Markov Model Assessment of Subjects' Clinical Skill Using the E-Pelvis Physical Simulator

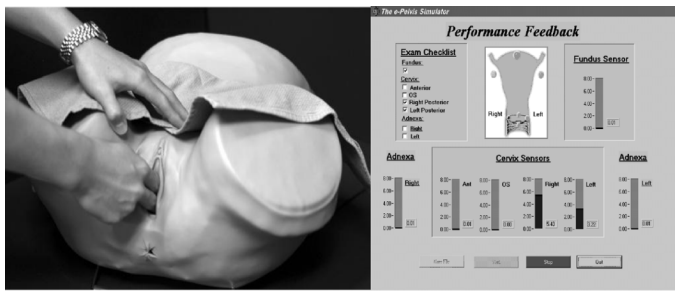
Thomas R. Mackel*, Jacob Rosen, *Member, IEEE*, and Carla M. Pugh

***Abstract*—Inherent difficulties evaluating clinical competence of physicians has led to the widespread use of subjective skill assessment techniques. Inspired by an analogy between spoken language and surgical procedure, a generalized methodology using Markov models (MMs), independent of the modality under study, was developed. The methodology applied to an endoscopic experiment in "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model" by J. Rosen *et al.* (*IEEE Trans. Biomed. Eng.*, Vol. 53, No. 3, pp. 399–413, Mar. 2006) is modified and applied to data collected with the E-Pelvis physical simulator. The simulator incorporates five contact pressure sensors located in key anatomical landmarks. Two 32-state fully connected MMs are used, one for each skill level. Each state corresponds to a unique five-dimensional signature of contact pressures. Statistical distances measured between models representing subjects with different skill levels are sensitive enough to provide an objective measure of medical skill level. The method was tested with 41 expert subjects and 41 novice subjects in addition to the 30 subjects used for training the MM. Of the 82 subjects, 76 (92%) were classified correctly. Unique state transitions as well as pressure magnitudes for corresponding states were found to be skill dependent. The "white box" nature of the model provides insight into the examination process performed.**

***Index Terms*—Classification, E-Pelvis, Markov model (MM), pressure sensing, skill assessment.**

I. Introduction

MEDICINE is in the process of converting its apprenticeship-training model of "watch, do, teach" to a model borrowed from aviation of "watch, *simulate*, do, teach." In this model, simulation is an inherent part of the training process and should precede any exam treatment or procedure performed on patients by physicians in training. One of the most critical elements of a medical simulator (for review, see [1]) is the ability to assess competency of high-level decision making and low-level skill in performing the medical task. Inherent difficulties in evaluating clinical competence of physicians have led to the widespread use of subjective skill assessment techniques. Subjective evaluation techniques lead to inconsistent evaluation by different examiners. The methodology for assessing surgical skill as a subset of surgical ability [2], [3] is gradually shifting from subjective scoring of an expert, which may be a variably biased opinion using vague criteria, towards a more objective, quantitative analysis. The ultimate aim is therefore to develop a modality-independent methodology for objectively assessing medical competency. The methodology may be incorporated into a simulator (physical or virtual reality) or a surgical robot, or track the performance during a medical procedure performed on a patient and provide objective and unbiased assessment based on quantitative data resulting in part from the physical interaction between the physician and treated modality.



(a) (b)

Fig. 1. E-Pelvis Simulator. (a) Simulated pelvic exam with the physical simulator. (b) Graphical user interface.

The field of speech recognition has shown that Markov modeling (MM) and hidden Markov models (HMMs) are effective methods for deconstructing and understanding speech data [4], [5]. These methods have been widely developed and used in a variety of other fields, such as visual recognition of gestures and facial expressions [6], [7], DNA and protein modeling [8], surgical tools in an MIS setup [9], [10], and robotic teleoperation [11]–[15]. Inspired by an analogy between the structure of a medical procedure and that of spoken language [16], MMs were adapted from the field of automatic speech recognition (ASR) for developing an objective skill assessment methodology of a medical procedure. A metric that represents the skill level of a subject can be determined by analyzing the subject's performance with respect to the performance of subjects of known skill levels. Previous studies applied the MM approach to skill evaluation of minimally invasive surgery [16]–[22]. Using an approach that is independent of the modality used by the physician, the aim of the current study is to utilize a similar MM approach in developing a methodology for objectively assessing clinical skills during a pelvic exam using data acquired with the E-Pelvis simulator [4], [23], [24]. The method outlined in [21] uses an MM with a discrete version of the "B" matrix to classify multiple levels of skill on endoscopic tools in an animal model,

Manuscript received March 3, 2006; revised February 17, 2007. This work was supported in part by the U.S. Army Medical Research and Materiel Command under Award Number W81XWH-04-1-0464. *Asterisk indicates corresponding author.*

*T. R. Mackel is with Rockwell Collins, Inc., Mail Stop 182-108, 400 Collins Road NE Cedar Rapids, IA 52498-0001 USA (e-mail: tmackel@gmail.com).

J. Rosen is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: rosen@u.washington.edu).

C. M. Pugh is with the Department of Surgery and the Center for Advanced Surgical Education, Northwestern University, Evanston, IL 60201 USA (e-mail: drpugh@northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2007.908338

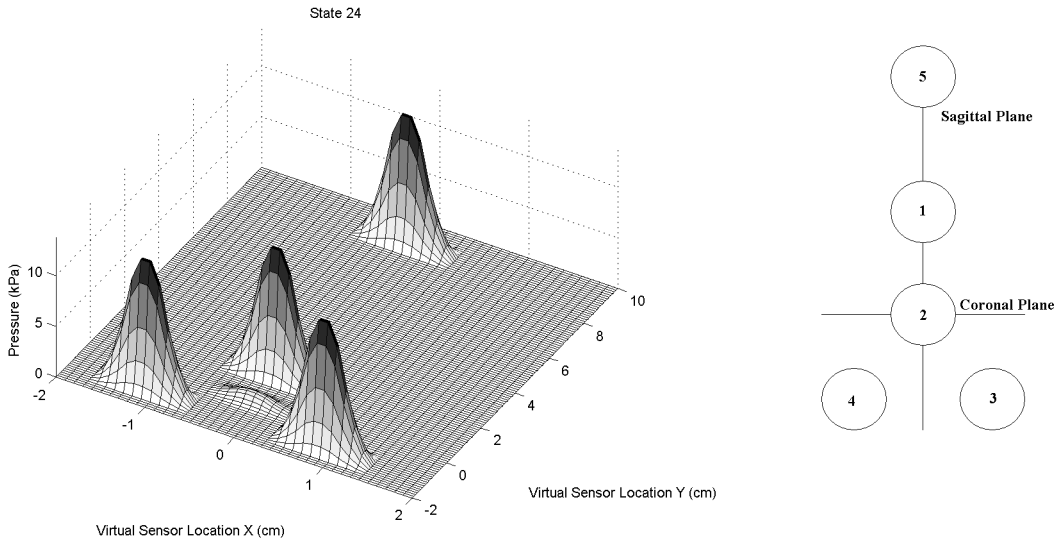


Fig. 2. Location of the five sensors on the E-Pelvis Simulator mannequin in the virtual environment.

whereas this work uses a continuous version of the "B" matrix to distinguish two levels of skill on a physical simulator.

II. Method

A. The E-Pelvis Physical Simulator and Database

The E-Pelvis, shown in Fig. 1, is a physical simulator that consists of a partial mannequin (umbilicus to mid thigh) constructed in the likeness of an adult human female [4], [23], [24]. The simulator sampled data at 30 Hz from five pressure-sensing resistors (FSR) located on key anatomical structures while the subjects performed pelvic examinations. The sensor locations are shown in Fig. 2. The examination can be reconstructed virtually from the pressure sensor data, as shown in Fig. 3, by mapping the sensor coordinates in a virtual cylindrical model of the E-pelvis simulator, which shows the intensity of pressure at each sensor location, onto a Cartesian coordinate system. One pressure unit (PU) as measured by the sensors is approximately equal to 6.9 kPa. Data recorded by subjects using this simulator were selected to test the evaluation methodology discussed herein. The 41 expert subjects were selected at random out of data collected from professional examiners. The 41 novice subjects were selected at random out of data collected from medical students. A different set of subjects from these groups, 15 experts and 15 novices, were selected at random from the remaining data to train the Markov models.

B. Data Collection

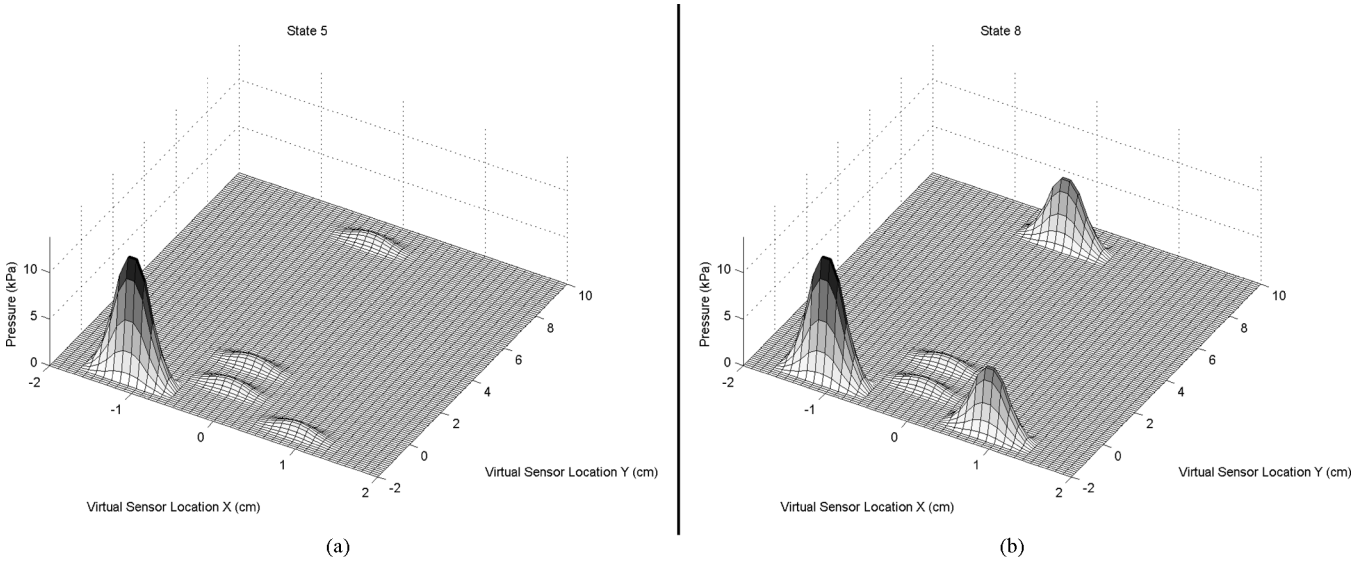
All of the second-year medical students in one school participated in the study to form one group. The group of professional examiners was a convenience sample of volunteers attending one of the largest annual meetings of OB/Gyn physicians. For both groups, the simulator was placed on a table that was 34 in from the floor which is the average examination height. The examiners were told the patient was there for an annual check up and has no complaints. The examiners were not informed as to the internal configuration of the model (i.e., normal or abnormal) and were asked to perform a complete internal pelvic examination. The internal examination begins with circumferential examination of the cervix. The next step is bimanual examination of the uterus and ovaries. This involves applying pressure on the lower abdominal wall and on the cervix simultaneously to trap the organ between two hands. There are no rules on proceeding from left to right, for example, or clockwise versus counterclockwise. No time constraints were placed on the exam. The mean and standard deviation of the collected data is shown in Table I.

TABLE I
Summary of Pressure Data Collected From Subjects. All Values are in Kilopascals (kPa)

	Sensor	1	2	3	4	5
Expert	Mean	0.1538	0.2937	0.7371	0.8459	0.4073
	Std. Dev.	1.1063	1.3086	1.5509	1.9363	1.429
Novice	Mean	0.1095	0.8616	0.5027	0.8061	0.2359
	Std. Dev.	0.7637	2.3172	1.3892	1.6639	1.0447

C. Data Analysis

An analogy between spoken language and minimally invasive surgery tasks [16]–[22] was extended to pelvic examination tasks. In the same way that a paragraph or book chapter can be broken down into single words, a medical procedure can be broken down into basic maneuvers or "states." Defining and analyzing these "states" is a key step in decomposing the medical procedure. The MM incorporating these states represents the medical procedure as a process, and as such it allows the construction of an objective medical performance. The MM is defined by 32 fully connected states. Each state is characterized by properties of a five-dimensional binary vector associated with the five pressure sensors incorporated into the physical simulator (Table II). The sensors are categorized as being either active or inactive, depending on the recorded pressure level. If the recorded pressure level exceeds 1 PU (6.9 kPa), then the



(a) (b)

Fig. 3. Amplitude of the surface corresponds to the pressure at each of the sensor locations. Screenshots of sensor response during pelvic exam. Left: high magnitude of pressure applied to sensor 3, Right: low magnitude of pressure applied to sensors 4 and 5, and high magnitude of pressure applied to sensor 3.

TABLE II
MAPPING FROM ACTIVE/INACTIVE SENSOR COMBINATIONS TO MARKOV MODEL STATE DEFINITIONS. SUBJECTS DO NOT USE ALL STATES. SENSORS WHICH ARE ACTIVE IN EACH STATE ARE MARKED WITH AN "X"

Sensor	State 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1																	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2									x	x	x	x	x	x	x	x									x	x	x	x	x	x	x	x
3					x	x	x	x					x	x	x	x					x	x	x	x					x	x	x	x
4			x	x			x	x			x	x			x	x			x	x			x	x			x	x			x	x
5		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x		x

sensor is considered active, otherwise it is considered inactive. All possible combinations of active and inactive for five sensors result in 2^5 total possible states. This means of sampling the continuous level data to discrete levels causes a necessary loss of information. The more states contained in the model, the less severe this loss is, but the more complex the model.

Each subject's performance, represented as a five-dimensional vector of length N, was encoded and therefore mapped into one of the MM states. In this way, different "pronunciations" of each state were observed in the pressure data measured by the E-Pelvis simulator. Data characterizing the performance of two categories of medical examiners, expert and novice, were analyzed using two 32-state fully-connected MMs (Fig. 4). Within each model, certain sequences of state transitions, known as Markov chains, are more probable than others. Many states were more commonly used than others, resulting in an uneven distribution of the data points between states.

D. Constructing the Markov Models

1) Transition Probability ("A*" Matrix):* The frequency transition matrix ("A" matrix [25]) defines the probability that a transition occurs between any two states. The elements of the "A" matrix were computed by counting the state transitions that occurred in the training subjects' data. The element in row i, column j, is found by counting the number of transitions from state i to state j, then dividing by the total number of transitions from state i to any state.

The transition probability (for model λ) is given by the frequency transition matrix, $\mathbf{A}_\lambda$. This probability takes into account the probability of transitioning from the previous state S_{n-1} to the current state S_n, and is found by directly indexing the transition matrix (1). The first data point $\mathbf{x}[0]$ is a special case, as the previous state is unknown, and $\mathrm{P_T}$ is assigned a value of 1 for this case, indicating the subject started in the idle state where none of the pressure sensors were touched.

$$P_{\mathrm{T}}(\mathbf{x}[n]|\lambda) = \mathbf{A}_\lambda[S_{n-1}, S_n]. \tag{1}$$

2) Membership Probability ("B*" Matrix):* A mean vector and covariance matrix ("B" matrix [25]—continuous version observation) is sufficient to model the probability density of each state as a multivariate Gaussian. The data points representative of each state were used to compute the 1×5 mean vector, μ, and 5×5 covariance matrix, Σ. The output of the probability density function (2) represents the likelihood of the continuous valued input data point being a member of the corresponding state

$$L(\mathbf{x}[n]|\lambda) = \frac{1}{(2\pi)^{p/2}\sqrt{||\Sigma||}} e^{-(\mathbf{x}[n]-\mu)^T\Sigma^{-1}(\mathbf{x}[n]-\mu)/2} \tag{2}$$

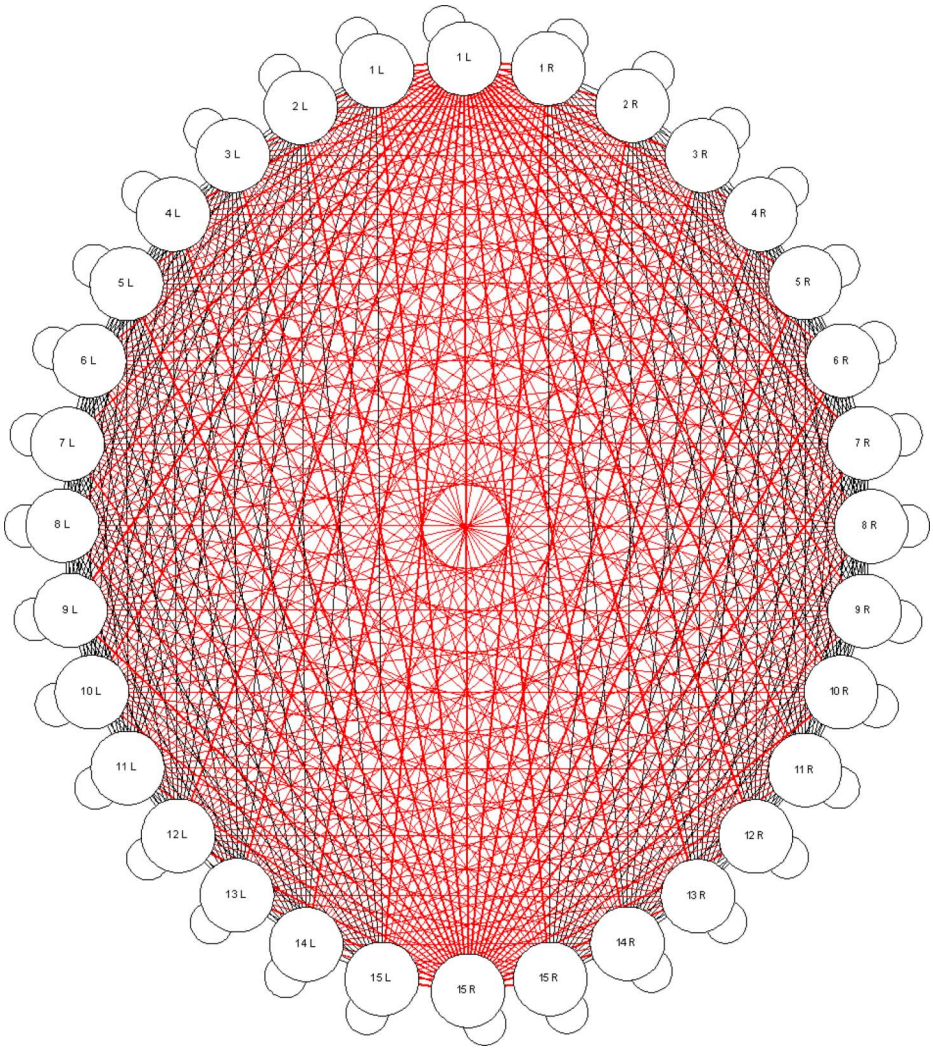


Fig. 4. One 32-state fully connected Markov model state diagram.

where p is the number of dimensions of data, 5 in the case of the E-Pelvis.

The membership probability is defined using the total probability rule (3)

$$P_M(\mathbf{x}[n]|\lambda) = \frac{p(\lambda)L(\mathbf{x}[n]|\lambda)}{\sum_{j=1}^{K} p(\lambda_j)L(\mathbf{x}[n]|\lambda_j)} \tag{3}$$

where $p(\lambda)$ represents the *a priori* probabilities of each model, $L(\mathbf{x}[n]|\lambda)$ is the likelihood of data point $\mathbf{x}[n]$ belonging to model λ, and K is the total number of models. In the case of this E-Pelvis simulator database, a novice model and an expert model are used, and $\mathrm{K} = 2$. Equation (3) is simplified by assuming identical *a priori* probabilities for each model.

$$P_M(\mathbf{x}[n] \mid \lambda) = \frac{L(\mathbf{x}[n] \mid \lambda)}{\sum_{j=1}^{2} L(\mathbf{x}[n] \mid \lambda_j)}. \tag{4}$$

3) Bayesian Classifier: Bayes' Decision Rule was used to classify an unknown subject as either an expert or novice. If there are two classes, A and B, this rule states to choose class A if $\mathrm{P(A)} > \mathrm{P(B)}$, choose class B otherwise. Define observation vector O as a sequence of N data points $\mathbf{x}[n]$. Let P(A) be the probability of a sequence of data points arising from an expert subject model, $\mathrm{P}(O|\lambda_{ES})$, and P(B) to be the probability of a sequence of data points arising from a novice subject model, $\mathrm{P}(O|\lambda_{NS})$.

$\mathrm{P}(\mathbf{x}[n]|\lambda)$ is the product of the *membership probability* $\mathrm{P}_M(\mathbf{x}[n]|\lambda)$ ("B" matrix) and the *transition probability* $\mathrm{P}_\mathrm{T}(\mathbf{x}[n]|\lambda)$ ("A" matrix). The probability that model λ would generate an observation sequence O, $\mathrm{P}(O|\lambda)$, is the product of probabilities that each data point $\mathbf{x}[\mathrm{i}]$ was produced by model λ, $\mathrm{P}(\mathbf{x}[\mathrm{i}]|\lambda)$ (5)

$$P(O|\lambda) = \prod_{i=1}^{N} P(\mathbf{x}[i]|\lambda). \tag{5}$$

Given a sequence of data associated with a specific subject, the above method can be used to estimate the probability that the MM of a class generated the sequence. The subject can be classified as a member of the class whose model results in the highest probability.

4) Objective Subject Comparison: More data are collected during slower examinations. This penalizes both models by increasing the length of the observation vector O, hence increasing the number of factors used to compute $\mathrm{P}(O|\lambda)$. As a result, the $\mathrm{P}(O|\lambda)$ of one subject can not be directly compared to the $\mathrm{P}(O|\lambda)$ of another. Subjects' behavior relative to one another cannot be measured without the use of a common benchmark. One method uses a third MM, trained from the subject's own data samples, $\mathrm{P}(O|\lambda_{OS})$, which is compared to the novice model and the expert model [17]. Two statistical factors [expert skill factor (ESF) and novice skill factor (NSF)] can be defined as

$$\mathrm{NSF} = \frac{\log(P(O|\lambda_{OS}))}{\log(P(O|\lambda_{NS}))} \tag{6}$$

$$\mathrm{ESF} = \frac{\log(P(O|\lambda_{OS}))}{\log(P(O|\lambda_{ES}))} \tag{7}$$

where O is an observation vector representing the subject's performance, λ_{OS} is a subject model trained by the data O, and λ_{ES} and λ_{NS} are models trained by data from experts and novices, respectively.

There are two methods of finding $\mathrm{P}(O|\lambda_{OS})$. In one method, the membership probability is affected by the presence of the subject model. Using this method, $P_M(O|\lambda_{OS}) + P_M(O|\lambda_{ES}) + P_M(O|\lambda_{NS}) = 1$ for each observation point. Most points result in a high-probability match to the subject model, since the subject model was constructed from the data points. Prior to the inclusion of the subject model, whichever class model most closely fit the data (the "correct" model) had a high-probability match to many data points. The class model that did not closely fit the data (the "incorrect" model) had a high-probability match to few data points. After the inclusion of the subject model, points that formerly resulted in a high-probability match to one of the class models instead have a high-probability match to the subject model. The correct model is penalized more than the incorrect model because more points had matched the correct model before the inclusion of the subject model. Therefore, the distinction between the expert and novice membership probabilities becomes somewhat more obscured when using this method.

In the other method, the membership probability is unaffected by the subject model. $P_M(O|\lambda_{OS}) = P_M(O|\lambda_{ES}) + P_M(O|\lambda_{NS}) = 1$ for each observation point. Only the subject model's transition matrix influences the overall probability value of $P(O|\lambda_{OS})$. This method prevents the subject model itself from biasing the classification results in favor of the incorrect model and is used to compute skill factors for this reason.

Given the "white box" nature of the MM, analyzing the models provides insight into the process in which the pelvic exam is performed through examining the Markov chains. A Markov chain is a sequence of states of an MM. Observing the most probable transitions (excluding same-state transitions) within each model's transition matrix can identify Markov chains. Markov chains are characteristics of the model that can distinguish between subjects of different skill level. The most-probable transitions are known as top-level chains. Some top-level chains are distinct to each skill level, and do not occur even as a subset of a chain of either skill level.

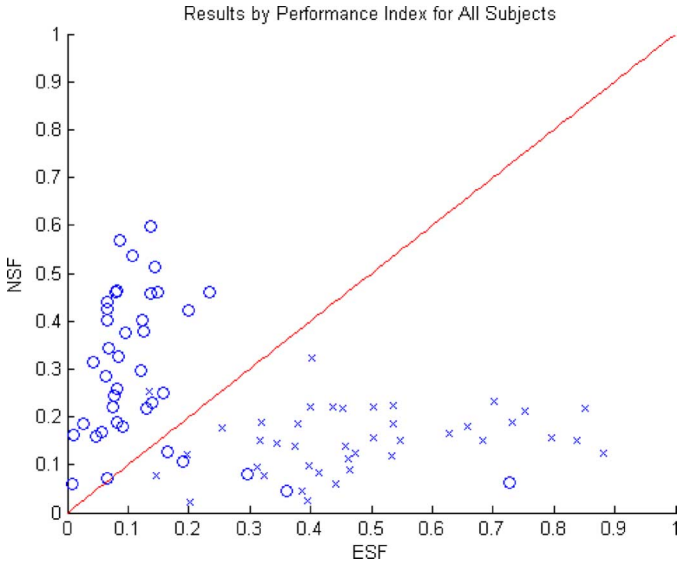


Fig. 5. Performance index derived by plotting the novice skill factor against the expert skill factor, determined by model outputs, for each subject. Data points marked with an "x" are known to be expert subjects, those marked with a "o" are novice. The line that is drawn is a decision boundary used by the algorithm to decide the class of the plotted subject. Ideally, all of the "x's" would be on one side of this line, and all of the "o's" on the other.

A special class of chain is referred to as an "impractical" chain. The etymological analogy of an "impractical" chain is a phrase or sentence that is missing phoneme(s). These chains may still provide useful information, but some information is lost beyond the physical data collection capabilities of the simulator. A sensor which is below the activation threshold at data sample n but is above the threshold at sample $n+1$ is activating at sample n. Let the activation or deactivation of a sensor be called an action. In a continuous-time real-world scenario, it is possibly only in theory for multiple actions to occur at the exact same instant in time. Chains that illustrate multiple actions occurring simultaneously are recognized as "impractical" chains. They are impractical because in the real world, according to Newtonian physics and given an accurate enough measurement device, there will always be a finite duration of time between the activation of one sensor and the deactivation of another. If the duration is short enough, multiple different actions will appear to be simultaneous in the sampled data due to the finite sampling rate of the simulator and electrical noise in the measurement. One such example is the expert chain 22-8-7-3-1. Referring to Table II, sensors 1, 3, and 5 are active in state 22, while sensors 3, 4, and 5 are active in state 8. State 22 cannot pass instantaneously to State 8 because it is not possible to deactivate sensor 1 and to activate sensor 4 at the same instant in time. Another example is the chain 24-5-1. These chains are missing some high-frequency information that has been lost beyond the simulator's sampling bandwidth. This lost information may not be useful, but the possibility of useful information being lost exists.

Two-factor analysis of variance (ANOVA) was performed three times for each of the 32 states. The first factor used is

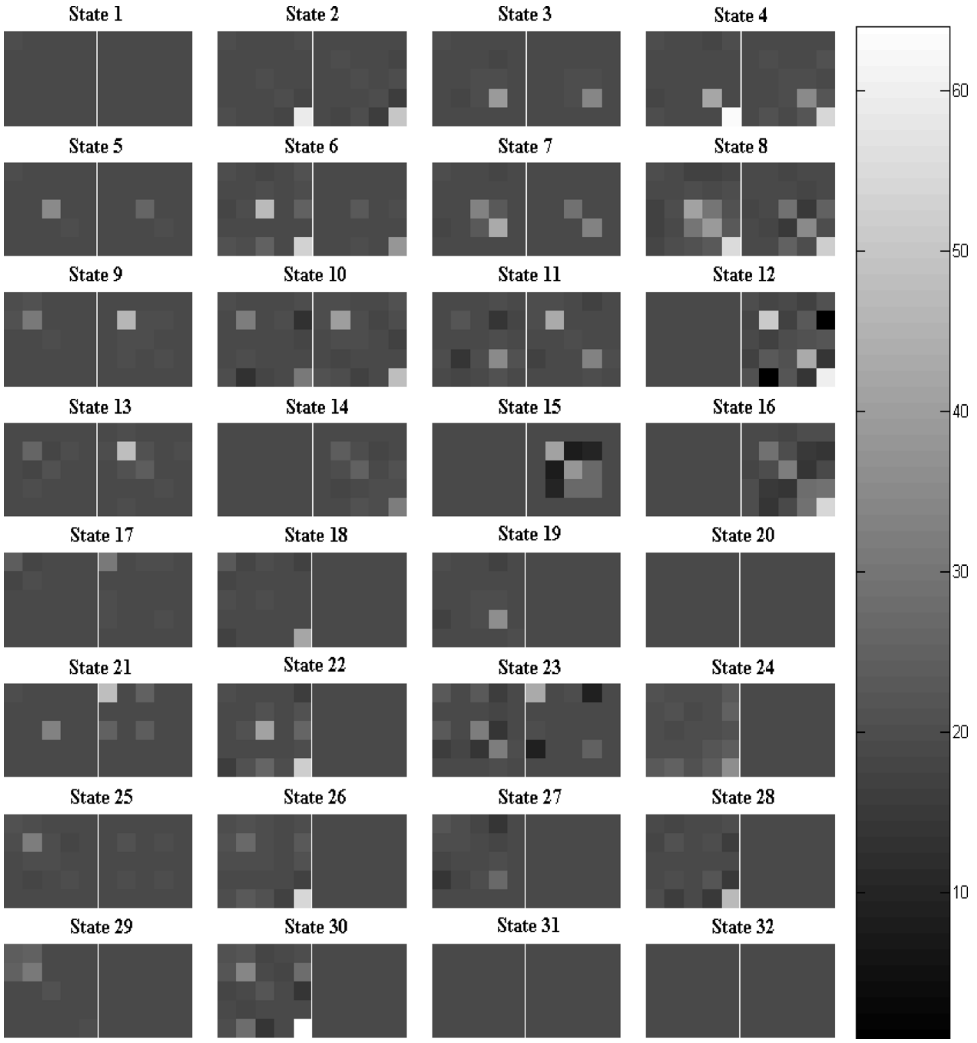


Fig. 6. 5×5 Covariance matrices ("B" matrices) for each state showing covariance between sensor readings. Below each state label are two 5×5 image representations of two covariance matrices, the expert matrix is on the left, while the novice matrix is on the right. The intensity of the pixel in row j, column k corresponds to the numerical value of the covariance between sensors j and k. The top-left pixel is in row 1, column 1, while the bottom-right pixel is row 5, column 5.

the sensor number (from 1 to 5), and the second factor is class (group A and group B). Groups A and B were changed for each of the ANOVA experiments. In the first experiment, Group A and B were each composed of different random selections of known expert data points. In the second, Group A was composed a random selection of expert data points, and Group B was composed of a random selection of novice data points. In the third, Groups A and B were composed of random selections of novice data points. The ANOVA interaction significance demonstrates that when the random data points which comprise Groups A and B are taken from different classes of subjects, there is a greater statistical difference between groups than when the data points are taken from the same class of subject.

III. Results

Using the MMs to compute Bayesian classifier probabilities, 40 of the 41 expert subjects (97.6%), and 36 of the 41 novice subjects (87.8%) were correctly classified using this method. Overall, 76 of the 82 subjects tested (92.7%) were correctly classified. These subjects were not used in the training of the MMs. Misclassification is considered if a subject known as an expert by training is classified by the MM as a novice and vise versa. One may note that more novices were misclassified as experts then experts misclassified as novices. The greater a subject's ESF, the more closely the subject matches the expert model, and similarly for the NSF. Plotting these two skill factors against each other yields a performance index, plotted in Fig. 5. It is possible for a subject to exhibit a higher NSF than another subject, yet be classified as an expert while the other subject is not. This is because the decision is based on the proportional relationship between an individual's NSF and ESF. In Fig. 6, the statistical covariance matrix between the five sensors is plotted as an image. There are two covariance matrices plotted for each state, one for each skill level, and 32 states are represented, for a total of 64 images. Each element of the 5×5 covariance matrix is treated as a pixel of the image, and the intensity of the pixel corresponds to the value of the covariance. The solid-color images are those for which there was very little or no subject data.

The Markov chains reveal aspects that differentiate between expert and novice subjects. All five sensors are inactive in State 1, and it is treated as an "idle" state in which all of the most probable sequences of transitions (top-level chains) end. Table III

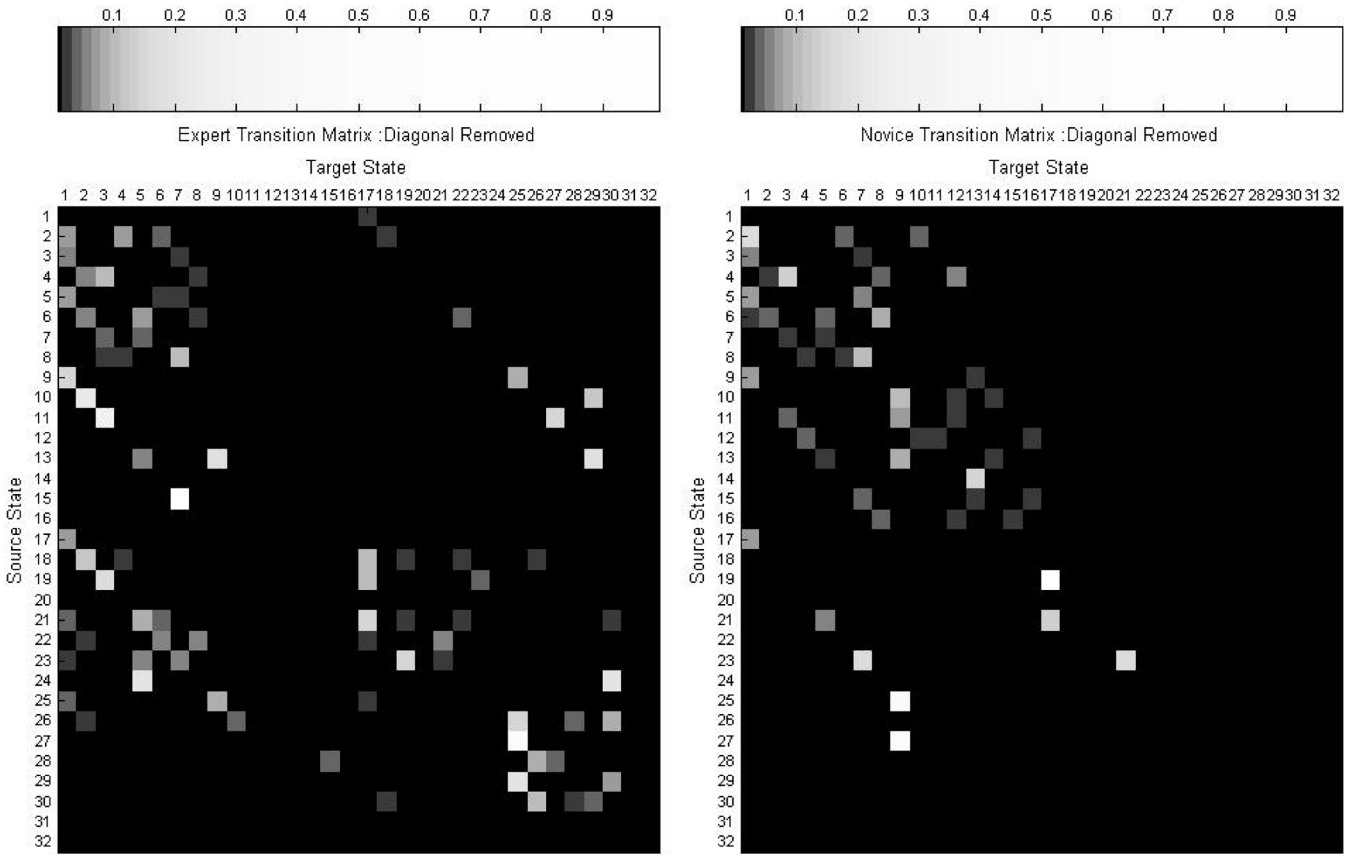


Fig. 7. Transition matrices of expert and novice Markov models—indicates the probability of a subject transitioning from a source state (column) to any target state (row). Each row sums to 1.

TABLE III
TOP-LEVEL MARKOV CHAINS

Expert Chains (19 chains total)	*Novice Chains* (13 chains total)
{2-4,15-7,11,23-19}-3-1	{6-8,15,16-8,23}-7-3-1
13-{N,29-25}-9-1	{19,23-21}-17-1
{18,23-19,22-21}-17-1	{10,11,14-13,25,27}-9-1
22-{6-5,8-7-3}-1	12-4-3-1
24-5-1	5-1
{11-27,24-30-26,28-26,29}-25-9-1	
{10,18}-2-4-3-1	
10-29-25-9-1	

lists all of the top-level chains found in the training data. Transitions are indicated by a hyphen ("-"). A table entry which states "15-7-3-1" would indicate that, for a given class of subject, a data point currently in State 15 is most likely to transition into State 7 next, followed by State 3 and then finally State 1. In some cases, there are multiple states or sequences that are approximately equally likely to occur. These multiple paths are denoted as comma separated entries within brackets ("{·}"). A table entry of "7-{5,3}-1" would indicate that the transition from State 7 to 5 has nearly the same probability as the transition from State 7 to 3, and both chains "7-5-1" and "7-3-1" are counted as top-level chains. A table entry of "N" is used to indicate a null transition.

Each top-level chain is determined by looking at the transition matrix (Fig. 7). Each state that was used by the subject is considered. Starting with State 2, the most likely transition from that state is State 4. In State 4, the most likely transition is to State 3. In State 3, the most likely transition is to State 1. All chains end when they arrive at State 1 (the idle state). This process yields the chain 2-4-3-1. The process is then repeated for every state used by the subject, instead of just State 2. At the end, all chains derived via this process and, are a subset of a longer chain belonging to the same class, are removed from the list of top-level chains. All the information contained in the subset chain, that is, the string of most probable transitions from a starting state is already contained in the longer chain, i.e., 22-6-5-1 contains all of the information that 6-5-1 contains, so it is not necessary to record both chains.

Table IV lists distinct top-level chains. Distinct chains are those that do not occur in both expert and novice models, even as a subset of another longer chain that is unique to each level. The removal of chains that are a subset of another chain is expanded to include chains from both classes instead of just chains belonging to the same class. Analyzing the results listed in Table IV indicates that the chains performed by the experts and novices follow two types: 1) decreasing numbers of sensors simultaneously activated (e.g., 4,3,2,1,0 or 3,2,1,0) and 2) alternating numbers of active sensors (e.g., 2,1,2,1,0 or 2,3,2,1,0). Out of the most frequently used chain by the expert 64% were type 1) and 36% were Type 2) whereas with the novice group 89% were type 1) and 11% were type 2).

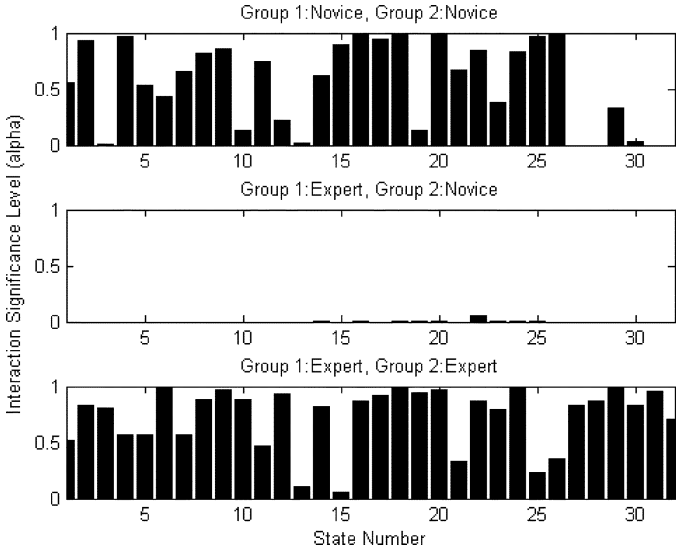


Fig. 8. Each of the 32 states is listed along the x-axis. The height of each bar corresponds to the ANOVA Interaction Significance Results for that state. Smaller values indicate greater difference between groups. Bottom: expert data compared to expert data; Middle: expert data compared to novice data; Top: novice data compared to novice data.

TABLE IV
DISTINCT TOP-LEVEL MARKOV CHAINS

Distinct Expert	*Distinct Novice*
Top-Level Chains (11)	*Top-Level Chains (9)*
11-3-1	10-9-1
18-17-1	11-9-1
22-6-5-1	12-4-3-1
23-19-3-1	23-21-17-1
23-19-17-1	14-13-9-1
22-21-17-1	23-7-3-1
13-29-25-9-1	15-7-3-1
11-27-25-9-1	16-8-7-3-1
28-26-25-9-1	6-8-7-3-1
10-2-4-3-1	
18-2-4-3-1	

ANOVA was performed to compare the data from each novice state with the corresponding expert state. The results pertaining to the interaction significance are shown in Fig. 8 for each state in graphical form. Three bar graphs are present. Each bar graph corresponds to a different set of input group data for the ANOVA algorithm. In the top graph, the ANOVA compared input data from novice subjects to input data from novice subjects. In the middle graph, novice subjects were compared to expert subject. In the bottom graph, experts were compared to experts. In each graph, there is one bar for each of the 32 states. Each bar shows the relative degree of interaction significance obtained from the ANOVA for that state. The interaction significance represents the statistical similarity between groups. It considers how each group interacts with each of the five sensors differently. Smaller interaction significance corresponds to a greater difference between how the groups interact with the sensors. The interaction significance is much smaller when the subjects compared during the ANOVA are members of different groups.

IV. DISCUSSION

An objective assessment methodology based on Markov model was developed and tested with 112 subjects while performing a pelvic exam with a simulator (E-pelvis). The error rate of this MM methodology is similar to the error rate of many commercially available multiclass ASR systems. Analyzing the internal structure of the Markov model indicates that the magnitudes of pressures applied, the states that were used, and the state transitions (Markov chains) that were utilized are all skill dependent. Twenty top-level Markov chains distinct to only one model are shown in Table IV. Each chain represents a particular expression of a physical human/simulator interaction. The interactions made by an expert in accomplishing the pelvic exam task may differ from the interactions made by a novice. Revisiting the spoken language analogy, it is possible to form both a grammatically correct sentence and a grammatically incorrect sentence that express a similar meaning. In the E-Pelvis examination simulator, given that a subject has entered State 11, the most common expert motion is the chain 11-3-1, while the most common novice motion is the chain 11-9-1. A potential model simplification can be achieved by treating the Markov chains as the fundamental model states, instead of the 32 states defined in this experiment. This may lead to a relatively large reduction in the number of states currently used by the model (32 states) with a relatively small amount of additional information loss due to quantization of the data.

One of the most interesting results of this analysis is to demonstrate that the information needed to distinguish between skill levels is in fact present in the recorded data. This is verified in a number of ways: 1) the output of the Bayesian classification algorithm, 2) analysis of the Markov model components, and 3) statistical analysis of the data (ANOVA). They all indicated that sufficient amount of distinguishing information was present in the data; i.e., the classification algorithm successfully differentiated the two skill levels under study, and the ANOVA showed such a significant difference in output when subjects with different skill level (versus same skill level) were compared.

The use of MMs for data analysis has been successfully applied to speech recognition. Extension of this concept to objective medical skill assessment has lead to a successful Bayesian dichotomous classification method. Subjects classified in this manner can be compared to one another using their performance indices. The distinct chains identified in the models may help to reduce the number of states in the models. The strength of this methodology is that it is independent of the modality under study. It was previously used to assess surgical skill in a minimally invasive surgical setup using the BlueDRAGON [21], and it is currently applied to data collected using the E-Pelvis as a physical simulator. Similarly, the same methodology can be incorporated into a surgical robot as a supervisory controller that could detect potentially dangerous mistakes by a human or computer operator.

REFERENCES

[1] A. Liu, F. Tendick, K. Cleary, and C. R. Kaufmann, "A survey of surgical simulation: Applications, technology, and education," *Presence: Teleoperators Virtual Environments*, vol. 12, no. 6, pp. 599–614, Dec. 2003.

[2] R. Satava *et al.*, "Metrics final report: Developing quantitative measurements through surgical simulation," in *Metrics for Objective Assessment of Surgical Skills Workshop*, 2001, CD-ROM.
[3] A. G. Gallagher and R. M. Satava, "Virtual reality as a metric for the assessment of laparoscopic psychomotor skills," *Surgical Endoscopy*, vol. 16, no. 12, pp. 1746–1752, Dec. 2002.
[4] C. M. Pugh and P. Youngblood, "Development and validation of assessment measures for a newly developed physical examination simulator," *J. Amer. Med. Inf. Assoc.*, vol. 9, no. 5, pp. 448–460, Sep.-Oct. 2002.
[5] J. Berkely, S. Weghorst, H. Gladstone, G. Raugi, and M. Ganter, "Fast finite element modeling for surgical simulation," *Studies Health Technol. Inf.*, vol. 62, pp. 55–61, 1999.
[6] C. S. Tseng, Y. Y. Lee, Y. P. Chan, S. S. Wu, and A. W. Chiu, "A PC-based surgical simulator for laparoscopic surgery," *Studies Health Technol. Inf.*, vol. 50, pp. 55–60, 1998.
[7] H. Delingette, S. Cotin, and N. Ayache, "Efficient linear elastic models of soft tissues for real-time surgery simulation," *Studies Health Technol. Inf.*, vol. 62, pp. 100–101, 1999.
[8] C. Basdogan, C. H. Ho, and M. A. Srinivasan, "Simulation of tissue cutting and bleeding for laparoscopic surgery using auxiliary surfaces," *Studies Health Technol. Inf.*, vol. 62, pp. 38–44, 1999.
[9] C. Richards, J. Rosen, B. Hannaford, M. MacFarlane, C. Pellegrini, and M. Sinanan, "Skills evaluation in minimally invasive surgery using force/torque signatures," *Surgical Endoscopy*, vol. 14, no. 9, pp. 791–798, Sep. 2000.
[10] E. Acosta, B. Temkin, T. M. Krummel, and W. L. Heinrichs, "G2H–graphics-to-haptic virtual environment development tool for PC's," *Studies Health Technol. Inf.*, vol. 70, pp. 1–3, 2000.
[11] Y. Akatsuka, T. Shibasaki, A. Saito, A. Kosaka, H. Matsuzaki, T. Asano, and Y. Furuhashi, "Navigation system for neurosurgery with PC platform," *Studies Health Technol. Inf.*, vol. 70, pp. 10–16, 2000.
[12] J. Berkley, P. Oppenheimer, S. Weghorst, D. Berg, G. Raugi, D. Haynor, M. Ganter, C. Brooking, and G. Turkiyyah, "Creating fast finite element models from medical images," *Studies Health Technol. Inf.*, vol. 70, pp. 26–32, 2000.
[13] N. El-Khalili, K. Brodlie, and D. Kessel, "WebSTer: A web-based surgical training system," *Studies Health Technol. Inf.*, vol. 70, pp. 69–75, 2000.
[14] R. Friedl, M. Preisack, M. Schefer, W. Klas, J. Tremper, T. Rose, J. Bay, J. Albers, P. Engels, P. Guilliard, C. F. Vahl, and A. Hannekum, "CardioOp: An integrated approach to teleteaching in cardiac surgery," *Studies Health Technol. Inf.*, vol. 70, pp. 76–82, 2000.
[15] E. Gobbetti, M. Tuveri, G. Zanetti, and A. Zorcolo, "Catheter insertion simulation with co-registered direct volume rendering and haptic feedback," *Studies Health Technol. Inf.*, vol. 70, pp. 96–98, 2000.
[16] J. Rosen, L. Chang, J. D. Brown, B. Hannaford, M. Sinanan, and R. Satava, "Minimally invasive surgery task decomposition – etymology of endoscopic suturing," *Studies Health Technol. Inf.*, vol. 94, pp. 295–301, 2003.
[17] J. Rosen, B. Hannaford, C. G. Richards, and M. N. Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 5, pp. 579–591, May 2001.
[18] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan, "Objective evaluation of laparoscopic skills based on haptic information and tool/tissue interactions," *Comput. Aided Surgery*, vol. 7, no. 1, pp. 49–61, Jul. 2002.
[19] J. Rosen, J. D. Brown, M. Barreca, L. Chang, B. Hannaford, and M. Sinanan, "The blue DRAGON – A system for monitoring the kinematics and the dynamics of endoscopic tools in minimally invasive surgery for objective laparoscopic skill assessment," *Studies Health Technol. Inf.*, vol. 85, pp. 412–418, 2002.
[20] J. Rosen, J. D. Brown, L. Chang, M. Barreca, M. Sinanan, and B. Hannaford, "The blue DRAGON – A system for measuring the kinematics and the dynamics of minimally invasive surgical tools *in–vivo*," in *Proc. 2002 IEEE Int. Conf. Robotics Autom.*, 2002, vol. 2, pp. 1876–1881.
[21] J. Rosen, J. D. Brown, L. Chang, M. Sinanan, and B. Hannaford, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 399–413, Mar. 2006.
[22] T. M. Kowalewski, J. Rosen, L. Chang, M. Sinanan, and B. Hannaford, "Optimization of a vector quantization codebook for objective evaluation of surgical skill," *Studies Health Technol. Inf.*, vol. 98, pp. 174–179, 2004.
[23] C. M. Pugh *et al.*, "The effect of simulator use on learning and self-assessment: The case of Stanford University's E-Pelvis simulator," *Studies Health Technol. Inf.*, vol. 81, pp. 396–400, 2001.
[24] C. M. Pugh and J. Rosen, "Qualitative and quantitative analysis of pressure sensor data acquired by the E-Pelvis simulator during simulated pelvic examinations," *Studies Health Technol. Inf.*, vol. 85, pp. 376–379, 2002.
[25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

Thomas R. Mackel received the M.S. degree in electrical engineering from the Rose-Hulman Institute of Technology, Terre Haute, IN, in 2004.

His research interests include parametric probability-based modeling and passive radio frequency identification devices. He is currently working in Avionic Systems Engineering at Rockwell-Collins, Inc., Cedar Rapids, IA.

Jacob Rosen (M'02) received the B.Sc. degree in mechanical engineering and the M.Sc. and Ph.D. degrees in biomedical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1987, 1993, and 1997, respectively.

Since 1997, he has been with the University of Washington, Seattle, with an appointment of Research Associate Professor of Electrical Engineering since 2006, and codirector of the Biorobotics Lab, and adjunct appointments with the Departments of Surgery and Mechanical Engineering. His research interests focus on medical robotics, biorobotics, human-centered robotics, surgical robotics, wearable robotics, rehabilitation robotics, neural control, and human–machine interface.

Carla M. Pugh received the undergraduate degree in neurobiology from the University of California, Berkeley, and the medical degree from the Howard University School of Medicine, Washington, D.C. Upon completion of her surgical training at Howard University Hospital, she received the Ph.D. degree in education from Stanford University, Stanford, CA.

She is currently Assistant Professor of Surgery and Director of the Center for Advanced Surgical Education at Northwestern University, Evanston, IL. She also holds an appointment in the School of Education at Northwestern. She holds a patent on the method of simulation used to design the pelvic exam simulator and is currently engaged in the design of other simulators using similar technology. She is also working with the National Board of Medical Examiners (NBME) to support their interests in using her simulators as assessment tools on the United States Medical Licensing Examination. She holds an appointment at the Telemedicine and Advanced Technology Research Center (TATRC) as Special Assistant to the Director.