Skill Assessment for Robotic Surgery: Language of Surgery

Carol Reiley  
Johns Hopkins University  
Advised by: Gregory D. Hager  
4/19/2010
Thesis

Robotic Minimally Invasive Surgery (RMIS) has the potential revolutionize our understanding of modeling, teaching and evaluating human manipulation skills.
Skill Learning On Robotic Surgery

• Our goal: develop a method for objective evaluation of technical skill in surgery
  – Analyze motion to better understand surgical skill using segmented motion blocks (surgemes).
  – Classify different users with various skill and experience
  – Provide feedback in an intuitive and inexpensive method
How Is Surgery Taught?

Sir William Halsted, JHU 1889
Apprentice style graded responsibility
“see one, do one, teach one”
An October 8, 2003 JAMA study from the U.S. government’s Agency for Healthcare Research and Quality (AHRQ) documented 32,000 mostly surgery-related deaths costing $9 billion and accounting for 2.4 million extra days in the hospital in 2000.
We MUST!
Richard Reznick, “Teaching Surgical Skills - Changes in the Wind”. NEJM 2006

- Pressures from government and insurance companies to reduce cost of deaths due to iatrogenic causes.
- Economic pressures on medical schools to reduce the costs of training surgeons.
- All within new labor laws of limiting resident work hours.

| Table 1: Types of laparoscopic instruction incorporated into surgical skills laboratories |
|-------------------------------------------------|-----------------|-----------------|-----------------|
| Answer option | Not helpful (%) | Very helpful (%) | N (%) of responses* |
| Observation of procedure by instructor | 21 | 79 | 183 (91) |
| Viewing of instructional videos | 33 | 67 | 166 (82) |
| Discussion of instrumentation and laparoscopic theory | 21 | 79 | 164 (81) |
| Basic dissection techniques | 13 | 87 | 174 (86) |
| Basic intracorporeal suturing techniques | 3 | 97 | 195 (97) |
| Use of surgical simulators | 11 | 89 | 176 (87) |
| Live animal wet labs | 7 | 93 | 170 (84) |

*Number who answered question = 202

LAPAROSCOPIC SKILL ACQUISITION IN THE SURGICAL LABORATORY: A NATIONAL SURVEY OF GENERAL SURGERY RESIDENTS. VERGIS, QURESHI, JIMENEZ, GREEN, PRYOR, SCHLACHTA, A. OKRAINEC Open Medicine 3(3).

Copyright GD Hager, 2010
Methods for Dexterous Assessment

Objective Structured Assessments of Technical Skills (OSATS)

Figure 1. Examples of OSATS Stations.
Examinees rotate through multiple stations, where they perform elements of surgical tasks and are graded by expert examiners using global rating forms and task-specific checklists. These examples are drawn from an “inventory” of more than 40 such stations.
Dexterity Assessment In Simulation

Imperial College Surgical Assessment Device
Darzi et al., uses electromagnetic markers to track a subject’s hands during a standardized task.

Minimally Invasive Surgical Trainer - Virtual Reality
Movements of two standard laparoscopic instruments are tracked. Low level analysis of positions, forces and times.

Copyright GD Hager, 2010
Dexterity Assessment in Live Surgery
Why Robotic Surgery?

Manipulation
Data for Free!
Four Questions

- What data can we acquire for assessing or improving training and evaluation?
- How do we model surgical technique from empirical data?
- How do we evaluate and/or impart skill?
- How do we effectively validate these results?
Sample Motion Data From RMIS

- Benchtop surgical tasks
- 72-192 motion variables recorded from API
- 5-15 trials/user
Modeling: What Do Time and Motion Tell Us?
Modeling Structure Using Gestures


Copyright GD Hager, 2010
Suturing Gesture Vocabulary

1. Reach for needle
2. Position needle
3. Insert and push needle through tissue
4. Move to middle with needle (left hand)
5. Move to middle with needle (right hand)
6. Pull suture with left hand
7. Pull suture with right hand
8. Orient needle with both hands

Copyright GD Hager, 2010
Testing and Training Process

1. Training Data → Learned Time-series Model → Labels
2. Testing Data → Learned Time-series Model → Estimated Labels
3. Correct Labels → Accuracy Assessment

Copyright GD Hager, 2010
## Classifier vs Manual Segmentation


<table>
<thead>
<tr>
<th>n</th>
<th>Number of labeled classes</th>
<th>LDA output dimensions</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>3</td>
<td>91.26</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
<td>91.46</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
<td>91.14</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>3</td>
<td>91.06</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>91.34</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>3</td>
<td>92.09</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>4</td>
<td>91.92</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3</td>
<td>91.88</td>
</tr>
</tbody>
</table>

**Expectation Maximization + Bayes Classifier**
Classifier vs Segmentation

Manual Segmentation vs Classifier (92.92% correct)

- Blue solid - classifier
- Red dashed - manual segmentation

Copyright GD Hager, 2010
# Multi-User, Multi-Task Data Collection

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Medical Training</th>
<th>Da Vinci Training</th>
<th>Hours?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>-</td>
<td>&lt;1</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>10-15</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>100+</td>
</tr>
<tr>
<td>D</td>
<td>X</td>
<td>X</td>
<td>100+</td>
</tr>
<tr>
<td>E</td>
<td>X</td>
<td>X</td>
<td>100+</td>
</tr>
<tr>
<td>F</td>
<td>-</td>
<td>X</td>
<td>100+</td>
</tr>
<tr>
<td>G</td>
<td>-</td>
<td>X</td>
<td>&lt;10</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>X</td>
<td>&lt;10</td>
</tr>
<tr>
<td>I</td>
<td>-</td>
<td>-</td>
<td>&lt;1</td>
</tr>
</tbody>
</table>

Knot Tying  
Needle Passing  
Suturing
Expanded Vocabulary

(1) Initial idle position
(2) Reaching for needle
(3) Positioning Needle
(4) Inserting Needle/Push needle through tissue
(5) Transferring needle from left to right

(6) Moving to center with needle in gripper
(7) Pulling suture with left
(8) Orienting needle
(9) Right hand helps tighten suture
(10) Loosening more suture

(11) End task
(12) Reaching for needle with left
(13) Making C loop around right instrument
(14) Right hand reaches for suture
(15) Both hands pull

Copyright GD Hager, 2010
Results of Supervised State Labeling

Data-Derived Models for Segmentation with Application to Surgical Assessment and Training, B Varadarajan, Carol Reiley, H Lin, S Khudanpur, G Hager, Proc. MICCAI 2009

<table>
<thead>
<tr>
<th>LDA Dimension</th>
<th>Setup I</th>
<th>Setup II</th>
<th>Setup III</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>83%</td>
<td>82%</td>
<td>73%</td>
</tr>
<tr>
<td>15</td>
<td>86%</td>
<td>82%</td>
<td>71%</td>
</tr>
<tr>
<td>20</td>
<td>87%</td>
<td>83%</td>
<td>70%</td>
</tr>
</tbody>
</table>

- Multi-State SLR HMM per gesture
  - New notion of “dexeme”

- HLDA
  - A discriminative projection per state in the HMM

Copyright GD Hager, 2010
Can We Learn a Vocabulary?

- Start with a one-state HMM (N=1)
- **Concurrently** split each state into four
- *Choose N+Δ* states that maximize likelihood
- Continue until a desired number of states
What Gesture is HMM state #2?
What about HMM state #10?
Skill Assessment: Beyond Time and Motion

• Can we detect interesting differences in categories of users?
  – Experiment 0: Accuracy of skill classification

• Can we do so at the surgeme level?
  – Experiment 1: Surgeme level vs. Task Level HMM

• Does labeling matter?
  – Experiment 2: Task Level HMM with known states vs. Task Level HMM with unknown states
Task Level Evaluation

- Build one statistical model of each skill level for each trial using instrument velocities

- Can apply to labeled (KS) or unlabeled (BW) data

Motion Models
1 per skill level

Test Data

Compare Models to Data

Skill Category

Copyright GD Hager, 2010
Task Level Evaluation

• Trained 3 skill level models for each surgeme.

• Test each sequence of surgemes and return vote of most labels.

Test Data

Surgeme Models
1 per skill level per surgeme

Compare Models to Data

Skill Category Per Surgeme

Copyright GD Hager, 2010
Database

• 57 trials:
  • 19 from expert,
  • 19 from intermediate,
  • 19 from novice

• Four-throw suturing task
• 1011 total surgeme occurrences
• Average trial 45-130 seconds
Summary Results

• 100% correct classification on surgeme level
• 100% correct classification on task level, labeled data
• 95% correct classification on task level, unlabeled data
• Certain surgemes more indicative of skill than others

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Classification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c: Surgeme (E)</td>
<td>100%</td>
</tr>
<tr>
<td>1c: Surgeme (I)</td>
<td>100%</td>
</tr>
<tr>
<td>1c: Surgeme (N)</td>
<td>100%</td>
</tr>
<tr>
<td>2c: Task BW(E)</td>
<td>84%</td>
</tr>
<tr>
<td>2c: Task BW (I)</td>
<td>100%</td>
</tr>
<tr>
<td>2c: Task BW(N)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS(E)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS (I)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS(N)</td>
<td>100%</td>
</tr>
</tbody>
</table>
Applications Of Motion Models

Underlying hypothesis: Learned motion models of experts can be used for teaching, training, and automation of surgical actions.

Live Input → Automated Modeling → Skill Evaluation → Feedback

Copyright GD Hager, 2010
One Example

- Let the system learn its own model
- Define a distance between the resulting string
- Show the differences

User 1
1 2 3 2 3 3 4 5 6 8 9 9 9 3

User 2
1 2 3 3 2 3 3 3 4 5 8 7 9 9 3

Table 1: The average string distance between surgeons of three different skill levels while performing a 4-throw suture. The strings were created by training an unsupervised HMM. The data shows that experts are more similar and consistent than novices or those of intermediate skill.
Expert-Expert
Expert-Novice
Validation: Multi-site Data Collection

Rajesh Kumar, JHU

- Secure, anonymized, transparent, and systematic collection of procedure data for creation of a longitudinal archive of robotic surgery training with trainees of known surgical and robotic proficiency
- Analysis of system and surgical skill acquisition and identification of key robotic surgery skills
- Development of basic metrics of system operation, unique opportunity for creating methods of standardized assessment

Copyright GD Hager, 2010
Experimental Tasks

• Acquiring data from 4 tasks from users of 4 skill levels

• From robotic surgery training practicum (Intuitive Surgical)

• Experts : 2 data collection sessions

• Other 3 levels (novice/beginner/intermediate):
  – longitudinal data collection
  – 12 regularly spaced sessions over a year
Data Access/Archive

• Secure online archive
  – Two levels of authentication
  – Semantic support for collation/creation of new data sets
  – Easy browser based review
  – Online assessment including OSATS type analysis
Current Status

• Data collection ongoing at 3 centers
  – Currently 16 volunteers
  – Plan for 48 volunteers@ 6 centers

• Preliminary assessment of robot use now possible
  – Master workspace usage between an expert (top) and novice (bottom)

• Tasks metrics development and OSATS analysis now starting
  – Tasks completion times, and some errors can be automatically segmented
  – Learning methods of system skill assessment also in development.
The Future: Beyond Motion Analysis

<table>
<thead>
<tr>
<th>Skill Level</th>
<th>Throws</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>72/80</td>
<td>90.0</td>
</tr>
<tr>
<td>Inter/Novice</td>
<td>68/76</td>
<td>89.5</td>
</tr>
<tr>
<td>All</td>
<td>141/156</td>
<td>90.4</td>
</tr>
</tbody>
</table>

Video Analysis

Mirror Assembly

Eye Tracking in Da Vinci Robot

Copyright GD Hager, 2010
Beyond the Benchtop

How To Understand This Connection?
Detection of gestures in live surgery with 81% accuracy

Towards “Real-time” Tool-tissue Interaction Detection in Robotically Assisted Laparoscopy
Discussion

• Assumption:
  – Human skill can be modeled through HMMs.
  – Experts will use fewer motions and execute task more efficiently

• What we learned
  – Subtask level provides more information
  – Manually labeled surgemes results comparable to unlabeled
Key Publications Thus Far


• Decomposition of Robotic Surgical Tasks: An Analysis of Subtasks and Their Correlation to Skill, MICCAI 2009 workshop (accepted)

• Task Versus Subtask Level Skill Modeling in Robotic Minimally Invasive Surgery, MICCAI 2009 (accepted)
The Future: Beyond Teaching and Training

• Understanding the effect of:
  – Distance
  – Changes in interface
  – Collaboration/expert interaction

• Intelligent Assistance
  – “I know what you’re trying to do”
  – Supervisory interaction rather than “hands on”

• **Fundamental** understanding of human manipulative activity
Many Thanks!

• Collaborators
  – Gregory Hager
  – David Yuh MD
  – Grace Chen MD
  – Rajesh Kumar
  – Rene Vidal
  – Sanjeev Khudanpur

• Students
  – Carol Reiley
  – Henry Lin
  – Balakrishnan Varadarajan
  – Nicolas Padoy
  – Many undergraduate labelers

• Funding
  – NSF IIS 0534359
  – NSF CDI 0941362
  – NSF CPS 0931805
  – NIH R21 EB009143
  – NSF EEC 9731748
  – NSF EEC 0646678
  – NSF MRI 0722943
  – NIH R42 RR019159

• Intuitive Surgical
Suturing Trial

Expert
Suturing Trial

Expert
1. Reach for needle

2. Position needle

3. Insert and push needle through tissue

4. Move to middle with needle (left hand)

5. Move to middle with needle (right hand)

6. Pull suture with left hand

7. Pull suture with right hand

8. Orient needle with both hands
Prior Work on Motion Segmentation

- Example classifier to manual segmentation result

**Single User**

**Multiple Users**
da Vinci Data Output

192 values
34 for each Master manipulator (2)
38 for each Patient manipulator (3)
10 other
23 data packets per second
High-quality stereo vision
Use 14 velocity subset
Methods to Assess Skill

• Descriptive Statistics
• Skill Modeling through Language Modeling
  – Task Based
  – Surgeme Based
Skill Modeling

- Data filtering and vector quantization techniques to discretize input data
- Build skill models
Task Level Data Preprocessing

• Preprocessing to discrete signals
  – Fast Fourier Transform over each 14 velocity vectors
    • 400 ms sliding window shifted every 200 ms
    • Take lower 4 coefficients
  – K-means of 64 clusters
  – New 56 dim discrete vector
    • (14 velocity channels x 4 coefficients)
Statistical differences between models

• HMMs built for each group (expert, intermediate, novice)
• Which skill model is most likely to generate given test sequence?

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

Or similarly...

\[ D(\lambda_s, \lambda_{test}) = \frac{1}{T_{test}} \min (\xi(\lambda_i, \lambda_{test}), \xi(\lambda_e, \lambda_{test}), \xi(\lambda_n, \lambda_{test})) \]
Experimental Study

- 57 trials: 19 from expert, 19 from intermediate, 19 from novice
- Four-throw suturing task
- 1011 total surgeme occurrences
- Average trial 45-130 seconds
Experiment 1 Results
Surgeme level (1c) vs. Task Level HMM (2c)

- 100% correct classification on surgeme level
- 95% correct classification on task level
- Certain surgemes more indicative of skill than others

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Classification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c: Surgeme (E)</td>
<td>100%</td>
</tr>
<tr>
<td>1c: Surgeme (I)</td>
<td>100%</td>
</tr>
<tr>
<td>1c: Surgeme (N)</td>
<td>100%</td>
</tr>
<tr>
<td>2c: Task BW (E)</td>
<td>84%</td>
</tr>
<tr>
<td>2c: Task BW (I)</td>
<td>100%</td>
</tr>
<tr>
<td>2c: Task BW (N)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS (E)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS (I)</td>
<td>100%</td>
</tr>
<tr>
<td>2a: Task KS (N)</td>
<td>100%</td>
</tr>
</tbody>
</table>
Task decomposition of surgemes according to time
Experiment 2 Results
Task Level HMM with known states (2a) vs. Task Level HMM with unknown states (2c)

• 100% with known states; 94% correct classification with unknown states
• Unlabeled data does almost as well as labeled
0-idle

1-Reaching for needle

2-Position needle

3-Insert Needle

5-Move to center

6-Pull suture

7-Transfer needle

8-Orient needle

0-Drop suture

(a) Exp.KS

(b) Int.KS

(c) Nov.KS

(d) Exp.BW

(e) Int.BW

(f) Nov.BW
a) Suturing

b) Needle Transferring

c) Knot Tying

d) Suturing (live)
Timeline and Deliverables

• Phase 1: Modeling expert surgeons

• Phase 2: Offline skill feedback of humans using robotic system
  – Human subject study evaluating training with feedback versus current techniques

• Phase 3: Skill Learning on a robot
Task Based Method

• Train one statistical model of each skill level for each trial using patient side tool velocities.
• Leave 1 trial out cross validation

Motion Data

\[ x_{(1,t)} \]

\[ m(s) \]

\[ x(t) \]

Data Preprocessing

\[ o_{\text{test}} \]

\[ \lambda_{\text{test}} \]

\[ \lambda_{s1}, \lambda_{s2}, \lambda_{s3} \]

Trained expertise models for task (3)

\[ D(\lambda_{s}, \lambda_{\text{test}}) \]

Skill Label
Surgeme Level Methods

- K-means with 8 clusters
- Trained 3 skill level models for each surgeme. Test each sequence of surgeme and return vote of most labels.
- Leave one trial out cross validation

<table>
<thead>
<tr>
<th>Time/surgeme</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Transcription file

<table>
<thead>
<tr>
<th>Time/surgeme/skill</th>
<th>0</th>
<th>1</th>
<th>exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>5</td>
<td>int</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>exp</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>novice</td>
<td></td>
</tr>
</tbody>
</table>

New transcription file
Classification Methods

• Linear Discriminant Analysis (LDA) with Single Gaussian
• LDA + Gaussian Mixture Model (GMM)
• 3-state Hidden Markov Model (HMM)
• Maximum Likelihood Linear Regression (MLLR)
  – Supervised
  – Unsupervised
Results

- Leave one user out cross validation
- Supervised: Surgeme start/stop events manually defined
- Unsupervised: Surgeme start/stop events automatically derived