Surgical Task and Skill Classification from Eye Tracking and Tool Motion in Minimally Invasive Surgery

Narges Ahmidi[1], Gregory D. Hager[2], Lisa Ishii[3], Gabor Fichtinger[1], Gary L. Gallia[4], and Masaru Ishii[3]

[1] Queen's University, Kingston, ON K7L3N6, Canada
{narges,gabor}@cs.queensu.ca
[2] Johns Hopkins University, Baltimore, MD 21211
hager@cs.jhu.edu
[3] Johns Hopkins Medical Institutions, Baltimore, MD 21287
{learnes2,mishii3}@jhmi.edu
[4] Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD 21287
ggallia1@jhmi.edu

Abstract. In the context of minimally invasive surgery, clinical risks are highly associated with surgeons' skill in manipulating surgical tools and their knowledge of the closed anatomy. A quantitative surgical skill assessment can reduce faulty procedures and prevent some surgical risks. In this paper focusing on sinus surgery, we present two methods to identify both skill level and task type by recording motion data of surgical tools as well as the surgeon's eye gaze location on the screen. We generate a total of 14 discrete Hidden Markov Models for seven surgical tasks at both expert and novice levels using a repeated *k*-fold evaluation method. The dataset contains 95 expert and 139 novice trials of surgery over a cadaver. The results reveal two insights: eye-gaze data contains skill related structures; and adding this info to the surgical tool motion data improves skill assessment by 13.2% and 5.3% for expert and novice levels, respectively. The proposed system quantifies surgeon's skill level with an accuracy of 82.5% and surgical task type of 77.8%.

1 Introduction

The performance of a minimally invasive surgery highly depends on surgeons' dexterity in using surgical tools and their knowledge of the anatomy. This fact highlights the significance of Objective surgical skill evaluation.

A procedure of Functional Endoscopic Sinus Surgery (FESS) involves inserting an endoscope with a tiny camera on the end into the sinus cavity to provide the surgeon with a clear view of the surgical field and the ability to use instruments in treating the pathology. The surgeon's performance is limited due to indirect observation of the anatomy and inflexibility of the tools' movements inside the sinus cavity. FESS involves even higher risks due to the sinus' close proximity to the brain, major arteries and critical tissues such as optic nerves.

A skill evaluation system reveals the characteristics hidden in motion data to accurately associate given test sequences to their corresponding skill level. They have

T. Jiang et al. (Eds.): MICCAI 2010, Part III, LNCS 6363, pp. 295–302, 2010.

been assessed in many studies by either tracking the surgeon's body motion in the operating room [1] or hand motion while performing a specific surgical task [2][3][4]. The Imperial College Surgical Assessment Device (ICSAD) system tracks the surgeon's hand motions during a surgery using electromagnetic (EM) markers [4]. In their system, they use a simple feature vector of the number of movements, hand speed, and procedure time to define an observed motion. Authors in [5] propose a system for scoring an image-guided percutaneous needle-based surgery by a feature vector of the successful trials, distance to target, and number of needle retractions. Here, we focus on the analysis of kinematic parameters of motion including translation and rotation of both the tool and the camera.

In laparoscopic surgeries, skill level is evaluated by measuring force and motion data [3]. The promising results with tele-operated robotic systems [2][6][7] show that Hidden Markov Models (HMM) enable us to recognize skill level and subtasks from motion data. Results in [8] show that rotated view of camera in laparoscopic surgeries increases the complexity of the task. In this paper, motion data is not recorded from a robotic system. We collect surgical tool motion data by attaching EM sensors to them.

An infrared-based eye tracking system can be used to measure the gaze position as an important factor in skill evaluation [9]. These trackers are used in other fields such as psychology [10]. Eye-gaze information plays a key role in eye-hand coordination, performance and adjustment of the surgical tools. To the best of our knowledge there is no prior published work in surgical skill evaluation featuring this system.

Our work differs from previous studies in that we generate 14 models (two skill levels in seven surgical tasks) and used them to recognize skill level and task type. Motions of the surgical tools and the surgeon's eye-gaze are recorded while performing different FESS tasks. We address the two following questions: First, is there any skill indicative structure in surgeon's eye-gaze motions? Second, how significant is the addition of eye-gaze data to surgical skill evaluation performance?

2 Experiment Setup

In this experiment, subjects are asked to find and touch a given anatomy inside the sinus cavity of a cadaver by using an endoscope and a nasal pointer. A group of seven anatomical targets are defined: Anterior Genu (AG), Eustachian Tube (ET), Fossa of Rosenmuller (FR), Opto Carotid Recess (OCR), Optic Nerve (ON), Pituitary Gland (PIT), and Superior Turbinate (ST). Figure 1 is the block diagram of the proposed experiment which illustrates the data collection setup and the proposed methodology.

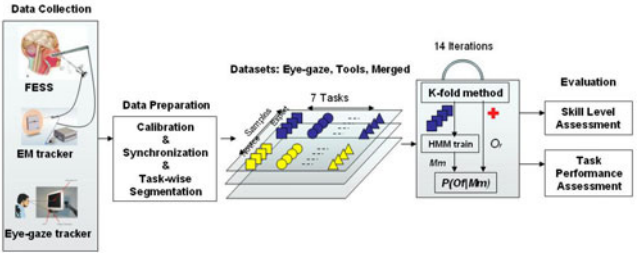

Fig. 1. System block diagram

Motion data from different trials of surgical tasks were collected from a group of 11 different subjects in two levels: 5 expert and 6 novices (Table 1). We defined an expert as a surgeon possessing knowledge of sinus anatomy structure and operation of the endoscope. Our novice subjects were those with no prior endoscopic experience.

Each subject performs two trials of surgeries. Each trial consists of 14 tasks: two sets of all seven tasks in random order. The duration of tasks vary between 5 to 46 seconds. We discard tasks involving irregular procedures (i.e. cleaning the endoscope tip, or leaving tools idle on the bed). The number of trials for each task in two possible skill levels is listed in the rows of Table 1. A total of 95 tasks are collected from expert surgeons and 139 tasks from novice surgeons.

The motion data of each trial is recorded using two trackers and a video stream: (1) The EM tracker collects motion data of both the endoscope and the nasal pointer at a frequency of 40 Hz, using two 5-DOF coil sensors attached to them. Sixteen variables are recorded per frame: a time stamp, a 3D translation and the four parameters of a rotation quaternion per sensor. (2) The eye-gaze tracker records 2D eye-gaze locations on the monitor (800x1200 pixel) at a frequency of 50 Hz. Three parameters are recorded per frame: a time stamp and a 2D eye-gaze location on the screen. (3) The video stream (352x240 pixel, 30 fps) is recorded from the endoscope tower.

Table 1. Number of performed tasks per skill level

Skill \ Task	AG	ET	FR	OCR	ON	PIT	ST
Expert	11	14	14	14	11	14	17
Novice	15	19	21	24	16	23	21

Later, we perform a pivot calibration for EM-tracker dataset and an eye-gaze calibration in order to register identified pupil position to the ground truth. In this experiment we initially study skill information of each of these trackers individually and then assess their aggregated performance.

3 Methodology: Hypotheses and Tests

In this paper we answer the aforementioned key questions: First, is there any skill indicative structure in surgeon's eye-gaze motions? Second, how significant is the addition of eye-gaze data to surgical skill evaluation performance?

We run two tests to find the answer for each question: a *Skill Level Assessment* (SLA) and a *Task Performance Assessment* (TPA). The former evaluates the skill level of the surgeon given a particular task, while the latter identifies the surgical task for a given expertise level. An additional *pre-observed test* is run on each dataset to verify that the generated models recognize their trained sequences properly. Referring to Figure 1, 14 configurations of SLA and 49 trials of TPA are carried out for a given skill level.

3.1 SLA: Skill Level Assessment

The intent of this test is to identify skill level of a given known task. To do so, we evaluate different trials of the task against HMMs of both expertise levels and measure True Positives, True Negatives, False Positives, and False Negatives ratios in a cross validation context. The Positives and Negatives are associated to expert and

novice skill level, respectively. Our first hypothesis is True Positives and True Negatives are higher than False Positives and False Negatives for each model. The more the disparity they have, the higher the skill level performance is.

3.2 TPA: Task Performance Assessment

We use this test to recognize the type of the performed task for a given expertise level. We find the most similar task to the test task by evaluating it against all HMMs in the same expertise level. A noise model REJ is generated from random parts of the datasets to measure the False Rejection Probability. Our second hypothesis is that the performance of TPA is high when we compare the test task against its matching model and is low when compared to other tasks models.

3.3 Data Preparation

To prepare motion data for training the task models, a *task-wise segmentation* is run on the synchronized datasets. We are able to split the videos based on the performed tasks. We need to mark the motion data to help synchronize with the corresponding video. During the clinical procedures, subjects are asked to look at a fixed point in the surgical field while touching it with the pointer tip for a few seconds in between the trials. To extract the fixed points in the motion data, we follow two assumptions: (1) the tool-tip is not in motion, and (2) the endoscope does not move significantly while touching the fixed points. The gradient of the motion is zero for a stationary object.

In equation 1, the function *f(x)* is used to extract those stationary moments in the tool and endoscope motion datasets. The function *g(x)* is a binary signal which represents the same moments in the video stream. Both *f* and *g* are variables of time and *S* is the convolution of fixed-point moments in the video and the tools. The global maxima of *S* is used to sync the tool motion data with the performed tasks in the video. We carry out the same procedure for the eye-gaze dataset to synchronize it with the performed task. This allows us to segment all of the datasets task-wise.

$$S = g(video) * \left[f(\nabla \vec{endoscope}).f(\nabla \vec{tool}) \right]$$
$$f(x) = \begin{cases} 1 & x \approx 0 \\ 0 & otherwise \end{cases} \qquad g(x) = \begin{cases} 1 & x = fixpo\,\text{int} \\ 0 & otherwise \end{cases} \tag{1}$$

In a clinical procedure, an expert surgeon tries to avoid critical tissues inside the sinus cavity by constantly monitoring them on the screen while holding the tools away from them inside the sinus. This leads to a group of imaginary points in the surgical field which constrain the surgeon's path toward the desired anatomical target. To discover the path, we apply *k-means* clustering algorithms to the motion datasets. We use a range of 4-9 for variable *k* in tool dataset, and 2-6 in eye-gaze dataset to determine the number of clusters with higher accuracy. Each frame of motion sequences is replaced by its corresponding cluster.

3.4 Evaluation Method

The HMMs are generated and tested using a *k*-fold method. The *k*-fold is run 14 times by changing the let-out task randomly. Each model is trained using the Baum-Welch algorithm with 100 iterations and an error tolerance of 0.01.

The HMMs are evaluated using Equation 2 to classify a test sequence. Probability function P is the log likelihood of an observation sequence (O_f) to a given model M_m, where the set O is generated from the trials of k-fold method. Comparing the resulting probabilities, the model C with the highest log likelihood is taken as the most probable source for that observed task. Then, we measure the percentage of similarity between the given test sequence O and a given HM model M_m by counting the number of identified models (Function V).

$$\forall O_f \in O \quad C(f) = \max_{m=1}^{8} P(O_f | M_m) \quad 1 < f < |O|$$
$$V(O \mid M_m) = \sum_{f=1}^{|O|} [C(f) == m] / |O| \tag{2}$$

4 Results and Discussions

Question 1: Is there any skill indicative structure in the surgeon's eye-gaze motions? As explained in the previous section, we execute both *SLA* and *TPA* tests for each skill level and performed task. Table 2 shows the results of skill level recognition for a given task. Additionally, Table 3 shows the accuracy of recognizing the type of the performed task for a given expertise level. To test the accuracy of the HMMs, the *pre-observed test* is run for each level of expertise and a result of 100% recognition is achieved for all the following datasets.

Our first hypothesis is confirmed by comparing the result of each column in Table 3. True Positives and True Negatives are significantly larger than False Positives and False Negatives, except for the expert level task PIT which is misclassified as the novice level. However Table 3 indicates that task PIT can be successfully classified at a given expertise level. The disparity between False and True classification in Table 2 shows that surgeon eye-gaze data includes structure for skill-level recognition.

Table 2. Skill Level Assessment of a given task, using eye-gaze dataset

Task / Skill	AG	ET	FR	OCR	ON	PIT	ST
TPR	85%	95%	88%	87%	75%	42%	82%
FNR	15%	5%	12%	13%	25%	58%	18%
FPR	12%	25%	11%	4%	23%	7%	19%
TNR	88%	75%	89%	96%	77%	93%	81%

Table 3. Performed Task Assessment for a given skill level, using eye-gaze dataset

Skill level	Expert								Novice							
Trained / Test	AG	ET	FR	OCR	ON	PIT	ST	REJ	AG	ET	FR	OCR	ON	PIT	ST	REJ
AG	91.7	-	-	8.3	8.3	-	-	-	85.8	7.1	-	-	7.1	-	21.4	7.1
ET	-	100	-	-	8.3	8.3	-	8.3	-	85.8	-	-	7.1	-	-	7.1
FR	-	-	75	-	8.4	8.3	-	-	-	-	100	-	7.1	-	-	7.1
OCR	-	-	8.3	75	-	-	8.3	-	7.1	-	-	92.9	7.1	-	-	-
ON	-	-	-	-	50	-	-	-	-	-	-	-	64.5	-	-	-
PIT	8.3	-	8.3	-	-	75	-	-	-	-	-	-	7.1	100	14.3	-
ST	-	-	8.4	16.7	-	8.4	92.7	-	7.1	-	-	-	-	-	64.3	-
Reject (REJ)	-	-	-	-	25	-	-	91.7	-	7.1	-	7.1	-	-	-	78.7
%accuracy	91.7	100	75	75	50	75	92.7	91.7	85.8	85.8	100	92.9	64.5	100	64.3	78.7

The columns of Table 3 show that the performed task is always classified to the correct HMM (the diagonal of each square repeated in last row). The notable difference between the recognized models and the REJ model confirms the second hypothesis that eye-gaze data can be used for task evaluation as well.

Question 2: Can eye-gaze data improve surgical skill evaluation?
First, we execute SLA and TPA for the motion data of the tools to measure the accuracy of the system. Then, we make a new system of HMMs by combining both eye-gaze and tool datasets. To quantify the improvement achieved, we compare the performance of these two systems.

- *System1: skill evaluation using tool datasets*

Tables 4 and 5 show the results for SLA and TPA using EM tracked tool datasets. In Table 4, the columns show that the skill level is recognized correctly for each task, except for task AG. It reveals that both expert and novice participants perform that task with the same level of expertise. Carotid artery is one of the largest objects in the nose and one of the most prominent so we would expect it to be the easiest to find. It is also the most understood even by novices, since they are taught specifically from the start to always identify this structure. The anatomy was distorted by removing part of the skull base (planum) to make the task harder. Therefore, some subjects misidentified structures.

Table 4. Skill Level Assessmet of a given task, using tools dataset

Skill \ Task	AG	ET	FR	OCR	ON	PIT	ST
TPR	50%	79%	75%	80%	75%	82%	73%
FNR	50%	21%	25%	20%	25%	18%	27%
FPR	31%	18%	11%	25%	11%	11%	25%
TNR	69%	82%	89%	75%	89%	89%	75%

Table 5. Performed Task Assessment for a given skill-level, using tools dataset

Skill level	Expert								Novice							
Trained \ Test	AG	ET	FR	OCR	ON	PIT	ST	REJ	AG	ET	FR	OCR	ON	PIT	ST	REJ
AG	100	43	50	28.6	42.9	35.7	14.4	50	64.4	7.1	-	7.1	-	7.1	-	14.3
ET	-	57	-	-	-	-	-	7.1	14.3	64.4	-	7.1	-	7.1	-	7.1
FR	-	-	50	-	-	-	7.1	-	-	14.3	78.6	-	7.1	7.1	7.1	-
OCR	-	-	-	64.3	-	-	7.1	-	7.1	-	-	71.6	-	-	7.1	-
ON	-	-	-	-	50	-	7.1	-	7.1	7.1	-	7.1	78.7	-	21.4	-
PIT	-	-	-	-	-	64.3	21.4	-	-	-	7.1	-	7.1	78.7	-	-
ST	-	-	-	7.1	7.1	-	42.9	7.1	7.1	-	-	7.1	-	-	50	-
Reject (REJ)	-	-	-	-	-	-	-	35.8	-	7.1	14.3	-	7.1	-	14.4	78.6
%accuracy	100	57	50	64.3	50	64.3	42.9	35.8	64.4	64.4	78.6	71.6	78.7	78.7	50	78.6

Overall accuracy of the models in skill level detection is 73.4% and 81.1% for expert and novice surgeons, respectively (Table 8). The columns of Table 5 show that tasks are always recognizable for novice datasets. The overall identification rate for task type recognition is 58.1% and 70.6% for expert and novice groups, respectively.

- *System2: skill evaluation using combined dataset*

Tables 6 and 7 show the results of executing SLA and TPA over the merged dataset. Table 8 shows that expertise level is identifiable in 82.9% of cases for expert surgeons and 82% for novice surgeons. The columns of Table 7 show that all of the tasks in the same level of expertise are recognized correctly.

Table 6. Skill Level Assessment of a given task, using merged dataset

Task / Skill	AG	ET	FR	OCR	ON	PIT	ST
TPR	50%	89%	89%	85%	89%	85%	93%
FNR	50%	11%	11%	15%	11%	15%	7%
FPR	46%	13%	13%	13%	10%	13%	18%
TNR	54%	87%	87%	87%	90%	87%	82%

Table 7. Performed Task Assessment for a given skill-level, using merged dataset

Skill level	Expert								Novice							
Trained / Test	AG	ET	FR	OCR	ON	PIT	ST	REJ	AG	ET	FR	OCR	ON	PIT	ST	REJ
AG	100	27.3	27.3	27.3	18.2	27.3	36.4	36.3	71.4	-	-	-	-	-	-	-
ET	-	72.7	-	-	--	-	-	-	-	78.6	-	-	-	-	-	-
FR	-	-	72.7	-	-	-	-	-	-	-	78.6	-	-	-	-	-
OCR	-	-	-	72.7	-	-	-	-	-	-	-	71.4	-	-	-	-
ON	-	-	-	-	81.8	-	-	-	-	--	-	-	78.6	-	-	-
PIT	-	-	-	-	-	72.7	-	-	-	-	-	-	-	71.4	-	-
ST	-	-	-	-	-	-	63.6	-	-	-	-	-	-	-	85.7	-
Reject (REJ)	-	-	--	-	-	-	-	63.6	-	-	-	-	-	-	-	78.6
%accuracy	100	72.7	72.7	72.7	81.8	72.7	63.6	63.6	100	78.6	78.6	71.4	78.6	71.4	85.7	78.6

Table 8. Performance improvement in skill assessment adding eye-gaze dataset

Test	Skill	Surgical tools dataset	Merged with eye-gaze dataset	Performance Improvement
Avg TPA	Expert	58.1%	74.98%	16.9%
	Novice	70.7%	80.36%	9.7%
Avg SLA	Expert	73.4%	82.9%	9.5%
	Novice	81.1%	82%	0.9%

Each row in Table 8 is the average result of the corresponding method over the related models (an average of first and last row of SLA, and last row of TPA). Comparing the performance of the two systems reveals that adding eye-gaze information improves SLA performance by 9.5% for expert and 0.9% for novice surgeons. The average performance of TPA is 75% and 80.4% for expert and novice surgeons, respectively. This indicates that the new system improves the TPA performance by 16.9% for experts and by 9.7% for novice surgeons.

5 Conclusion

Precision of sinus surgery is critical due to its close proximity to the brain, major arteries and critical tissues. The work presented here is a promising statistical method to assess skill of a surgeon while performing a Functional Endoscopic Sinus Surgery. We make HMM models for seven different surgeries in two level of expertise using the eye-gaze locations and the surgical tools motions. Two metrics of SLA and TPA access skill level and task type of a given surgery. Results reveal that eye-gaze data contains skill related structures; and combining it with the surgical tool motion data improves the classifier performance. The proposed system improves SLA performance 9.5% for experts and 0.9% for novices, on average. Besides, TPA performance is improved 16.9%

for experts and 9.7% for novice surgeons. The proposed system quantifies surgeon's skill level with an accuracy of 82.5% and surgical task type of 77.8%.

Currently, we are exploring the methods to improve performance of the proposed skill assessment techniques. One is to unify coordination systems of the tools and eye-gaze motions by registering them to the CT image volume of the target anatomy.

Acknowledgments

The authors gratefully acknowledge support from NIH 5R01CA118371, NSF CDI-0941362 and NSF CPS 0931805 as well as Johns Hopkins internal funding. They would like to thank Daniel Abretske, Daniel Mirota, and Kelleher Guerin for their time and help with the data collection.

References

1. Padoy, N., Mateus, D., Weinland, D., Berger, M.O., Navab, N.: Workflow Monitoring based on 3D Motion Features. In: ICCV Workshop on Video-oriented Object and Event Classification, Kyoto, Japan (September 2009)
2. Lin, H.C., Shafran, I., Yuh, D., Hager, G.D.: Towards Automatic Skill Evaluation: Detection and Segmentation of Robot-Assisted Surgical Motions. Computer Aided Surgery 11(5), 220–230 (2006)
3. Rosen, J., Brown, J.D., Chang, L., Sinanan, M., Hannaford, B.: Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Trans. Biomed. Eng. 53(3), 399–413 (2006)
4. Munz, Y., Almoudaris, A., Moorthy, K., Dosis, A., Liddle, A., Darzi, A.: Curriculum-based solo virtual reality training for laparoscopic intracorporeal knot tying: objective assessment of the transfer of skill from virtual reality to reality. The American Journal of Surgery 193(6), 774–783 (2007)
5. Ahmidi, N., U-Thainual, P., Vikal, S., Mousavi, P., Iordachita, I., Fichtinger, G.: A System for Performance Analysis of Surgeon Dexterity in Percutaneous Needle-based Interventions, Imaging Network Ontario. University of Toronto, Canada (2008)
6. Reiley, C., Hager, G.D.: Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 435–442. Springer, Heidelberg (2009)
7. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.D.: Data-Derived Models for Segmentation with Application to Surgical Assessment and Training. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 426–434. Springer, Heidelberg (2009)
8. Leong, J.J., Nicolaou, M., Atallah, L., Mylonas, G.P., Darzi, A.W., Yang, G.Z.: HMM assessment of quality of movement trajectory in laparoscopic surgery. Comput. Aided Surg. 12, 335–346 (2007)
9. Nicolaou, M., James, A., Darzi, A., Yang, G.Z.: A Study of Saccade Transition for Attention Segregation and Task Strategy in Laparoscopic Surgery. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3217, pp. 97–104. Springer, Heidelberg (2004)
10. Martín-Loeches, M., Schacht, A., Casado, P., Hohlfeld, A., Abdel Rahman, R., Sommer, W.: Rules and heuristics during sentence comprehension: evidences from a dual-task brain potential study. Journal of Cognitive Neuroscience 21(7), 1380–1395 (2009)