Published in final edited form as:

Int J Comput Assist Radiol Surg. 2012 January ; 7(1): 1–11. doi:10.1007/s11548-011-0566-4.

Automatic Scoring of Virtual Mastoidectomies Using Expert Examples

Thomas Kerwin,

Ohio Supercomputer Center, Columbus, Ohio, USA . Department of Computer Science and Engineering, Ohio State University, Columbus, Ohio, USA

Gregory Wiet,

Department of Otolaryngology and Biomedical Informatics, Nationwide Children's Hospital, Columbus, Ohio, USA. The Ohio State University Medical Center, Columbus, Ohio, USA

Don Stredney, and

Ohio Supercomputer Center, Columbus, Ohio, USA

Han-Wei Shen

Department of Computer Science and Engineering, Ohio State University, Columbus, Ohio, USA

Thomas Kerwin: kerwin@osc.edu; Gregory Wiet: gregory.wiet@nationwidechildrens.org; Don Stredney: don@osc.edu; Han-Wei Shen: hwshen@cse.ohio-state.edu

Abstract

Purpose—Automatic scoring of resident performance on a virtual mastoidectomy simulation system is needed to achieve consistent and efficient evaluations. By not requiring immediate expert intervention, the system provides a completely objective assessment of performance as well as a self-driven user assessment mechanism.

Methods—An iconic temporal bone with surgically important regions defined into a fully partitioned segmented dataset was created. Comparisons between expert-drilled bones and student-drilled bones were computed based on gradations with both Euclidean and Earth Mover's Distance. Using the features derived from these comparisons, a decision tree was constructed. This decision tree was used to determine scores of resident surgical performance. The algorithm was applied on multiple expert comparison bones and the scores averaged to provide reliability metric.

Results—The reliability metrics for the multi-grade scoring system are better in some cases than previously reported binary classification metrics. The two scoring methods given provide a trade-off between accuracy and speed.

Conclusions—Comparison of virtually drilled bones with expert examples on a voxel level provides sufficient information to score them and provide several specific quality metrics. By merging scores from different expert examples, two related metrics were developed; one is slightly faster and less accurate, while a second is more accurate but takes more processing time.

Keywords

Automatic evaluation; Objective assessment; Mastoidectomy; Surgical simulation; Temporal bone

1 Introduction

An integral and essential part of surgical training is evaluation. This is true whether the training is through a virtual simulation system, a physical simulation (i.e. a cadaveric specimen), or supervised interaction with real patients. In one-to-one training methods, an experienced surgeon provides immediate feedback to the resident. However, this type of

training time is intensive, costly, often unavailable, and can potentially be influenced by the evaluator's personal biases. With a virtual simulation system, objective evaluation of performance and active feedback can be provided to the user. We have previously reported on the development of temporal bone simulator system for learning the anatomy and surgical techniques associated with a basic mastoidectomy [2, 9]. This system employs volume rendering of temporal bone data acquired from CT scans, delivers haptic feedback during drilling using a 3D joystick, and plays aural feedback to the user with modulated drilling sounds. The system provides a realistic multi-modal environment for learning surgical technique. A screenshot and photo of the simulator system can be seen in Fig. 1. Our intent is not to replace one-on-one training, but to create a system that serves as additional training experience without the need for direct supervision by an attending surgeon. Part of the goal of this system is to give feedback to users in a way that can support its use in an educational curriculum by providing both formative and summative evaluations.

A multi-institutional trial consisting of eight institutions was designed to test the efficacy of temporal bone surgical training in a simulator versus traditional training in a cadaveric laboratory [23]. All participants were asked to drill the same virtual bone (referred to as bone ID number 9413R) before and after receiving training in performing the surgical techniques associated with a mastoidectomy. This task was in addition to drilling other bones, real and virtual. Experts performed the same exercise on that same virtual bone to provide a standard for comparison with the trainee drilled bones. We will refer to a data volume consisting of the end product of a resident performing a mastoidectomy in the simulator as a resident bone and a data volume from an expert performing the same task on the simulator as an expert bone. The combination of expert and resident bones forms our dataset. This set of varying surgical performances starting from identical initial conditions forms the basis of our analysis.

Virtual simulation performance in otology and other surgical specialties have generally been evaluated by direct observation by trained experts in the given domain, and many different scales have been proposed for grading technical skills performance. Our work thus far has been based on the Welling scale [3], which is designed to measure performance in specific tasks in a temporal bone dissection lab executed on a cadaveric specimen. Other scales have been presented as well, including one by Laeeq et al. [11], which has similar goals but also attempts to expand applicability to surgical performance. A composite scale of available metrics on temporal bone dissection/mastoidectomy has been published with a classification schema for potential application to computer scoring [22]. Since these types of scales are meant to be applied by experienced surgeons, they often contain terms that are extremely difficult to quantify. The definition of "proper saucerization" in the Welling scale is a good example of this problem. Additionally, use of such terms potentially introduces error in expert evaluation: since there is no clearly defined quantitative definition, evaluations are subject to the expert's own interpretation which may vary from person to person and even from time to time. In order for drilled mastoid bones to be scored on these metrics by an algorithm, we require either a precise definition of these types of terms or a data-driven approach. This article details a data-driven method to automatically score virtually drilled bones based on expert evaluation of these types of subjective metrics.

2 Related Work

Much of the existing literature on automatic scoring systems for surgical simulation deals with hand-motion analysis. This type of analysis usually incorporates hidden Markov models to classify expert from non-expert surgical performances. Murphy [14], Cotin et al. [5] and Megali et al. [13] have demonstrated that the use of hand motion analysis to describe and score surgical performances is a useful method of evaluating simulation performance. However, there are limits to hand motion analysis. Porte et al. [16] have shown that expert

feedback is better than motion economy feedback for one month retention of suturing skills. While both methods are shown to give improvement in post-test validity, they conclude that long-term skills are better learned by feedback that is more salient than through simple hand-motion analysis. Therefore, it is reasonable to integrate other types of automated feedback into evaluation systems along with hand-motion-based scores.

The most comprehensive work specifically on algorithmic analysis of mastoidectomy simulation performances to date is by Sewell et al. [19] The broad goals in their work are the same as in ours – to develop metrics to score mastoidectomy performances. In their work, they developed methods to distinguish between experts and novices using a simulation system. In this work, we use automated analysis of the final product of a virtual mastoidectomy to duplicate the same results as an expert analysis of the same final product. This type of analysis is important if simulation systems are to be used in the certification process. Repeatability and reliability is a key goal of assessment by simulation systems, and these properties are critical for integration into certification exams [20].

Part of Sewell's analysis included using mutual information to make a classifier that chooses between expert and novice performances based on the 1000 most informative voxels to make that decision in a training set. He also uses the sum of voxels that had at least 0.8 probability of expert removal that were not removed by the user and voxels that had a 0.2 or less probability of expert removal that were removed by the user. This type of analysis is similar to what we propose with the *added* and *removed* functions as described in Sec. 3.3.

Rosen et al. [18] used hidden Markov model analysis for binary classification of novice and expert surgeons performing a laparoscopy procedure. By defining motion signatures through both direct measurement of the forces on the tool and through video analysis, they were able to achieve a accuracy of 0.875. Rosen et al. [17] use Markov models to analyze knot tying performance with a more advanced laproscopic system (Blue DRAGON) which is able to record 27 channels of data during the procedure. Mackel et al. [12] used a similar framework to classify the users of a physical pelvic examination simulator with accuracy of 0.927. Cristancho et al. [6] have developed a technique using video recording and tool position to score laproscopic skills. Ahmidi et al. [1] used a combination of eye tracking and tool position to classify between expert and novice surgeons in endoscopic sinus surgery with an accuracy of 0.825.

An important difference between our work and the previous work on hand motion-based analysis of surgical procedure is that we use final product analysis – the end result of the surgery – rather than procedural analysis to evaluate the surgery. In mastoidectomy procedures, review of a novice during training is commonly performed by examining the end product of a training task on a cadaver specimen in a dissection lab. Our work uses existing surgical metrics that are currently used in training. We describe a voxel-based algorithm to evaluate a portion of those metrics using an automatic system.

We use earth mover's distance (EMD) as a metric when comparing parts of an expert-drilled bone with a student-drilled bone. The EMD has been used to great effect in other regions of image and volume analysis, especially with histogram comparisons. In work by Janoos et al. [8], an approximation of the EMD is used to cluster brain activity data as recorded by fMRI. Sun and Lei [21] outline a method to classify images acquired using optical coherence tomography that uses EMD as a processing step on features before using a classification algorithm. EMD is a flexible metric and has been used in speaker identification [10] as well as in audio searching [24].

3 Methods

3.1 Data acquisition

Under The Ohio State University Institutional Review Board (IRB) approval, as well as approval by each individual institution's IRB, residents in otolaryngology with a wide range of experience in performing mastoidectomies were asked to use a surgical simulator developed by our team. After performing a non-surgical virtual drilling task to give them a feel for the simulation environment, we asked residents to perform a complete mastoidectomy on bone *9413R*. Forty subjects drilled the bone twice (before and after training), therefore 80 different versions of this bone were available for analysis. Although our simulator records the drilling motions used during the procedure, we consider only the final product of the drilling for this study.

In order to determine meaningful differences between the users, we first create a fully partitioned volume that reflects anatomical distinctions between the regions that are relevant to surgery. The volume *9413R* was hand segmented by experts into 37 segments that have relevance to mastoid surgery. Most of the voxels containing bone were not assigned a segment. The result of this segmentation is shown in Fig. 2. However, we require all voxels to be assigned to a region instead of only the voxels that are part of critical structures. Since the drilling performed during the mastoidectomy is in regions that are in close proximity to the important anatomical structures that have been segmented, an intermediate goal is to completely partition the volume. We could divide the volume into blocks along a regular grid, but these blocks lack a relationship to the anatomy and would be clinically meaningless.

In order to completely divide the bone into meaningful regions, we use a voxelized Voronoi tessellation of the segmented volume. We want to give all voxels an id number that corresponds to the nearest segmented structure. In other words, for all voxels v in the volume, we find the nearest segmented voxel p to v . We then assign v to the Voronoi cell associated with the segment id of p : $v \in \text{cell}(\text{id}(p))$. In contrast with the normal Voronoi tessellation, we have many voxels sharing the same value for p .

Distance fields are employed to find the final tessellation. For each segment, we calculate a distance field that contains the distance to the nearest segment. Then, for each voxel, we then select the distance that is lowest out of all the segment distances. This algorithm assigns segment Voronoi cells to all voxels in the volume. Since we are not interested in voxels that are not bone, the id volume is then masked by the intensity values of the original bone volume. All voxels that are not associated with bone are assigned an id of 0. In this way, all voxels are assigned an id on proximity to anatomical segments. The result of this process is shown in Fig. 2.

This segmentation of the volume relies on a previous segmentation by an expert of the voxels assigned to each structure. These expert segmentations are common, but do not give a structure id for every voxel. Our technique does not explicitly take into account anatomical boundaries, but it finds a full tessellation of the bone based on key anatomical areas or subregions of surgical significance.

3.2 Expert data

Based on a survey performed on members of the American Neurotology Society by Wan et al. [22], we identified criteria considered important to a correct mastoidectomy procedure that were not captured directly by existing scoring mechanisms. We asked an expert in otolaryngology to visually rate the results of the virtual mastoidectomy procedures on a scale of one to five (with one being 'poor' and five being 'good') on five separate criteria:

1. Antrum entered
2. Posterior canal wall thinned
3. Appropriate depth of cavity
4. Complete saucerization
5. Overall performance

We selected the above metrics because they can be difficult to quantify and they are more readily analyzed by final product analysis. Other metrics such as “maintains burr visibility” or “does not penetrate facial nerve” are important factors to measure performance as well, and have been examined extensively by Sewell et al. [19] Those metrics can be determined from simulation data fairly easily and of course should be incorporated into any system that grades mastoidectomy procedures completely. However, our work focuses on a subset of important criteria that should be incorporated with other metrics for a complete final score. The metrics considered here, especially *complete saucerization*, are considered important but constructing an algorithmic test for them is quite difficult. The inclusion of the *overall performance* metric in our study is for reference only, and we recommend the addition of burr visibility and violation of critical structures to be included in a final product analysis, as well as incorporating hand-motion analysis scores.

A surgeon was asked to provide us with final products of mastoidectomies performed on 9413R. We used four examples that the surgeon considered his best, after familiarizing himself with the simulator. From these four example volumes, we constructed three composite volumes: minimum, maximum and mean. The minimum contains the voxels that were removed by all expert examples while the maximum removes any voxels from the original dataset that were removed by any expert example. The mean simply takes the per-voxel mean of all the expert examples.

3.3 Extracting distance features

Our goal is to construct an algorithm that provides a score for a resident bone based on the expert examples. An important step is extracting a set of features out of the millions of voxels in the volume. Using a set of features rather than raw voxels, we can use a machine learning algorithm to compute a classification based on the features. Sewell used a set of “most significant voxels” in his classification. In our analysis, we use four distance measures between the previously calculated segment regions as our feature set.

The distance measures fall into two categories: euclidean and earth mover’s distance. We find two euclidean metrics: voxels removed and voxels added. Both of these can be calculated as a sum of a pair-wise operation between the two voxel sets. The position of the voxels has no bearing on these metrics, besides their occupancy in a region or not. The definitions for these functions are found in Eqs. 1 and 2. S is the resident volume, E is the expert volume and R designates a subset of the volume. occ (Eq. 3) is a binary function that determines if the voxel is in the selected region, behaving like a mask. These two functions are found for all 37 values of R , corresponding to each of the segmented regions. An efficient algorithm can calculate both of these metrics simultaneously in $O(n)$ time, where n is the number of voxels in the volume.

$$removed(S, E, R) = \sum occ(i, R) \max(e_i - s_i, 0) \quad (1)$$

$$added(S, E, R) = \sum occ(i, R) \max(s_i - e_i, 0) \quad (2)$$

$$occ(i, R) = \begin{cases} 1 & \text{if } i \in R \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The *removed* function (Eq. 1) describes the number of voxels that have been removed in the resident's drilled volume but not in the expert's drilled volume (i.e., excess drilling by the resident). Conversely, the *added* function (Eq. 2) describes the number of voxels that have been removed in the expert's drilled volume but not in the resident's drilled volume (i.e., not enough drilling by the resident).

The second category of distance measures consists of the earth mover's distance (EMD) and a signed version of the EMD. We use a fast implementation of EMD described by Pele and Werman [15], for which there is source code publicly available. The EMD was originally designed to compare probability distributions, but in this case we apply it directly to the amount of bone in the volume. This measure can be thought of as the total work that needs to be done to change one volume into another, by moving the voxels. In our case, it is better to think of the work in moving the drilling operation from one place to another in the bone. The work for moving the drilling operation from one voxel to another to match the expert volume is based on the distance between them, called the ground distance in the EMD algorithm. In our case, the ground distance used is the Euclidean distance between the two voxels.

Although the earth mover's distance is normally used for histogram comparison, it has some properties, both intuitively and mathematically, that make it a good candidate for volume comparison features. The EMD between two distributions increases as the work needed to change one distribution to the other increases as well. The work in our case is the amount of drilling, since this is the only operation that the users can perform on the bone. There are two types of drilling work when comparing a resident's performance to the expert: drilling that *should have* been done, but was not, and drilling that *should not have* been done, but was.

The EMD algorithm finds the minimum cost to transport material in voxels in the final resident volume from places that should have been drilled to places that were drilled and should not have been. Any remaining voxels that have a discrepancy between the resident and the expert bone are added as a penalty to the final cost. The idea behind using the EMD is that the choice of drilling or not drilling performed by the resident is a locally bounded decision: if extra drilling occurs in a spot that is close to a place where experts drilled, then the penalty should be lower than when extra drilling occurs far away from expert drilling. Although real bone cannot be moved from one place to another in a cost based manner that the EMD is based on, this cost is an abstraction for the magnitude of the error when drilling in an incorrect spot. Like its use in histogram comparison, the EMD captures a quality of similarity that direct pixel-wise vector euclidean distance does not. This idea is valid for metrics that deal with shape of the drilled cavity.

Computation of the EMD is expensive. The thresholded version of the EMD that we use [15] has a computational complexity of $O(N^2 U \log N)$ where U depends on the threshold value used. Pele and Werman's experiments were on 2D images containing 1538, three channel pixels, and their search on 773 images took around 6 seconds. Our dataset has many fewer 3D images, but each image has many voxels. Some of the partitions have over 50,000 voxels, even after removing identical voxels in the two volumes to be compared. Due to the

complexity, calculating the EMD completely on these partitions is not practical. We do not want users of our automated assessment tool to wait days to find out they did something wrong in the simulation.

To improve the performance of the algorithm, we subdivide our segment-based partitions of the volume into clusters of around 5000 voxels. The clusters are determined by k -means clustering, giving k a value of $\lfloor N/5000 \rfloor + 1$. Each cluster comparison takes around 15 seconds to compute.

We also compute a signed EMD (sEMD) value. If the total mass of the expert bone is less than the total mass of the resident bone, $sEMD = EMD$, if not, $sEMD = -EMD$. This is obviously not a metric, but this quantity does reflect the asymmetry between expert and resident. In this model, we are not computing distances between different resident bones or different expert bones, just between experts and residents. The sEMD measure captures the distinction between too much removal and too little removal.

3.4 Determining appropriate classifications

Using the previously described distance functions, we generate a feature vector. The vector has values from all four distance measures on all 37 segments. However, we eliminate any measure in the vectors that has no variance across all 80 samples. Doing this results in a feature vector that has a length around 50 for each sample for our data. With different input volumes, there might be a greater or fewer number of elements with zero variance, leading to a different length of the feature vector. From the feature vector and the expert scores for each bone, we can use machine learning techniques to determine a decision method that converts an arbitrary feature vector (as from a new resident performance) to a score for each of the five scored measures.

Initial attempts at classification using simple linear regression were not promising. A decision tree approach delivered much better results. Unbiased recursive partitioning [7] was used to construct the trees. This algorithm only splits groups of data elements into different nodes if the split has a p -value less than a minimum threshold. For our purposes we considered a value of $p < 0.05$ to be sufficient, although many of the splits had a value of $p < 0.01$.

The computed decision tree is applied like a filter to the user bone drilling performances, based on the values of the feature vector. One of the trees is shown in Figure 3. The top of the diagram shows the decision questions. The first question asked is “Is the value of the feature mean Posterior Canal Wall added greater than 45075”. If so, then the bone drilling result is filtered to Node 5. This is a terminal node, so a score is assigned; in this case the score is 2. The score assigned for this and each terminal node is determined by the score that has the plurality out of all the expert scores that were assigned to that bin. A histogram of the expert scores assigned to that bin can be seen on the bottom of the figure. There are some expert scores that are 1 and some that are 3, but most are 2, so this is the assigned score for this bin. If, however, the answer to the first question is no, then the tree algorithm goes to Node 2 and another question is asked, continuing down the tree. In this way, all feature vectors are assigned scores. Example decision trees computed from the composite final scoring method are shown in Figure 5.

Most of the trees had only enough information to classify the bones into three separate categories. This is due to the lack of examples for some of the bins. For example, only one bone was given a score of 1 by the human reviewer for *antrum entered* and only four bones were given a score of 4 for *overall score*. It is likely that with more examples these

categories would be better represented, and a decision tree could be computed that outputs the full range of values.

3.5 Evaluation

Two approaches were used to calculate final scores. In the first approach, feature vectors were constructed using the four distance measures between each resident bone and the minimum, maximum, and mean expert bone, as described in Sec. 3.1. We call this the *composite method*, since composite volumes were made from the expert examples. These feature vectors were used to optimize a decision tree. In the second approach, feature vectors were constructed for comparison between each resident bone and each expert bone. Decision trees were then constructed for each expert comparison, and a form of voting was used to determine the final score. We investigated two methods for the voting for the ordinal scoring. One takes the mean of all expert sub-scores for the final score (the *mean method*), while the other one uses the median of all expert sub-scores (the *median method*). For larger numbers of expert sub-scores, other types of voting algorithms other than these two may be appropriate. In the case of the binary classification, we use only a majority test, rounding up for ties.

We use two statistical approaches to evaluate the quality of our ordinal automated assessment scores: correlation and inter-rater reliability. The correlation method used is Spearman's rank correlation coefficient, which is a measure of monotonic association between two variables. For inter-rater reliability, we use Cohen's kappa, which is the most common method of determining reliability ratings between two sets of graders. For the binary classification task, we computed the accuracy measure, which is the percentage of correctly classified items. Table 1 shows the quality assessment scores. These scores were generated using leave-one-out cross-validation.

We use a sunflower plot [4] to demonstrate the correlation between the subjective scores determined by the expert and the computed scores in Figs. 4 and 6. In this type of plot, more petals represent more items assigned to a location on the plot.

4 Results

We evaluated several ways to score the virtual surgical performances using our framework, various scoring scales and calculation methods. We employed two scoring scales, a four-rank ordinal score and a binary classification. The results of these methods can be seen in Table 1. We collected data on a five-rank ordinal scale but for all metrics except for *overall score*, we had less than 0.07 of the data values fall into the 1 category. The data for categories 1 and 2 were merged for these metrics. For overall score, we had the same problem with categories 4 and 5 and these were merged. We also performed binary classification, by merging the categories further: the two higher categories merged into one as well as the two lower categories. This binary classification task is common in the literature on automatic evaluation of surgical simulator performance, while a ordinal classification is less common.

A statistical comparison between the original expert subjective scores and the computed scores shows validity for our approach. The range of inter-rater reliability found for the Welling scale [3], is 0.49–0.64. The *complete saucerization* and *antrum entered* metrics achieve scores in this range. The *overall score* metric falls in this range with the median voting method. The other metrics are under this range, with *depth of cavity* in the composite scoring method falling well below. Correlation scores, determined by Spearman's method, are moderate, with *complete saucerization* as well as *antrum entered* again being quite strongly correlated, but *depth of cavity* being comparatively weak.

The median and mean methods gave much better results than the composite method does. These methods are slower to compute than the composite method. Using the median method (or the mean method), one decision tree must be followed for each expert bone. Using the composite method, only one decision tree is used. The preprocessing time for the voting methods is longer than it is for the composite method when the number of expert bones is four or more, as it is in our tests.

The results for the binary classification task are shown in the right two columns in Table 1. For this task, the expert scores were divided into two groups, expert and non-expert. Values of four and above were counted as an expert performance; three and lower were considered non-expert. The decision tree method was then applied. The fraction of correct answers are reported as accuracy (ACC). Similarly to the ordinal results, the majority voting method achieved higher accuracy than did the composite method, with the majority method reaching 81%–89% accuracy for the individual metrics and 75% accuracy for overall score.

In Fig. 4, a two-dimensional histogram comparing the computed score and the expert-given subjective score for all the trainee-drilled bones for the metric of *complete saucerization* using the composite scoring method. The items are concentrated along the diagonal, which indicates a high coincidence of the subjective scores with the automated scores. Fig. 6 shows the remaining plots for the composite scoring method; the plots for the mean and median scoring method are not dramatically different. Not all the metric categories are represented by the automated scores, since there was not enough data gathered for those bins to make significant decisions, as explained in Sec. 3.4.

5 Discussion

5.1 General comments

Many of the criteria seem intuitively likely due to the anatomical basis of the metric. The *posterior canal wall* metric decision tree example shown in Fig. 3 is automatically calculated based on the *added* function (Eq. 2) applied to the posterior canal wall region as well as the facial canal nerve region, which is adjacent. Likewise, the *antrum entered* automated metric depends only on the voxels of the mastoid antrum region. Some of the metrics depend on computed regional EMD values, while other metrics use only the simpler Euclidean distances for classification.

However, because of the incomplete diversity of the data that we have obtained through the study, the decision trees we generate can have counter-intuitive results. In Fig. 5, for the *overall score* metric, the tree gives a higher score to bones that have a higher value for mean Facial Canal Nerve added. Intuitively, more added voxels should result in a lower score, not a higher score. The result here is due to factors in the data that are correlated to performance but are not causal. Many of the bones that our expert scored as a 2 overall are from students that did not drill away enough bone. Many of the bones scored as a 1 were over-drilled. Because of this, the algorithm constructed a decision tree to classify these categories using this difference. With a larger and more diverse set of examples, issues like this will be reduced.

Most of the decision trees had only enough information to classify the bones into three separate categories. This is due to the lack of examples for some of the bins. For example, only one bone was given a score of 1 by the human reviewer for *antrum entered* and only four bones were given a score of 4 for *overall score*. It is likely that with more examples these categories would be better represented, and a decision tree could be computed that outputs the full range of values.

Problems with the *depth of cavity* metrics could be due to the lack of stereo vision in the test environment used to gather these resident bones. Due to hardware limitations at the time, we were not able to deploy our systems with a 3D stereo display device. Some users complained about difficulties in determining depth during drilling. Even though the analysis described in this article is theoretically independent of the quality of the simulator, it is influenced by the training set, and for this metric the training set might not have been sufficient. With the use of 3D stereo in the next revision of our simulation system, we will see if correlation will improve for this metric.

With a tool to give automated assessment on difficult to define metrics such as *complete saucerization*, we can develop simulation environments that give feedback to residents during the early stages of their training. Although expert assessment is still needed during the course of study, these types of tools, along with hand-motion analysis, could accelerate training for formative development. More studies are required to determine the correlation of the use of automated assessment tools as a part of simulation use in a surgical curriculum to actual performance and outcomes on patients. This article presents evidence that an algorithm can be used to assess shape-based results from bone drilling procedures, but refinements will no doubt be needed as new evidence and more data comes in from larger-scale testing.

5.2 Limitations

A limitation to this technique is that it must be done for each dataset. We have constructed decision trees for bone *9143R*. We can apply the same procedure to other bones. However, we will require both representative expert final products from a mastoidectomy procedure and expert grading of sample resident bones. This takes about four hours total: two hours for drilling four mastoid bones at around 30 minutes per bone and around two hours to grade 80 bones, since it takes between one and two minutes per bone (on average) to grade, based on the time taken by our expert. The more time-consuming part is the acquisition of the 80 resident-drilled bones. These do not have to be drilled all by different residents, as it was in our case, but a wide enough variance in performance is important to avoid over-fitting of the data. Once this work is performed, automatic grading of each bone can be done without any more user interaction.

We performed an analysis using the same feature vectors and decision tree approach as in Sec. 3 but targeting violation based metrics, but this was unsuccessful. An expert graded the bone on violations of the tegmen, sigmoid sinus, facial nerve, and lateral canal. Most of the resultant decision trees had only one node, which means that there was not enough information in the feature vectors to justify splitting the dataset at a low enough p -value. The metrics that we used originally are shape-based metrics, and these can be analyzed well by the distance measures that we have chosen, on regional partitions. However, violation based metrics are more accurately based on exact voxel analysis based on the strict segmentation boundaries provided by the experts.

The k -means clustering necessary for efficient implementation of the EMD algorithm, so that it completes in a reasonable time, does add an artificial separation between regions that is undesirable. Although a k -means-based partitioning approach appears to be more desirable than a rectangular tessellation of the structure, the effects of this partitioning on the classification outcome have not been tested. In addition, more anatomical structures can be defined. The inclusion or exclusion of structures from the complete partitioning of the bone will affect the grading performance, but it is not clear if the inclusion of more structures always results in better performance.

6 Conclusion

To our knowledge, the use of EMD for the comparison at the voxel level of virtual surgical volumes has not been attempted before and deserves further study. Furthermore, our techniques provide automatic scores on shape-based performance metrics that can be difficult to quantify in other ways rather than economy of motion metrics that are not procedure specific. The use of a simulation system to obtain multiple expert and trainee performances from a single original dataset removes any noise in the system due to inexact registration between multiple anatomical specimens. This gives more confidence and objectivity in the resulting scores. However, further efforts include the creation of a complete objective scoring system for mastoidectomy simulation. We wish to include these metrics in an assessment module for our simulation system, and plan to do this in the next large-scale test of the system.

We have demonstrated a method of performing automatic scoring for a mastoidectomy simulator. Using decision trees and feature vectors generated from several distance measures, ratings on a multi-level scale can be given to users of a simulation system without action from an expert.

Acknowledgments

This work is supported by a grant from the National Institute of Deafness and Other Communication Disorders, of the National Institutes of Health, 1 R01 DC06458-01A1.

References

1. Ahmidi, N.; Hager, GD.; Ishii, L.; Fichtinger, G.; Gallia, GL.; Ishii, M. Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. In: Jiang, T.; Navab, N.; Pluim, JP.; Viergever, MA., editors. MICCAI. 2010. p. 295-302.
2. Bryan J, Stredney D, Wiet G, Sessanna D. Virtual temporal bone dissection: a case study. IEEE Visualization. 2001:497–500.
3. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. The Laryngoscope. 2007; 117(10):1803–8.10.1097/MLG.0b013e31811edd7a [PubMed: 17721407]
4. Cleveland WS, McGill R. The Many Faces of a Scatterplot. Journal of the American Statistical Association. 1984; 79(388):807– 822.
5. Cotin S, Stylopoulos N, Ottensmeyer MP, Neumann PF, Rattner D, Dawson S. Metrics for Laparoscopic Skills Trainers: The Weakest Link! MICCAI. 2002:35–43.
6. Cristancho SM, Hodgson AJ, Panton ONM, Meneghetti A, Warnock G, Qayumi K. Intraoperative monitoring of laparoscopic skill development based on quantitative measures. Surgical endoscopy. 2009; 23(10):2181–90.10.1007/s00464-008-0246-9 [PubMed: 19116747]
7. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006; 15(3):651–674.10.1198/106186006X133933
8. Janoos, F.; Machiraju, R.; Sammet, S.; Knopp, M.; Mórocz, I. Un-supervised Learning of Brain States from fMRI Data. In: Jiang, T.; Navab, N.; Pluim, J.; Viergever, M., editors. MICCAI, Lecture Notes in Computer Science. Vol. 6362. Springer; Berlin/Heidelberg: 2010. p. 201-208.
9. Kerwin T, Shen HW, Stredney D. Enhancing realism of wet surfaces in temporal bone surgical simulation. IEEE Transactions on Visualization and Computer Graphics. 2009; 15(5):747–758.10.1109/TVCG.2009.31 [PubMed: 19590102]
10. Kuroiwa S, Umeda Y, Tsuge S, Ren F. Nonparametric Speaker Recognition Method Using Earth Mover's Distance. IE-ICE - Transactions on Information and Systems. 2006:1074–1081.

11. Laeeq K, Bhatti NI, Carey JP, Della Santina CC, Limb CJ, Niparko JK, Minor LB, Francis HW. Pilot testing of an assessment tool for competency in mastoidectomy. *The Laryngoscope*. 2009; 119(12):2402–10.10.1002/lary.20678 [PubMed: 19885831]
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. *MMVR*. 2006; 119:355–60.
13. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. *IEEE Transactions on Biomedical Engineering*. 2006; 53(10):1911–9.10.1109/TBME.2006.881784 [PubMed: 17019854]
14. Murphy, TE. Master's thesis. Johns Hopkins University; 2004. Towards Objective Surgical Skill Evaluation with Hidden Markov Model-based Motion Recognition.
15. Pele, O.; Werman, M. Fast and Robust Earth Mover's Distances. *International Conference on Computer Vision*; Kyoto, Japan. 2009.
16. Porte MC, Xeroulis G, Reznick RK, Dubrowski A. Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. *American journal of surgery*. 2007; 193(1):105–10.10. 1016/j.amjsurg.2006.03.016 [PubMed: 17188099]
17. Rosen J, Brown JD, Chang L, Sinanan MN, Hannaford B. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. *IEEE Transactions on Biomedical Engineering*. 2006; 53(3):399–413.10.1109/TBME.2005.869771 [PubMed: 16532766]
18. 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 Transactions on Biomedical Engineering*. 2001; 48(5):579–91.10.1109/10.918597 [PubMed: 11341532]
19. Sewell C, Morris D, Blevins NH, Dutta S, Agrawal S, Barbagli F, Salisbury K. Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery*. 2008; 13(2):63–81.10.1080/10929080801957712 [PubMed: 18317956]
20. Shaffer DW, Gordon J, Bennett N. Learning, Testing, and the Evaluation of Learning Environments in Medicine: Global Performance Assessment in Medical Education. *Interactive Learning Environments*. 2004; 12(3):167–178.10.1080/10494820512331383409
21. Sun Y, Lei M. Method for optical coherence tomography image classification using local features and earth mover's distance. *Journal of Biomedical Optics*. 2009; 14(5):054, 037–6.
22. Wan D, Wiet GJ, Welling DB, Kerwin T, Stredney D. Creating a cross-institutional grading scale for temporal bone dissection. *The Laryngoscope*. 2010; 120(7):1422–7.10.1002/lary.20957 [PubMed: 20564671]
23. Wiet, GJ. Triological Society Thesis. 2010. Virtual Temporal Bone Dissection System: Development and Testing. (Submitted)
24. Yuxin, P.; Cuihua, F.; Xiaou, C. Using Earth Movers Distance for Audio Clip Retrieval. In: Zhuang, Y.; Yang, SQ.; Rui, Y.; He, Q., editors. *Advances in Multimedia Information Processing, Lecture Notes in Computer Science*. Vol. 4261. Springer Berlin Heidelberg; Berlin, Heidelberg: 2006. p. 405-413.

(a) Photo of the simulator in-use

(b) Screenshot of the simulator

Fig. 1.
The surgical simulator used in the study.

Fig. 2.

The regions created from the 3D Voronoi tessellation of the segments in the *9413R* dataset. This is a Right Lateral view of the regional anatomy.

Fig. 3.

The resultant decision tree for *Posterior canal wall thinned* based on the composite feature vectors.

Fig. 4.

2D histogram petal plot for complete saucerization scores using the composite method. The number of petals equal the number of items in that particular bin. A single dot represents one item. Most of the items fall along the diagonal, which means that the item's computed score and expert-given subjective score are equal. More petal plots of the composite method can be found in Fig. 6.

Fig. 5.

The decision trees determined from the data to give the best division between classes.

Fig. 6.
The remainder of petal plots of the scores from the composite score method. A continuation of Fig. 4.

For the ordinal classification task, the the inter-rater reliability (Cohen's kappa) and correlation (Spearman's rho) are given for each metric. The mean and median columns show the scores from computing a final score from the four separate expert scores. The composite column shows the results from the evaluator trained using the three composite datasets. For the binary classification task, overall accuracy (ACC) is given for the majority voting and composite scoring methods. Please see Sec. 4 for more detail.

Table 1

	Ordinal Classification				Binary Classification			
	Mean		Median		Composite		Majority	
	κ	ρ	κ	ρ	κ	ρ	ACC	ACC
Complete saucerization	0.51	0.85	0.54	0.85	0.61	0.79	0.83	0.80
Antrum entered	0.56	0.84	0.46	0.75	0.32	0.69	0.89	0.76
Depth of cavity	0.36	0.70	0.37	0.68	0.13	0.60	0.85	0.63
Posterior canal wall thinned	0.31	0.75	0.38	0.76	0.31	0.71	0.81	0.80
Overall score	0.47	0.80	0.50	0.66	0.32	0.46	0.75	0.45