Fuzzy Classification: Towards Evaluating Performance on a Surgical Simulator

Jeff HUANG, Shahram PAYANDEH (*PhD*), Peter DORIS (*MD*), and
Ima HAJSHIRMOHAMMADI (*BSc*)
Simon Fraser University and Surrey Memorial Hospital
Vancouver, CANADA

Abstract. Computer-based surgical simulators such as the MIST-VR (www.mentice.com) are able to provide scoring metrics such as time taken to complete a task, number of errors made, and economy of movement. Using MIST-VR's basic metrics, we explored the possibility of classifying skill levels using fuzzy logic. Our objective was to create a fuzzy classifier capable of classifying the performance of a subject training on a surgical simulator into 1 of 3 categories: Novice, Intermediate, and Expert. To accomplish this, we needed to establish a baseline skill level for each category. We had four laparoscopic surgeons, four surgical assistants/residents and four non-surgical staff/students with no laparoscopic experience perform two basic tasks on the simulator involving the placement of a ball into a box. We have found, through this preliminary study, that the results were inconclusive. We suspected a number of issues such as the size of our sample space used to train our classifier, and the difficulty of the chosen tasks adversely affected our results.

1. Introduction

Laparoscopy, despite being beneficial to the patient when done properly, is more difficult than open surgery. The surgeon is working in a relatively confined space and has only a two-dimensional view of the patient's anatomy through a video screen, rather than a direct three-dimensional view. This increases the chance of complications. The introduction of laparoscopic cholecystectomy (LC) was associated with an increased number of complications; more precisely, there were three times as many complications than in traditional cholecystectomy [8]. More alarming is that the number of injuries does not seem to be decreasing with time [8]. Olsen has found that these injuries are still being reported 6 years after the introduction of LC [5]. He found that the majority of injuries were the result of misidentification of anatomy, which is preventable. The need to reduce the number of injuries related to laparoscopic procedures has lead to a review of how surgeons are trained.

All surgeons are taught laparoscopic techniques through the use of tools such as mechanical box trainers, inanimate models, animals, and surgical simulators. In residency programs, trainees are evaluated using In Training Evaluation Reports (ITERs) over a number of years. ITERs are completed at periodic intervals during training. These reports consist of oral and written examination, and comments from the overseeing surgeon; however, they do not include an objective assessment of technical skill, preventing a standard from being established [1]. Without a standard, newly certified surgeons may emerge without the necessary level of surgical proficiency.

One part to a complete solution would be enhancing the VR simulators to evaluate the trainee. The evaluation would be objective in nature, as the metrics are based on actual performance and not perception. Current simulators can only provide a few basic metrics such as time taken to complete a procedure, the number of errors made, and the efficiency of movements. We want to extend these metrics to the point where we can determine where the subject's skills lie in comparison to those in the general laparoscopic surgical community. These simulators would allow more practice time for surgeons, and advise them of their current skill level.

Fuzzy logic can be applied to achieve our goals. Operative procedures can be divided into a number of tasks, and each task can be evaluated according to criteria set out by an expert surgeon. Ota outlines how the task of dissecting a blood vessel can be evaluated by fuzzy logic [4].

Hence, this paper explores the possibility of extending the performance evaluations offered by a surgical simulator with the ability to compare the trainee's performance with general levels of skill. We propose the use of a fuzzy-based if-then rule system known as a fuzzy inference system (FIS) [3], or more specifically a fuzzy classifier, to promote objective assessment. A FIS is a system that maps an input to an output using fuzzy logic. Neural networks can be used to train and generate rules from the input data. These rules are the foundation of the fuzzy classifier.

Fuzzy classification has many applications and has been used in various automated medical diagnostic areas including breast cancer tumour diagnosis and prognosis by Hoffman [1]. Hoffman constructed a classifier using a breast cancer data set containing features such as the texture, smoothness and radius of cell nuclei from a breast lump. Each instance in the data was classified as malignant or benign, according to actual, known diagnosis of the patient. Using 15 if-then rules, a 95% correct classification resulted for the testing data set.

For our surgical simulator application, a fuzzy classifier can be constructed to evaluate a trainee's skill level based on their performance on surgical tasks in a VR trainer. This particular classifier evaluates the trainee in tasks mainly involved with the dominant hand.

2. Method

The general approach is to conduct a user study to get performance metrics from people of different skill levels. The data from the study will then be used to build our classifier.

2.1 Experimental Setup

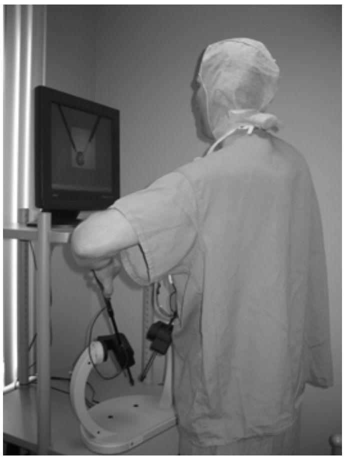

Figure 1: The Experimental Setup.

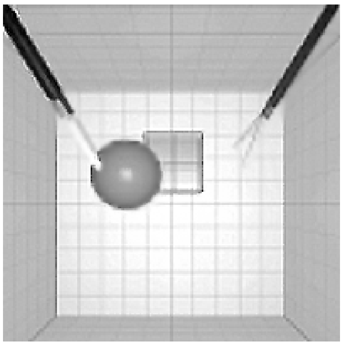

Figure 2: An image of AcquirePlace & TransferPlace

In order to create a fuzzy classifier, we conducted a user study at Surrey Memorial Hospital in B.C., Canada. Our setup (see Figure 1) consisted of an isolated room with a 19" eye-level LCD monitor, a two-handed laparoscopic device with needle-driver handles (Virtual Laparoscopic Interface, by Immersion Inc.), and a dual Intel Xeon 2.8Ghz computer.

Table 1: This is an example of the AcquirePlace metrics. It shows the performance for each hand for one trial.

	Scores	
	Left Hand	**Right Hand**
Time	33.2	22.0
Errors	29.0	27.0
Economy of Movement	2.2	2.6
Total Score	117.8	

The simulator used for our study was the Minimally Invasive Surgical Trainer – Virtual Reality (MIST-VR) by Mentice AB. This simulator has been shown to transfer skills from virtual reality into the operating room [7]. MIST-VR offers a number of tasks, combined within two modules: Coreskills and Suturing. The Coreskills module focuses on familiarizing the trainee with basic laparoscopic skills. It includes tasks focusing on target manipulation and placement, transferring objects between instruments, and diathermy. The Suturing module trains the surgeon to manipulate a needle and perform suturing. MIST-VR keeps track of the user's performance through a number of metrics. These metrics will be explained in the next paragraph.

The tasks used in our study were AcquirePlace and TransferPlace. Both tasks involved the manipulation of a ball with grippers. AcquirePlace consist of grasping the ball with a gripper, manipulating the ball into a 3-dimensional cube, and releasing it within the cube. TransferPlace is similar to AcquirePlace and consists of grasping the ball with one gripper, transferring the ball to the other gripper, manipulating the ball into a 3-dimensional cube, and releasing it within the cube. Both of these tasks use the following metrics: the time taken to complete a task, the number of errors made, the economy of the user's movements, and the user's overall score. Some possible errors are touching tools together, dropping the ball, and removing the ball from the cube without releasing it. The economy of movement is the ratio of actual tool movements during the task to the optimal calculated movement of the tools necessary to complete the task. These ratios are set by Mentice, and are unknown to us. The best results are the ones that approach the value of 1. The overall score is the sum of all the other metrics multiplied by their respective weights, which are set by the administrator of the simulator. For this study, all weights were set to 1. Hence, the overall score was simply the sum of all other metrics for the task. A sample of the scoring for a trial from the task AcquirePlace is shown in Table 1.

2.2 Experimental Design

Table 2: This is a detailed breakdown of the test groups.

	Group A	**Group B**
First Task	AcquirePlace	TransferPlace
Second Task	TransferPlace	AcquirePlace
# of Experts	2	2
# of Intermediates	2	2
# of Novices	2	2

Three groups of people were selected based on the following definitions: novices (little or no laparoscopic experience), intermediate (1-2 yrs of laparoscopic experience) and experts (people with 2+ years of laparoscopic experience). Our novices consisted of 1 SFU student and 3 operating room nurses; our intermediate group consisted of 4 surgical residents/assistants; and our experts consisted of 4 laparoscopic surgeons. We split our subjects into two groups in order to vary the task order. Group A did AcquirePlace first, then TransferPlace while group B did TransferPlace first, and then AcquirePlace. Two subjects from each classification were placed in each group. The group compositions are outlined in Table 2.

For each task, the presenter introduced the subject to the task, played a video of an error-free trial of the task, and explained what constituted as errors. The subject was then given two practice trials before performing 4 trials which were counted in the results. Each trial consisted of starting the task with the left hand, and then the right hand. At the end of the task, they were allowed to view their performance data.

2.3 Fuzzy Classifier Design

The purpose of the user study was to gather data necessary to create the classifiers for each task. We needed data from subjects that fit our three categories of skill. This data would be split in equal halves in order to form two representative data sets. One would be used for training, and the other would be used for testing the classifier. We created our classifier using Matlab's Fuzzy Toolbox's Adaptive Neuro-Fuzzy Inference System (ANFIS), which uses neural networks in order to train the classifier. ANFIS applies the learning abilities of neural networks[2] to detect trends in the training data. From these trends, ANFIS can generate the fuzzy if-then rules of the classifier. The benefit of using ANFIS is that if we do not have a conceptual view of what our system, membership functions and rules should be, ANFIS can create this for us automatically. The resulting system will be adapted to the data, and accommodate any variations [3].

ANFIS is limited to using Sugeno systems. In order to generate our classifier, we loaded our training data set and used subtractive clustering with the following parameters: *Range of Influence = 1; Squash Factor = 1.25; Accept Ratio = 0.5; and Reject Ratio = 0.15*. The benefit of using subtractive clustering is that it keeps the number of rules relatively low, requiring less computation. We then trained the FIS using the hybrid learning rule[2] for 200 epochs. A testing set is used to evaluate the validity of the trained classifier.

3. Results

3.1 Organizing the Resulting Data

In total, we had 12 participants who came in on a drop-in basis. This yielded 12 full sets of data. We first separated the data from AcquirePlace and TransferPlace. The data for each task was organized into 48 vectors in total (12 subjects who did 4 trials each). Each vector contained the performance metrics for the trial.

Then, the data had to be reorganized to exclude non-dominant hand scores. This involved removing extra data (all the data related to the non-dominant hand in each trial) and recalculating the score for just the dominant hand. Note that MIST calculates the scores using metrics for *both hands*. Since we were only interested in the dominant hand, we recalculated the scores using the metrics for only the dominant hand.

Table 3: Breakdown of our training and testing data sets.

Training Data Set	Testing Data Set
1 Group A Expert	1 Group A Expert
1 Group B Expert	1 Group B Expert
1 Group A Intermediate	1 Group A Intermediate
1 Group B Intermediate	1 Group B Intermediate
1 Group A Novice	1 Group A Novice
1 Group B Novice	1 Group B Novice

Table 4: Samples of vectors in a data set. A class of '1' represents an expert, '2' represents an intermediate, and '3' represents a novice.

Economy of Movement	Time	# of Errors	Score	Class
4	9.4	2	15.4	1
2.4	9.5	0	11.9	2
5.2	19.7	3	27.9	3

Next, we formed our training and testing data sets. Each data set contained 2 subjects per skill level category. For each of the two subjects, one was taken from group A and the other from group B. This is clarified in Table 3. Each subject did 4 trials, so s/he has 4 vectors associated with him/her; hence, each data set had 24 vectors: 8 experts, 8 intermediates, and 8 novices. For each vector, we added an extra tag called Class to indicate the correct classification. A class of ‘1’ represents an Expert, a ‘2’ represents an Intermediate, and a ‘3’ represents a Novice. This is shown in Table 4.

3.2 The Resulting Classifiers

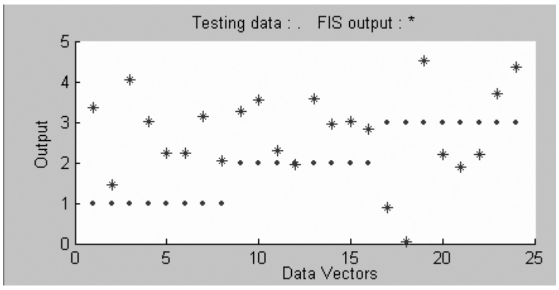

Figure 3: The AcquirePlace fuzzy classifier plot. The output (y-axis) represents the classification of the vector (x-axis).

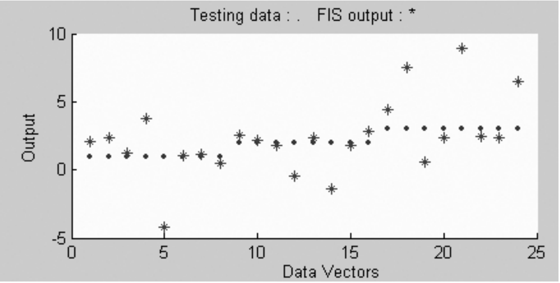

Figure 4: The TransferPlace fuzzy classifier plot. The output (y-axis) represents the classification of the vector (x-axis).

The plots in Figure 3 and Figure 4 show the plots for the classifiers. The •’s in the plots represent our desired output, which is the Class of each vector. The actual output of our classifier is represented by the *’s. Optimal results would be indicated by minimal separation between desired output points and their corresponding actual output points.

4. Discussion

By inspection of the classifier plots in Figure 3 and Figure 4, it is evident that our classifier does not offer the optimal results. In AcquirePlace, only 1 vector (#12) was correctly classified. The remainders were haphazardly classified into categories not defined.

However, TransferPlace offers some positive hopes. At least eight vectors are close enough to be considered classified correctly. Also, most vectors were within the defined classes 1, 2, or 3. This suggests that the classifier is having trouble distinguishing between our categories of skill level. Using more training data may correct this problem.

The underlying hypothesis for this study was that a fuzzy classifier could be used to properly distinguish various users and their levels of expertise. A number of factors influenced the conclusions in this paper: sample size and grouping methods; organization of the training and testing data sets; task selection; task difficulty; and formulation of the fuzzy classifier. Varying these factors in further studies may serve to provide more positive results.

5. Conclusion and Future work

Although the classifiers for the tasks AcquirePlace and TransferPlace were not accurate, there is still more variations and improvements to be made to produce a successful classifier. In our case, we believe a major factor affecting our result is that our sample did not represent the performance variation of our three categories of people. A larger sample size will be necessary for further studies. Using more difficult tasks may also improve chances of a working classifier by providing a larger skill gap between the classifications.

We are conducting further studies to collect data to be used for an improved classifier, focusing on suturing and knotting tasks, which is a difficult skill for laparoscopic practitioners. Tasks such as *StitchStart* and ContinuousSuture have metrics such as the distance between the actual needle hit point and the hidden target point, and a stitch deformation metric will present a larger gap between the categories of skill levels, and present the opportunity to create a working classifier.

Traversing beyond MIST-VR, there are other surgical simulators in progress. SFU is developing the Laparoscopic Training Environment (LTE) [6], in which other metrics can be built into tasks such as suturing. In this case, we could integrate the fuzzy inference system into the software. Also, the LTE supports a haptic-feedback laparoscopic device, which would give the subject a more realistic training experience.

References

[1] Hoffmann F. Boosting a genetic fuzzy classifier, IFSA/NAFIPS 2001, Vancouver, Canada.

[2] Jang JSR, Sun CT. Neuro-fuzzy modeling and control. *The Proceedings of the IEEE*. March 1995; 83; 378-406.

[3] The MathWorks. (2002). *Fuzzy Logic Toolbox User Guide*. The MathWorks, Inc.

[4] Ota D *et al*. Virtual reality in surgical education. Computers in Biology and Medicine. 1995; 25; 2; 127-137.

[5] Olsen D. Bile duct injuries during laparoscopic cholecystectomy. Surgical Endoscopy. 1997; 11; 133-138.

[6] Payandeh S *et al*. LTE: A multi-modal training environment for surgeons, In *Proc. of ACM Fifth International Conference on Multi-Modal Interface*. 2003; 301-302.

[7] Seymour NE *et al.* Virtual reality training improves operating room performance. Annals of Surgery. 2002; 236(4); 458-464.

[8] Walsh RM *et al*. Trends in bile duct injuries from laparoscopic cholecystectomy. 1997 Americas Hepato-Pancreato-Biliary Congress, Miami, Fla.

[9] Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: from selection to certification. Curr Prob Surg 2002; 39; 573-660.