf{x}$ of voxel removal choices, the probability that this was generated by an expert can be estimated using Bayes' Rule as:

$$p(y=1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=1)\,p(y=1)}{p(\mathbf{x})} =$$

$$\frac{\left(\prod_{i=1}^{n} p(x_i \mid y=1)\right)\left(\frac{1}{2}\right)}{\left(\prod_{i=1}^{n} p(x_i \mid y=1)\right)\left(\frac{1}{2}\right)+\left(\prod_{i=1}^{n} p(x_i \mid y=0)\right)\left(\frac{1}{2}\right)} =$$

$$\frac{\prod_{i=1}^{n}\left(1\{x_i=1\}\phi_{i|y=1}+1\{x_i=0\}(1-\phi_{i|y=1})\right)}{\prod_{i=1}^{n}\left(1\{x_i=1\}\phi_{i|y=1}+1\{x_i=0\}(1-\phi_{i|y=1})\right)+\prod_{i=1}^{n}\left(1\{x_i=1\}\phi_{i|y=0}+1\{x_i=0\}(1-\phi_{i|y=0})\right)}$$

However, since the bone mesh is so large, and so many voxels are likely to not be very informative (i.e., 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 n=1000 most informative voxels. The mutual information between each voxel x_i and the class labels y was calculated as

$$\mathrm{MI}(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\,p(y)}$$

with each of the probabilities estimated using their empirical distributions in the training set.

We then evaluated this metric by performing leave-one-out cross validation with our classifier. There was a statistically significant correlation ($r = 0.740$, $p<0.00001$) between the calculated probability estimates and a one-to-five subjective global score assigned by an instructing surgeon who watched video replays of the procedures (Figure 1).

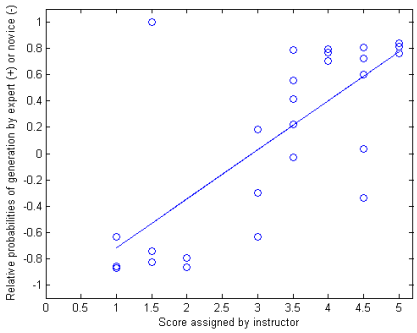

Figure 1. Correlation between instructor-assigned score and calculated probability of being in expert class.

This metric, along with all of our others, has been incorporated into our metrics console. When the simulator is run, all of the data is logged and can then be rendered

to video or loaded into the console, which computes metric scores and provides a number of visualizations intended to help the user highlight potential problem areas. The user can be shown red dots at the locations of each voxel he/she removed for which the expert removal probability was below a specified threshold value (Figure 2, left), or red dots for each voxel not removed for which the expert removal probability was above a specified threshold value (Figure 2, right). The percentage of low-expert-probability voxels removed can be plotted at specified intervals on a timeline, allowing the user to quickly fast-forward to mistakes.

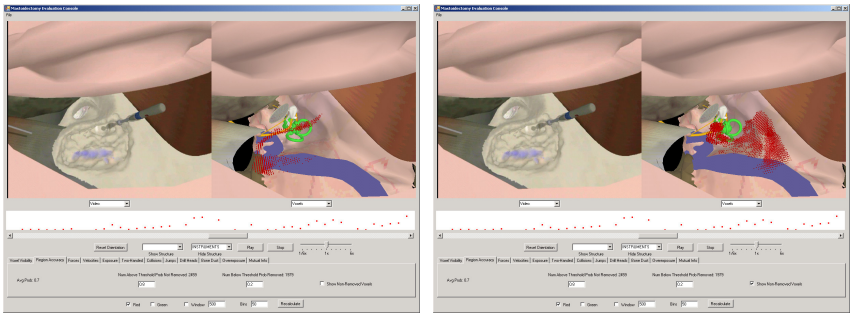

Figure 2. At left, improperly removed voxels shown in red. At right, improperly remaining voxels in red.

This analysis can also yield an interesting visualization of the most informative voxels, which provides useful insight into the regions of bone most likely to be removed by experienced surgeons but left by novices and vice versa. In Figure 3, the 1000 most informative voxels (based on the training data) are shown. Those more likely to be removed by experts are in gray (brighter corresponds to greater expert-novice discrepancy), while those more likely to be removed by novices are in red. There are more voxels of the former case presumably due to greater uniformity among experts than among novices.

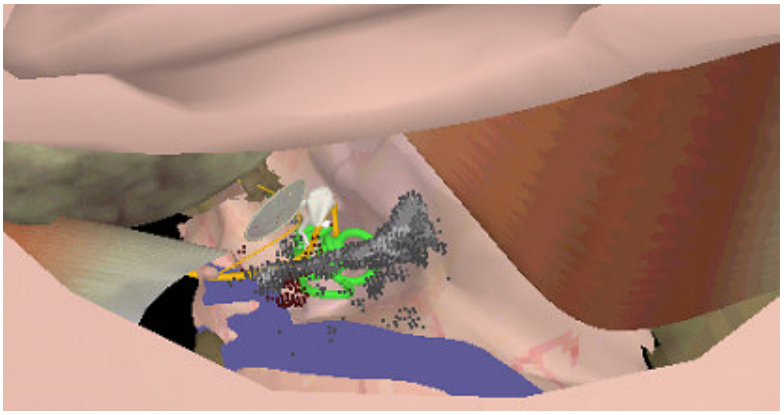

Figure 3. The 1000 most informative voxels (likely expert removals in gray, likely novice removals in red).

2. Other Metrics for Drilling Technique

Another key indicator of surgical skill is the exhibition of "purposeful" movements in drilling. An expert will almost invariably work locally, accomplishing a specific sub-goal (such as exposing a certain structure), and then move on to another task, and make smooth, continuous motions with the drill. A novice, on the other hand, is more likely to move around haphazardly without recognition for the localized subtasks that should be completed. Therefore, we have included a metric that reports the frequency of drill "jumps": the number of removed voxels per thousand that were more than a specified distance away from the previously removed voxel.

It is also important for a surgeon to know when to use each of his/her tools. In our simulator, the user can switch between 6mm and 3mm drill burrs. The smaller burr is intended for use near delicate structures, while the larger burr allows for quicker drilling in safer areas. The fraction of the time each burr is used to remove each voxel is learned from expert training data, and the user can be shown voxels he/she removed with the burr opposite the one used by the experts more than a specified fraction of the time for that voxel.

3. Metrics for Suctioning Technique

Good technique in the use of the suction involves removing bone dust as it is created in order to maintain visibility of the bone surface. In our simulator, particles of bone dust are generated as bone is removed, and a suctioning device (a second haptic device held in the opposite hand as the drill) is used to remove these particles. We have included a metric that highlights times in which the user was drilling while more than a specified number of dust particles obscured the surgical field. An example of excessive bone dust accumulation in the simulator is shown in Figure 4 (left side).

The coordination of the drill and the suction is an important element of good "two-handed technique." The suction should be kept near the drill when removing bone in order to prevent accumulation of bone dust and to properly cool the drilling surface (since the suction tool also provides irrigation). Therefore, another metric identifies voxels removed with the drill and suction more than a specified distance apart.

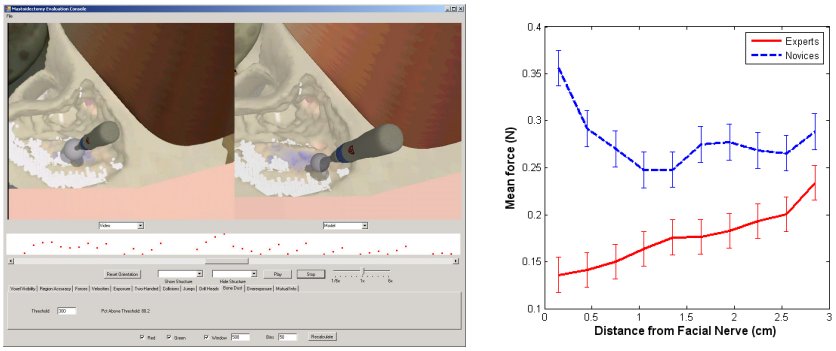

Figure 4. At left, excessive accumulated bone dust. At right, mean drill forces as facial nerve approached.

4. Metrics for Forces and Velocities

Applying appropriate forces and operating the drill at appropriate velocities are critical to safe drilling practice. In general, the magnitudes of these forces and velocities should decrease as vulnerable structures, such as the facial nerve, are approached, as shown in Figure 4 (right side), which shows force magnitudes as a function of distance from the facial nerve for experts and novices in our training data. The metrics console can highlight all voxels removed while applying force or velocity magnitudes above specified thresholds, or just such voxels within specified distances of critical structures. The values of these "safety thresholds" for forces and velocities can be assigned by an instructor or estimated from the expert training data.

5. Discussion

By considering all of these metrics together, as well as those we have previously proposed and ones yet to be developed, it is hoped that eventually the virtual instructor may be an adequate stand-in for the real instructor throughout much of the learning process, greatly reducing the time demands on instructors. It is therefore essential that we be able to establish that the feedback provided by these metrics mirrors that given by live instructors. We have conducted a study in which we have correlated the scores returned by these metrics with instructors' evaluations, the results of which are reported in [8]. We are also exploring additional metrics, including analyses of force, positional, and velocity profiles using time series classification.

References and Acknowledgements

Support was provided by NIH LM07295.

[1] 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.

[2] 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.

[3] 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 Transactions on Computer Graphics and Applications, November 2006.

[4] Cotin S, Stylopoulos N, Ottensmeyer M, Neumann P, Rattner D, Dawson S: Metrics for laparoscopic skills trainers: the weakest link! Proc of MICCAI, Lecture Notes in Computer Science 2002, 288: 35-43.

[5] Rosen J, Solazzo M, Hannaford B, Sinanan M: Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions. Proc of MMVR 2001.

[6] Huang J, Payandeh S, Doris P, Hajshirmohammadi I: Fuzzy classification: towards evaluating performance on a surgical simulator. Proc of MMVR 2005, pp. 194-200.

[7] Sahami M, Dumais S, Heckerman D, Horvitz E: A bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization, 1998.

[8] Sewell C, Morris D, Blevins N, Agrawal S, Dutta S, Barbagli F, Salisbury K: Validating metrics for a mastoidectomy simulator. To appear in proceedings of Medicine Meets Virtual Reality, February 2007.