Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery

Carol E. Reiley and Gregory D. Hager

Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
{creiley,hager}@jhu.edu

Abstract. Evaluating surgical skill is a time consuming, subjective, and difficult process. This paper compares two methods of identifying the skill level of a subject given motion data from a benchtop surgical task. In the first method, we build discrete Hidden Markov Models at the *task level*, and test against these models. In the second method, we build discrete Hidden Markov Models of surgical gestures, called *surgemes*, and evaluate skill at this level. We apply these techniques to 57 data sets collected from the da Vinci surgical system. Our current techniques have achieved accuracy levels of 100% using task level models and known gesture segmentation, 95% with task level models and unknown gesture segmentation, and 100% with the surgeme level models in correctly identifying the skill level. We observe that, although less accurate, the second method requires less prior label information. Also, the surgeme level classification provided more insights into what subjects did well, and what they did poorly.

1 Introduction

Human motion is stochastic in nature. A person performing a repeatable task multiple times (e.g. drawing a straight line) would generate different motion measurements (ie. forces, velocities, positions, etc.) despite the fact that the measurements represent the the same task performed with the same level of skill. The goal of skill modeling is to uncover and measure the underlying characteristics of skill hidden in measurable motion data. In this paper, we focus on modeling and assessing surgical technical skill. Current techniques for surgical skill assessment include descriptive statistics (time, path length, number of motions), morbidity rates, and checklists [1,2]. However, these methods require manual interpretation, lack flexibility, and are time consuming and labor intensive. We note that, robotic surgery, in particular, is known to have a steep learning curve and be difficult to teach [3]. However, robotic surgery is reported to be the fastest growing segment of computer aided surgical systems - an industry expected to grow to $2 billion by the year 2010 [4]. Thus, automated assessment and training will have a potentially high impact in this area.

In what follows, we make extensive use of Hidden Markov Models (HMMs) for statistical modeling of time-series motion data as a basis for skill assessment. HMMs are statistical models used to determine hidden parameters from observed data. An HMM can either be continuous or discrete. A discrete HMM, which is used in this work, is represented by $\lambda = (A, B, \Pi)$ where A is the state transition probability distribution matrix, B is the observation symbol probability distribution matrix, and Π is the initial

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part I, LNCS 5761, pp. 435–442, 2009.

state distribution [5]. They are extensively developed in the area of speech recognition [5] and have proved useful in studying teleoperation and human skill evaluation for non-surgical tasks [6,7,8,9]. Recently, HMMs have been used in laparoscopic [10] and virtual simulators [11,12] to classify the skill of a surgeon. Motivated by these studies, we propose building HMMs driven by expertise examples on teleoperated robotic systems.

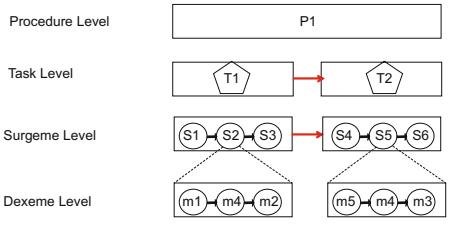


Fig. 1. A hierarchical decomposition of surgical motion. Modeling can be done on tasks (e.g. suturing), their decomposition into surgemes (e.g. needle pulling), and even more primitive motion elements called dexemes.

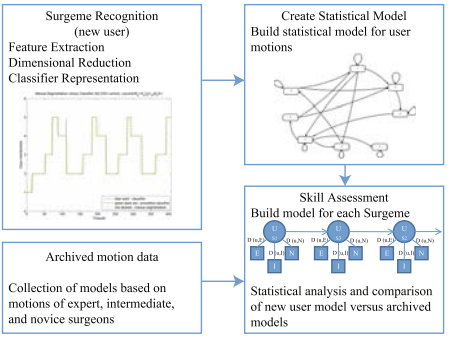


Fig. 2. Flowchart of building skill evaluation models

In prior work [13,14], it has been proposed that surgical tasks can be broken down into a hierarchy of more primitive gestures, sometimes referred to as "surgemes" (Figure 1). To the best of our knowledge, this is the first work that has analyzed surgical models on the surgeme level. Thus, we are evaluating if skill at the task level or at the surgeme level provides more accuracy or information. At either level, we may also explore whether knowing the underlying surgeme information at training and/or testing [10] provides significant additional information. Figure 2 is a general block diagram of building skill models.

In this paper, we develop skill-dependent HMM models for three levels of surgical expertise, novice, intermediate, and expert. We define an expert as a practicing surgeon with more than 100 hours of clinical surgical robotic experience. An intermediate as either a fellow or resident surgeon with less than 100 hours of surgical robotic experience, and a novice as a non-surgeon with no prior robotic experience. As shown in Table 1 this leads to six possible approaches to surgical skill evaluation. We then explore the relationships between three of these problem settings:

Table 1. Possible combinations of labeling information and level of skill evaluation

Training Testing	Labeled Labeled	Labeled Unlabeled	Unlabeled Unlabeled
Surgeme Level	1a	1b	1c
Task Level	2a	2b	2c

Experiment 1: Surgeme-Level Hidden Markov Models (1c - states: unknown dexemes) vs. Task-Level Hidden Markov Models (2c - states: unknown surgemes). We hypothesize that modeling skill on the surgeme level may provide insight to what portions of a task a subject performs proficiently or where he/she performs like a novice.

Experiment 2: Task-Level Hidden Markov Models With Known States (2a - states: motion surgemes defined in Experimental Setup) vs. Task-Level Hidden Markov Models With Unknown States (2c). Here, we investigate whether raw motion data can be modeled and evaluated for skill level *without* prior manual labeling.

2 Experimental Setup

Surgical Platform. The da Vinci surgical system by Intuitive Surgical Inc. [15] is a clinical, teleoperated robot used in operating rooms worldwide. It is a two handed manipulator with 7 degrees of freedom each. Due to its immersive interface, it provides a structured, well instrumented, unobtrusive environment for studying surgical motions. Using the da Vinci application programming interface (API), synchronized high-resolution video and motion data were recorded at 23 Hz and resampled to 40 Hz for data analysis. For the results reported below, we use a 14 variable subset of the available motion channels comprising of six joint velocity values and gripper information of the patient-side left and right robotic manipulators. Fifty seven trials of a four-throw suturing task were collected from a group of nine different surgeons categorized into three different expertise levels; 19 trials each for expert, intermediate, and novice. (Fig. 3).

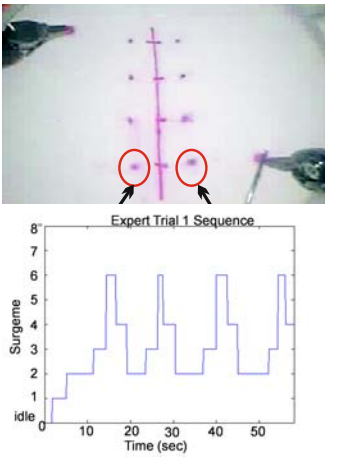


Fig. 3. (top) Experimental setup displaying surgeme 1 (reach for needle) in our motion vocabulary. (bottom) Example of manually labeled trial using surgemes

We make use of the motion vocabulary defined in [13] which consists of (0) idle position, (1) reach for needle, (2) position needle, (3) inset needle through tissue, (4) transferring needle from left to right, (5) moving to center with needle in gripper, (6) pulling suture with left hand, (7) pulling suture with right, and (8) orienting needles. Idle motion time at the start and end of the trial (motion 0) was not used for data analysis. In order to have ground truth for training and validation, our data set was manually segmented based on the above surgeme motions. Not all trials were required to use all surgemes in the motion vocabulary. Trials times varied between 45 to 130 seconds.

3 Task-Level Skill Modeling

The aim of this paper is to create an HMM $\lambda = (A, B, \Pi)$ that describes the surgical performance made by surgeons of various skill levels and to create a metric to evaluate surgical performance.

Data Preprocessing. The 14 vector continuous motion observations were postprocessed into a discrete alphabet using vector quantization techniques similar to [6]. We first apply a Short Time Fourier Transform (STFT) on each of the 14 velocity signals, $x(t)$, over a 400 ms window every 200 ms to filter the high frequency data.

$$STFT\{x(t)\} \equiv X(t,\omega) = \int_{-\infty}^{\infty} x(t')w(t'-t)e^{-j2\pi ft'}\,dt' \tag{1}$$

where $w(d) = 1$ if $0 < d < 200$ms and 0 otherwise. The STFT was chosen because it is able to extract useful features from time series data and, when applied over a sliding window in the time domain, was able to retain the time localization of events from the original data. The amplitudes of the lower 4 STFT coefficients were then concatenated to form a new 56 (14 velocity channels by 4 STFT coefficients) dimensional feature vector. Then we used the K-means algorithm [16] to search for a small number of cluster centers within each of the 9 states. These cluster centers form a so-called *codebook*. We chose a 64 symbol codebook based on empirical testing with values of 32, 64, and 128.

HMM Task Models. From the discretized signals, we trained two families of Hidden Markov Model for each skill level (expert λ_{s1}, intermediate λ_{s2}, and novice λ_{s3}) using the Matlab statistics toolbox [17]. In one family, we trained "true" Hidden Markov Models (2c) for each skill level using the Baum-Welch algorithm and no prior training labels. We chose 9 states to match the number of surgeme labels, but we allowed the generalized Expectation Maximization algorithm to uncover the best structure of the hidden states. For Hidden Markov Models with states as surgemes (2a), we concatenated each observation with its surgeme label. Because the underlying states were known, we did not need to perform Baum-Welch, but rather could directly calculate the transition and emission tables.

Evaluation Method. Given an observed output sequence O_{test} and a skill model λ_s, maximum log likelihood, $\log P(O_s|\lambda_s)$ was used to identify skill as

$$\lambda^* = \arg\max[\log P(O_{test}|\lambda_{se}), (\log P(O_{test}|\lambda_{si}), (\log P(O_{test}|\lambda_{sn})] \qquad (2)$$

where λ_{se},λ_{si},λ_{sn} are the expert, intermediate, and novice models for each trial, respectively.

Following [5,10], we also define a relative distance between a skill model λ_s and a model trained from the testing data λ_{test} based on the sequence of test observations

$$D(\lambda_s, \lambda_{test}) = \frac{1}{T_{test}} \min(\xi(\lambda_i, \lambda_{test}), \xi(\lambda_e, \lambda_{test}), \xi(\lambda_n, \lambda_{test})) \qquad (3)$$

where $\xi(\lambda_s, \lambda_{test}) = \log P(O_{test}|\lambda_{test}) - \log P(O_{test}|\lambda_s)$ and T_{test} is the length of the observation sequence O_{test}. This equation defines how well model λ_s matches observations generated by λ_{test}, relative to how well model λ_{test} matches observations generated by itself. Calculating the distance between a test sequence and skill levels: expert, intermediate, and novice yields three values. It is easy to see that the HMM model with minimum distance is also that with maximum log likelihood.

4 Surgeme-Level Skill Modeling

Data Preprocessing. The continuous motion values were discretized using a K-means algorithm with 8 cluster centers. The data was segmented into smaller motion blocks using the known manual surgeme labels, yielding a large set of observation subsequences for each surgeme.

HMM Surgeme Models. Three HMM expertise models for each surgeme were computed, totaling 24 skill models. We used 8 states after comparable results running the system with three, eight, and fourteen states.

Evaluation Method. In order to classify a test sequence as a particular skill level, we used Equation 2, where now λ_s is the trained surgeme expertise models and O_{test} is the observations from each surgeme occurrence. Each surgeme occurrence would have a corresponding skill label.

Each test trial now has a skill label associated to each surgeme occurrence. Majority voting was used on the test sequence labels, L_t, to classify the skill of that trial. In the event of a tie between skill levels, the maximum of the surgeme log-likelihood average was chosen.

5 Results

5.1 Experiment 1: Surgeme-Level HMM versus Task-Level HMM

o test the accuracy of the surgeme-based classification method, we performed a leave-one-out cross validation. In each round, one occurrence of a surgeme was left out for testing while the remaining occurrences of that surgeme was used to train an HMMs for each class. Out of the 1011 total surgeme occurrences, experts were correctly classified 75% (233/311), intermediates 59% (206/347), and novices 76% (268/353). For each expertise and surgeme, we present the number of correctly identified skill labels over the total number of occurrences of that surgeme (Table 2). The diagonal shows the correctly labeled surgemes. Reading across the matrix rows indicates the frequency of correctly classifying the skill of a surgeme and the frequency of it misclassified at another skill level. All rows sum to 1.

Table 2. Surgeme-based result: "Confusion Matrix"

	Exp.	Int.	Nov.	count
Exp. S1	0.50	0.28	0.22	18
Int. S1	0.33	0.67	0	18
Nov. S1	0.31	0	0.69	16
Exp. S2	0.76	0.12	0.12	76
Int. S2	0.16	0.78	0.06	77
Nov. S2	0.16	0.07	0.78	76
Exp. S3	0.79	0.17	0.04	76
Int. S3	0.35	0.53	0.12	75
Nov. S3	0.34	0.12	0.54	74
Exp. S4	0.89	0.02	0.09	57
Int. S4	0.00	0.78	0.22	27
Nov. S4	0.03	0.14	0.83	59
Exp. S5	-	0.25	0.75	4
Int. S5	0.11	0.79	0.11	19
Nov. S5	0.05	0.21	0.74	19
Exp. S6	0.71	0.08	0.22	78
Inter. S6	0.04	0.77	0.19	74
Nov. S6	0.05	0.17	0.79	42
Exp. S7	-	-	-	0
Inter. S7	-	0.92	0.08	36
Nov. S7	-	0.07	0.93	46
Exp. S8	-	-	1.0	2
Int. S8	-	0.76	0.24	21
Nov. S8	-	0.10	0.90	21

The data indicate that: (1) certain surgemes, such as 2, 4, 6, are indicative of skill based on high classification rates across all three skill levels, (2) other surgemes, such as 1, are not indicative of skill because the discrimination between skill levels is low, and (3) there are surgemes that are infrequently used by skill groups, such as surgemes 5,7,8 for an expert, suggesting that those are intermediate positioning moves. These results correlate to the distribution of time spent in each surgeme according to expertise. We found that regardless of the expertise level, most of the time was spent in surgeme 2 (Reaching for needle), then 3 (positioning needle), then 6 (Moving to center with needle in gripper). Incidentally, we also found that these surgemes were the ones that had the highest number of surgeme occurrences.

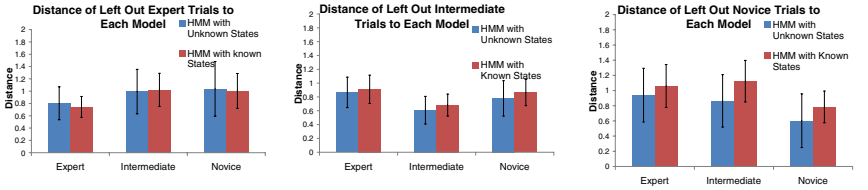


Fig. 4. Average distance, over 19 rounds, of the left out sequence to each class' model built using BW and KS

Intermediate skill was most difficult to classify correctly on the surgeme occurrence level as it had the highest percentage of misclassification overall, indicating that skill might be more accurately modeled as a scale instead of discrete classes. Assessing the surgemes in the context of their trials, we found that this evaluation system correctly labeled 100% of the trials. We found a 95% classification accuracy rate of the task-level with unknown states (BW) through 19 rounds of leave one out. In each round one trial from each class was left out. HMMs were built for each expertise and each left one trial out testing. Three expert trials were misclassified as a novice.

Table 3. Skill classification rates for trials using cross validation

Expertise	Classification Rate
1c: Surgeme (E)	100%
1c: Surgeme (I)	100%
1c: Surgeme (N)	100%
2c: Task BW(E)	84%
2c: Task BW (I)	100%
2c: Task BW(N)	100%
2a: Task KS(E)	100%
2a: Task KS (I)	100%
2a: Task KS(N)	100%

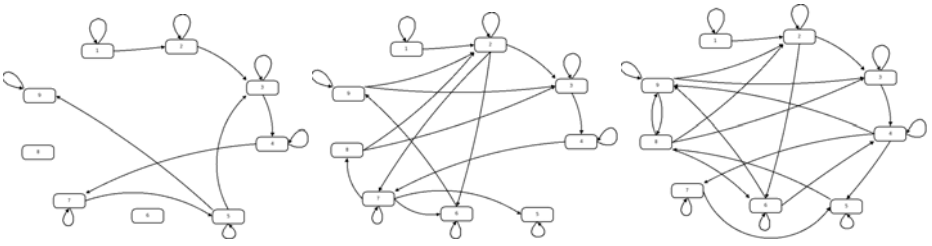
5.2 Experiment 2: Task-Level Hidden Markov Model with Known (KS) versus Unknown States (BW)

We performed 19 rounds of leave-one-out testing to determine the accuracy of this classification method. In each round, one trial from each class was left out. HMMs were built for each expertise and each left out test trial was tested using KS methods compared to BW.

Even evaluating the state transition diagrams, we see a qualitative difference between the movements of experts and novices (Figure 5). For the KS transition diagram, as surgeon skill increases; the graph of their movements becomes more directed. The expert surgeon accomplished a task using relatively few movements whereas the novice made more errors during the task causing them use extraneous motions and started over in the initial states.

There is still a difference between the expert and novice models using the Baum-Welch algorithm where the states are hidden. Even though the states do not correspond to the labeled surgemes it can still be observed that the expert model is less connected than the novice model. This shows that the movements of experts are more directed and less prone to erroneous or unnecessary movements.

Figure 4 shows the average distance over the 19 rounds of the expert, intermediate and novice trials being generated by each expertise model. The first of the three sets of



(a) Exp.KS (b) Int.KS (c) Nov.KS

Fig. 5. Expertise state transition diagram with known surgemes

bars represents the BW algorithm, and the second the KS algorithm. Each skill level had the minimum distance when compared against its own model. The intermediate skill level surgeon has a smaller distance to the expert than the novice surgeon. This seems to confirm that intermediate surgeons actually do perform the tasks in a manner that is closer to the expert than the novice does. KS yielded 100% trial classification accuracy which is an improvement over the 95% classification using BW.

6 Discussion and Conclusions

This paper addresses the problem of evaluating skill using continuous velocity data from the da Vinci system. Vector quantization techniques were used to discretize the data to train discrete Hidden Markov Models (HMMs). These experiments show that HMMs are a useful method to classify skill of an unknown trial based on maximum likelihoods to various trained skill models. Using the parsed motion segments, correct classification of each occurrence of a surgeme showed that surgemes that were more commonly used were more indicative of skill. Taking a scoring of all the labels during a trial, surgeme-level models achieved 100% correct classification over our 57 datasets. The task-level HMM with unknown states correctly classified 94.7% of the trials. We further analyzed the task-level HMMs by comparing the "true" HMM built from unlabeled data with HMMs using labeled surgemes as states. Interestingly, we found that classification of trials using HMMs surgemes as states increased to 100%. This indicates that we would be able to input raw data into the evaluator without any prior labels and correctly classify skill almost as well.

There are several extensions of this work. More importantly, this method can correctly identify skill but does not answer how a novice model can move towards an expert model. The categorization of an intermediate surgeon is somewhat ambiguous since it is a class that is between an expert and novice. It is interesting to note that the analysis was able to detect what appears to be a slight learning curve in the intermediate data. Thus, this may be an area of future investigation to pinpoint where novices and experts differ during a trial and reinforce correct technique. We will need to better understand the surgeme representation to analyze which surgemes are the most important and most indicative of skill. In the near future, we intend to evaluate our methods with a larger, more variable database of surgical motions we have recently collected.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 0534359 and the Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Moorthy, K., Munz, Y., Dosis, A., Hernandez, J., Martin, S., Bello, F., Rockall, T., Darzi, A.: Dexterity enhancement with robotic surgery. Surg. Endo. 18, 790–795 (2004)
2. Reznick, R., MacRae, H.: Teaching surgical skills-changes in the wind. New England Journal of Medicine 355(25), 2664–2669 (2006)
3. Xin, H., Zelek, J.S., Carnahan, H.: Laparoscopic surgery, perceptual limitations and force: A review. In: First Joint Eurohaptics Conf. and Symp. on Haptic Interfaces for Virtual Environment (2006)
4. Business, G.S.: Computer assisted surgical systems. A Global Strategic Business Report (2007)
5. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. IEEE 77(2), 257–286 (1989)
6. Yang, J., Xu, Y., Chen, M.: Hidden markov model approach to skill learning and its application to telerobotics. Technical report, Robotics Institute, Carnegie Mellon University (1993)
7. Hannaford, B., Lee, P.: Hidden markov model analysis of force/torque information in telemanipulation. Int. Journal of Robotics Research 10(5), 528–539 (1991)
8. Hovland, G., Sikka, P., McCarragher, B.: Skill acquisition from human demonstration using a hidden markov model. Int. Conf. on Rob. Automat. 10(5), 528–539 (1996)
9. Yu, W., Dubey, R., Pernalete, N.: Robotic therapy for persons with disabilities using hidden markov model based skill learning. In: IEEE Int. Conf. on Rob. Automat., pp. 2074–2079 (2004)
10. Rosen, J., Brown, J.D., Chang, L., Hannaford, M.N.S.B.: Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. IEEE Trans. in. Bio. Eng. 53(3), 399–413 (2006)
11. Murphy, T.E., Vignes, C.M., Yuh, D.D., Okamura, A.: Automatic motion recognition and skill evaluation for dynamic tasks. In: Eurohaptics, pp. 363–373 (2003)
12. Megali, G., Sinigaglia, S., Tonet, O., Dario, P.: Modelling and evaluation of surgical performance using hidden markov models. IEEE Trans. on Bio. Eng. 53(10), 1911–1919 (2006)
13. Lin, H.C., Shafran, I., Yuh, D.D., Hager, G.D.: Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Comp. Aid. Surg. 11(5), 220–223 (2006)
14. Cao, C.G., MacKenzie, R.D., Payandeh, S.: Task and motion analyses in endoscopic surgery. ASME IMECE, 583–590 (1996)
15. Guthart, G.S., Salisbury, J.K.: The intuitivetm telesurgery system: Overview and application. In: IEEE Int. Conf. on Rob. Automat., pp. 618–621 (2000)
16. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. Applied Statistics 8(1), 100–108 (1979)
17. Matlab: Statistics and pattern recognition toolkit