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”

Can We Do Better Today?

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


ORIGINAL CONTRIBUTION

Excess Length of Stay, Charges, and Mortality Attributable to Medical Injuries During Hospitalization

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	Response		
	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).

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

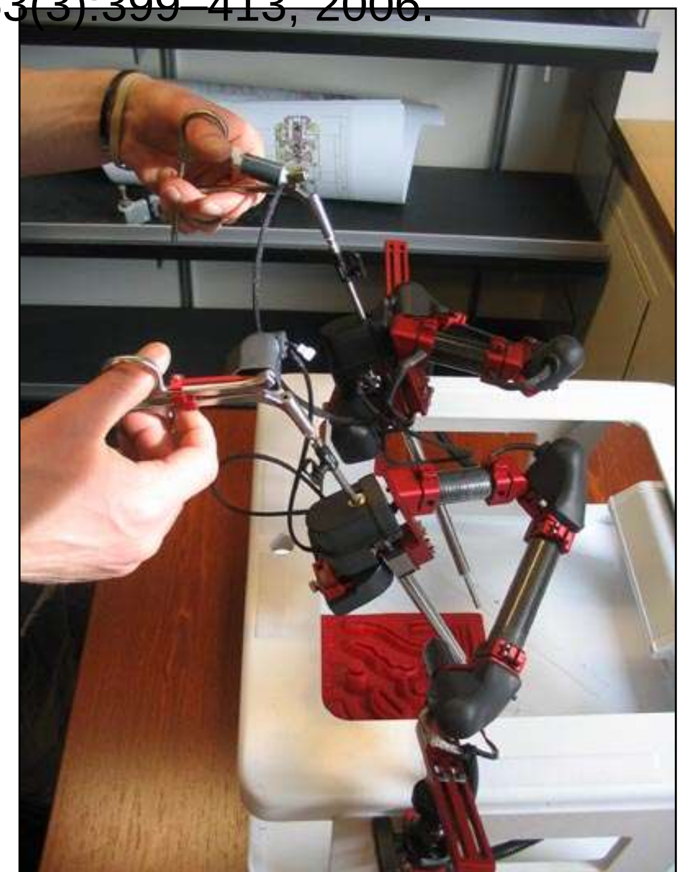
Darzi et al., uses electromagnetic markers
to track a subject's hands during a standardized task

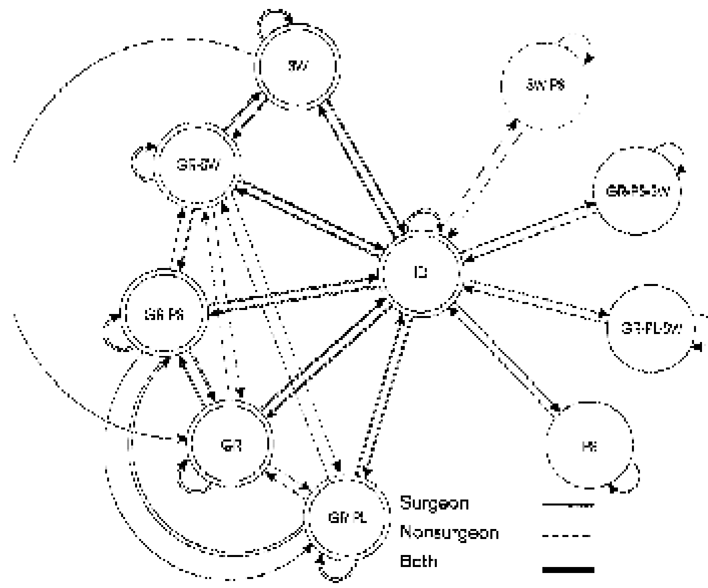
Minimally Invasive Surgical Trainer - Virtual Reality

Movements of two standard laparoscopic instruments are tracked. Low level analysis of positions, forces and times.

Dexterity Assessment in Live Surgery

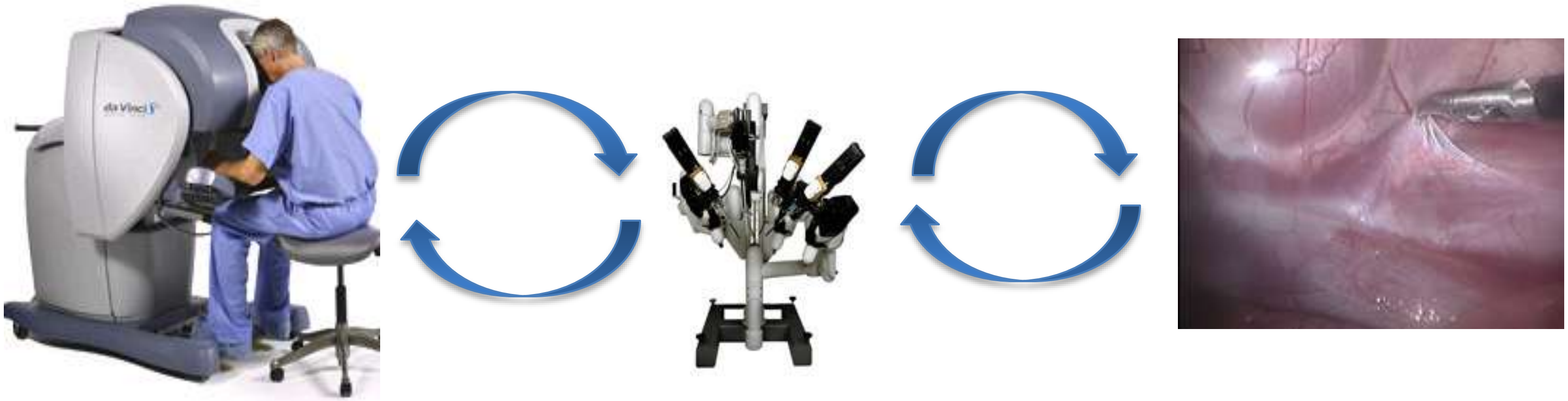
Rosen, Brown, Chang, Sinanan, Hannaford. **Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model.** IEEE Transactions in Biomedical Engineering, 53(3):399–413, 2006.



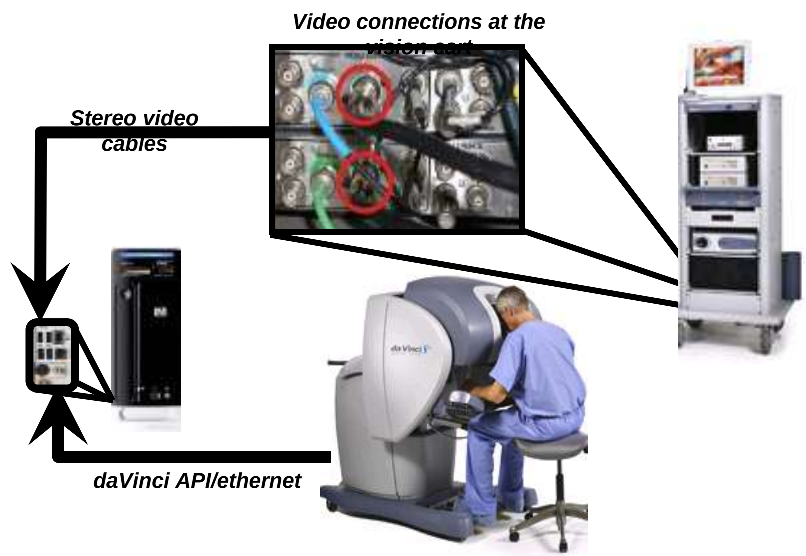


Courtesy of [BioRobotics Lab, University of Washington]

Why Robotic Surgery?



Manipulation Data for Free!



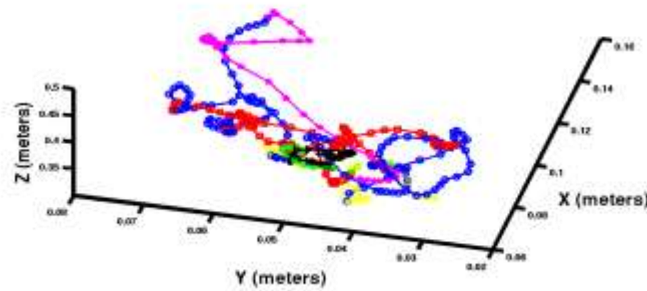
Four Questions

- What data can we acquire for assessing or improving training and evaluation?

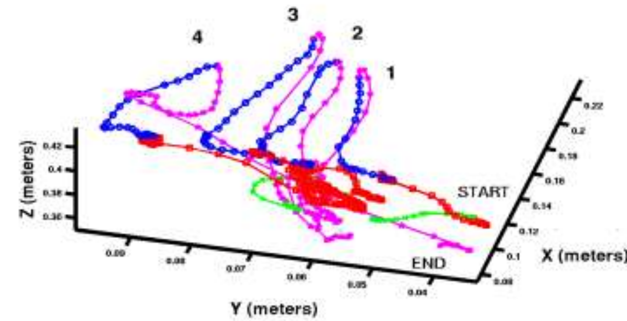
- How do we model surgical technique from empirical data?

- How do we evaluate and/or impart skill?

- How do we effectively validate these results?



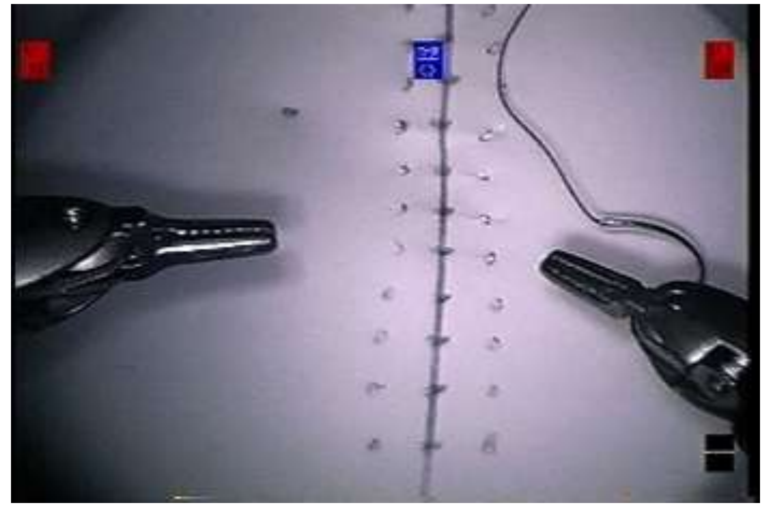
Intermediate Surgeon - trial 22

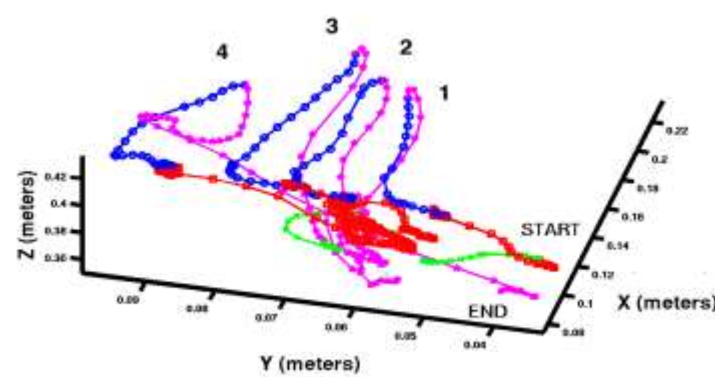


Expert Surgeon - trial 4

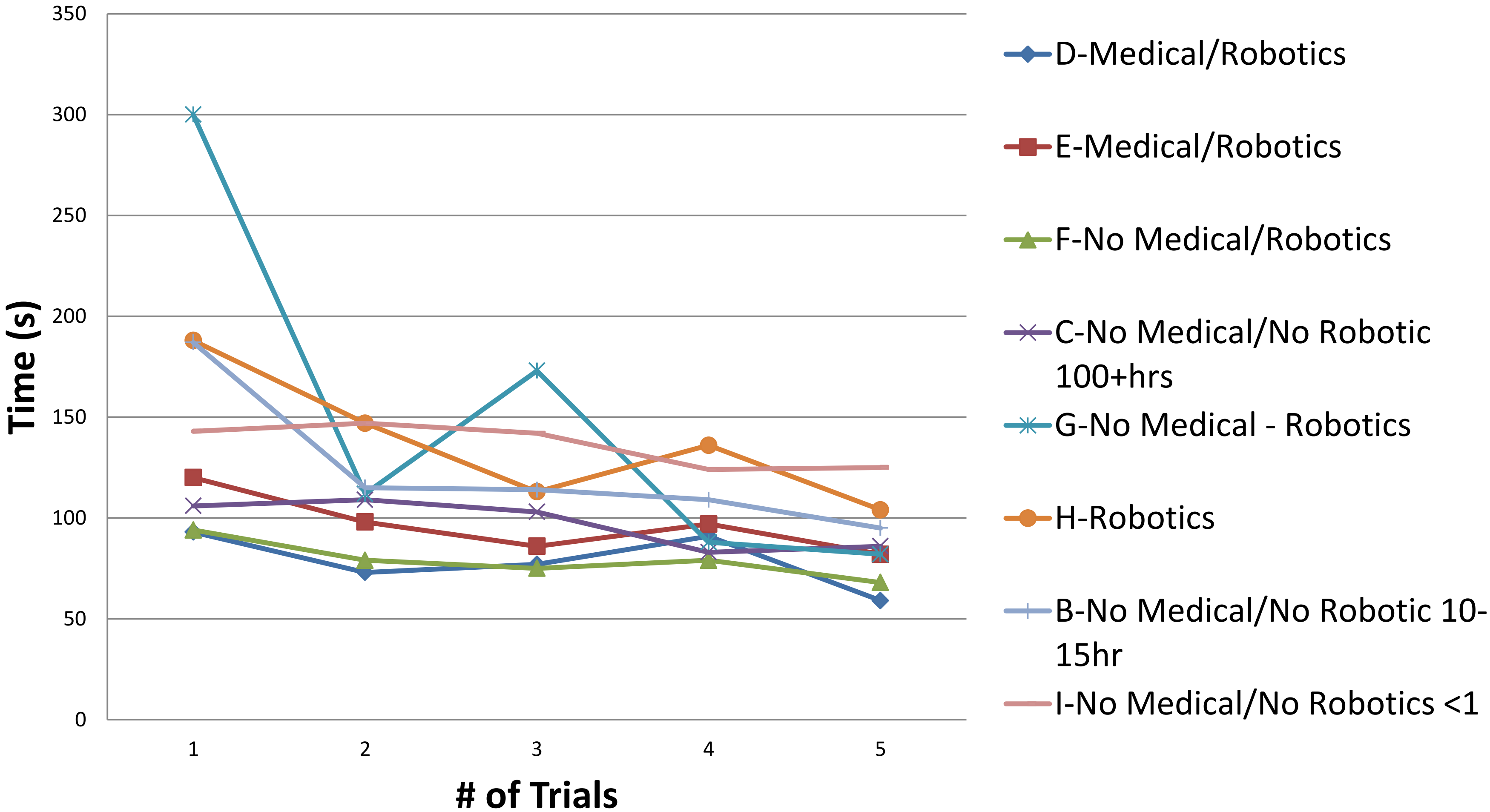
Sample Motion Data From RMIS

- Benchtop surgical tasks
- 72-192 motion variables recorded from API
- 5-15 trials/user



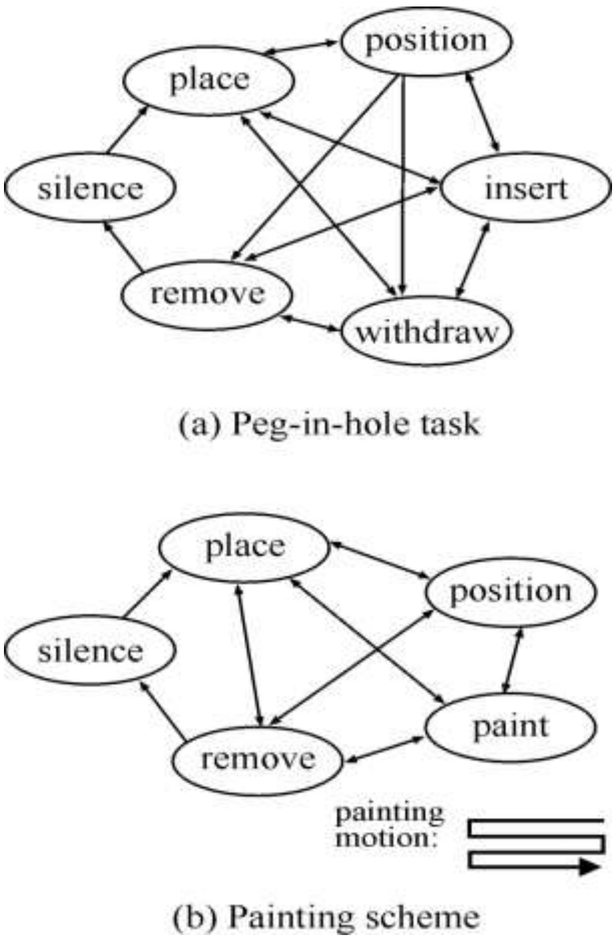


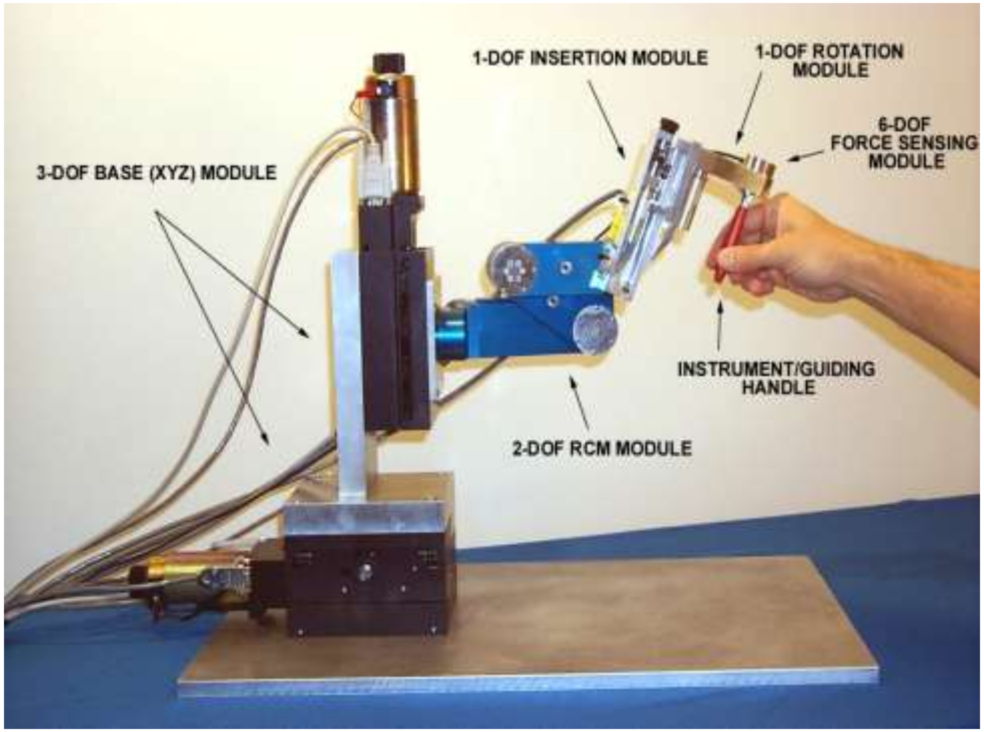
Modeling: What Do Time and Motion Tell Us?



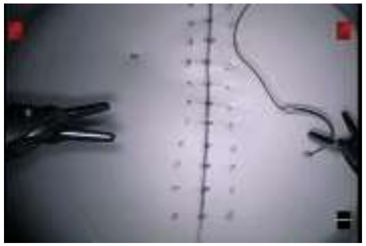
Modeling Structure Using Gestures

S. Hundtofte, A. Okamura, and G. Hager. Building a task language for segmentation and recognition of user input to cooperative manipulation systems. In *Proc. 10th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pages 225-230, 2002*

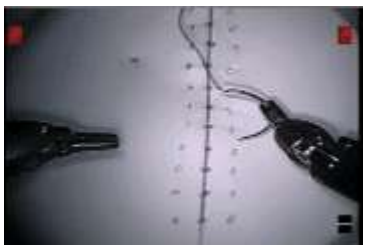




Suturing Gesture Vocabulary



1. Reach for needle

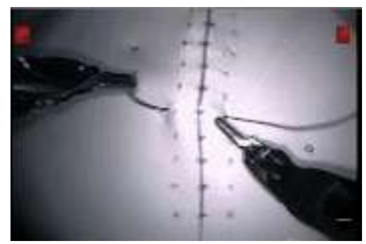


2. Position needle

3. Insert and push needle through tissue

4. Move to middle with needle (left hand)

5. Move to middle with needle (right hand)

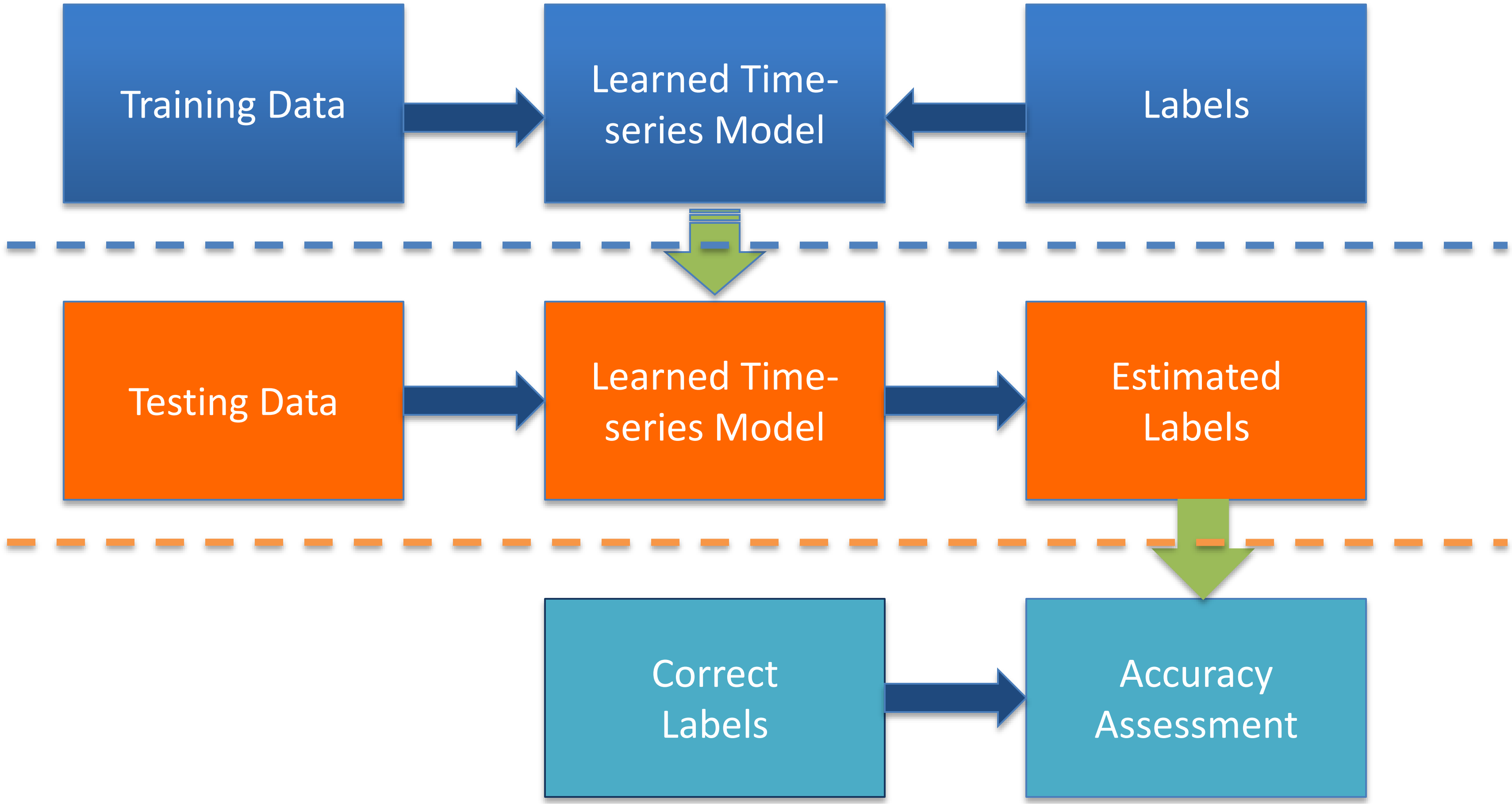


6. Pull suture with left hand

7. Pull suture with right hand

8. Orient needle with both hands

Testing and Training Process



Classifier vs Manual Segmentation

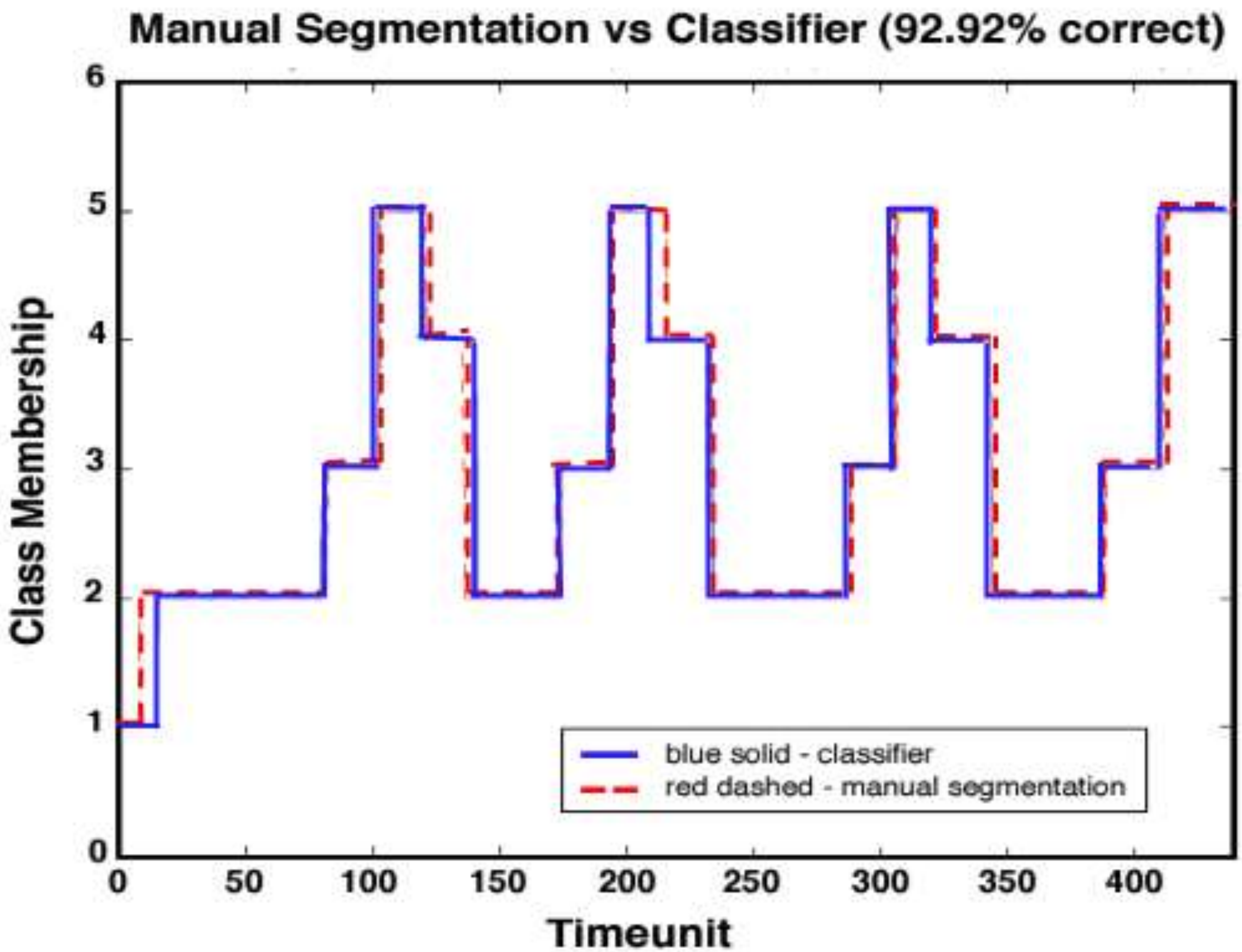
H Lin, I Shafran, D Yuh, G Hager. Towards Automatic Skill Evaluation: Detection and Segmentation of Robot-Assisted Surgical Motions. *Computer Aided Surgery, 11(5):220-230, September 2006.*

n		Number of labeled classes	LDA output dimensions		% correct
1		6	3		91.26
2		6	4		91.46
3		6	5		91.14
4		5	3		91.06
5		5	4		91.34
6		5	3		92.09
7		5	4		91.92
8		4	3		91.88

Expectation Maximization + Bayes Classifier

Henry Lin

Classifier vs Segmentation



Multi-User, Multi-Task Data Collection

Subject ID	Medical Training	Da Vinci Training	Hours?
A	-	-	<1
B	-	-	10-15
C	-	-	100+
D	X	X	100+
E	X	X	100+
F	-	X	100+
G	-	X	<10
H	-	X	<10
I	-	-	<1

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

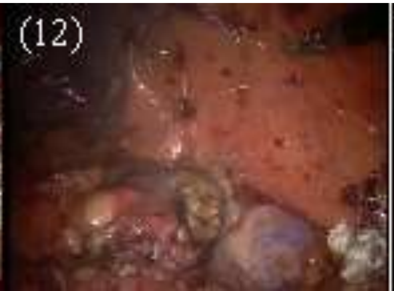
(8) Orienting needle

(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

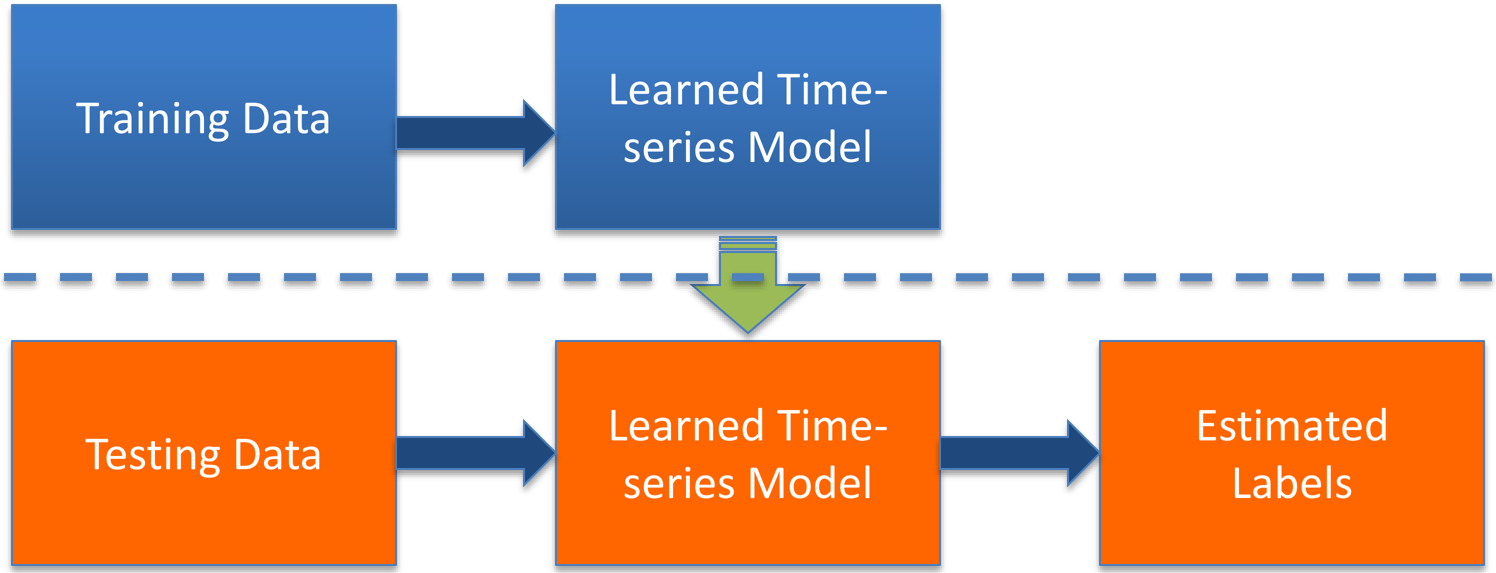
Results of Supervised State Labeling

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

LDA Dimension	Setup I	Setup II	Setup III
10	83%	82%	**73%**
15	86%	82%	71%
20	**87%**	**83%**	70%

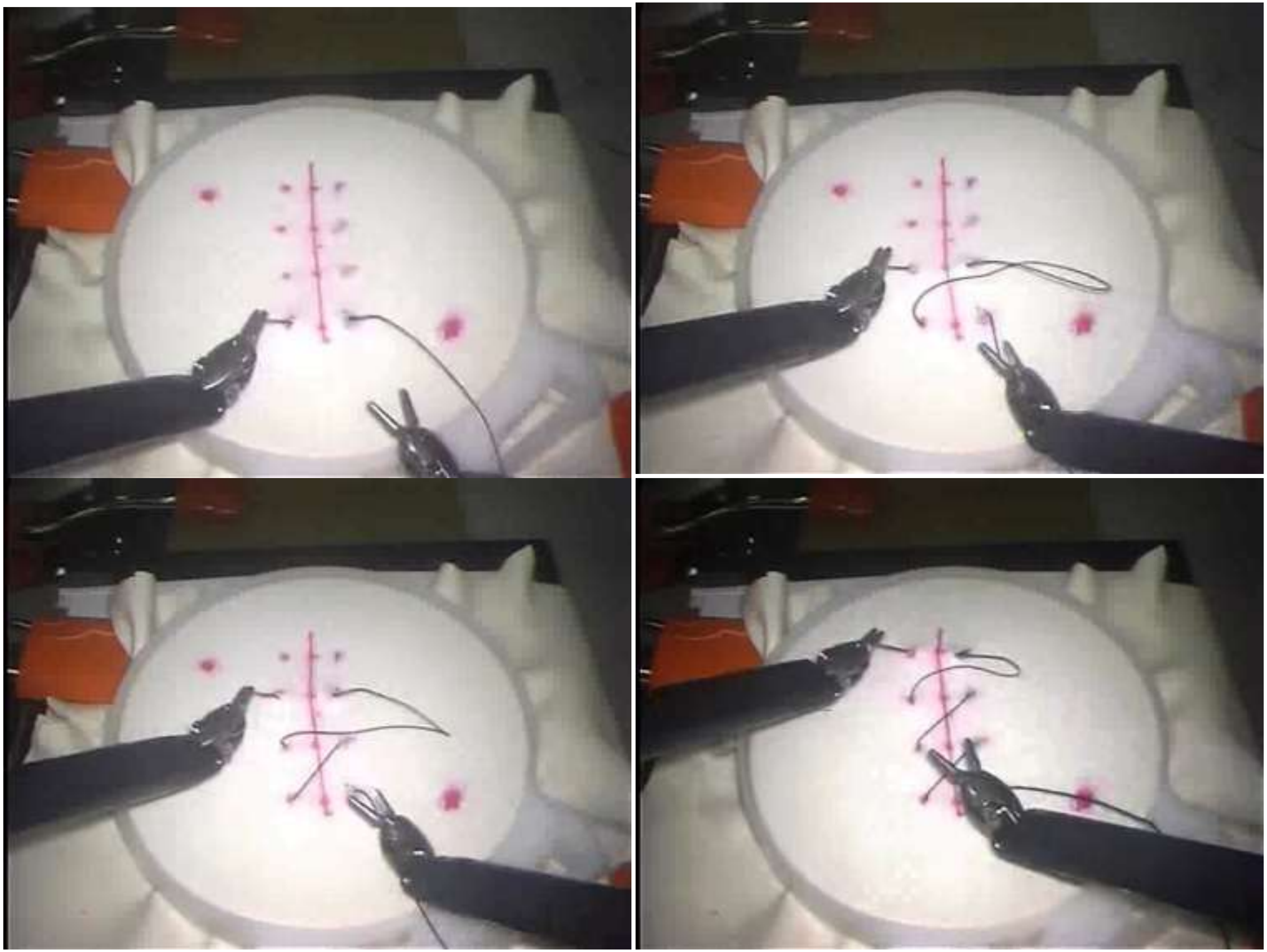
- Multi-State SLR HMM per gesture
  - New notion of “dexeme”
- HLDA
  - A discriminative projection per state in the HMM

Can We Learn a Vocabulary?

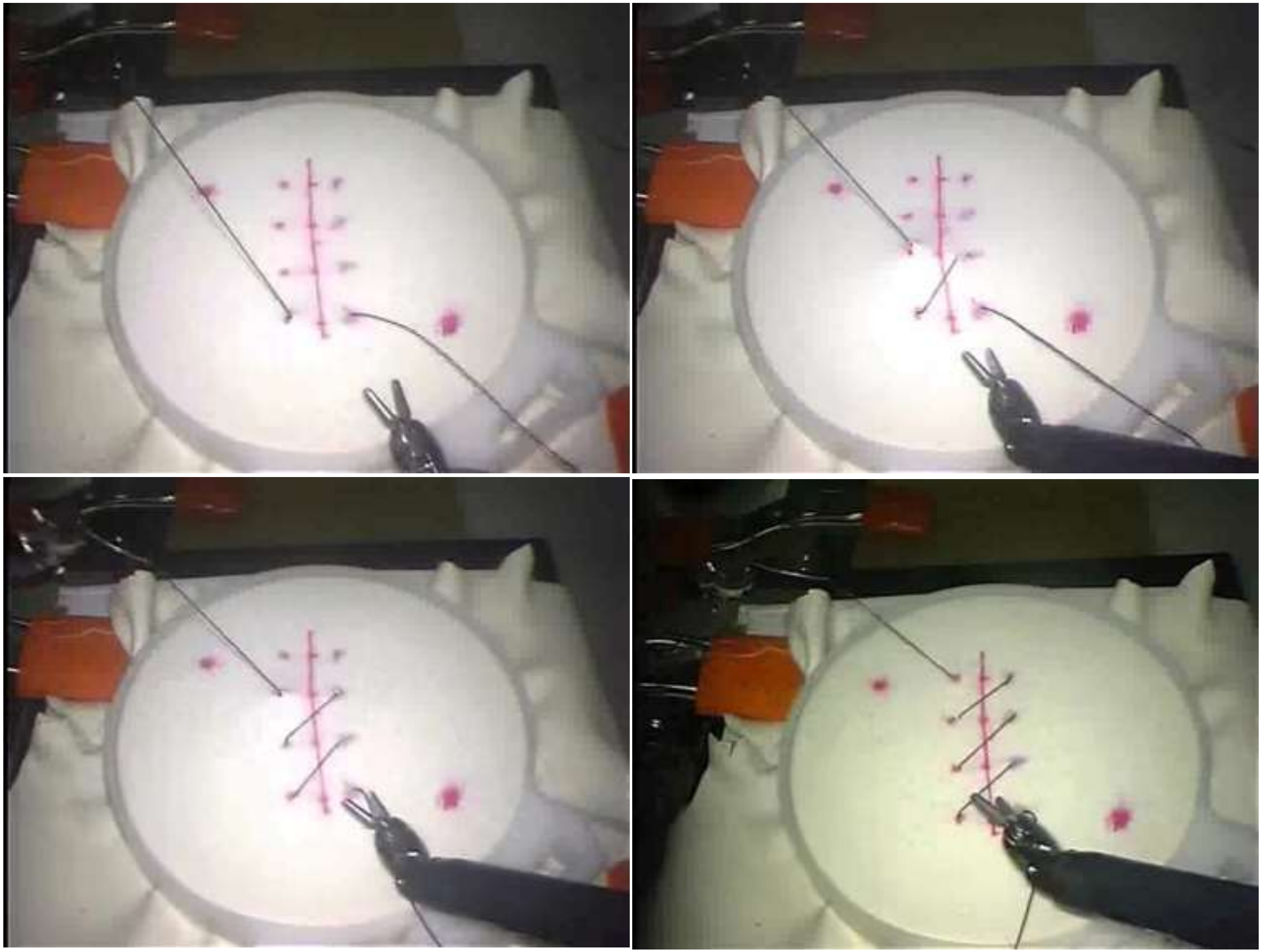


- Start with a one-state HMM (N=1)
- *Concurrently* split each state into four
- *Choose N+Δ* states that maximize likelihood
- Continue until a desired number of states

What Gesture is HMM state #2?

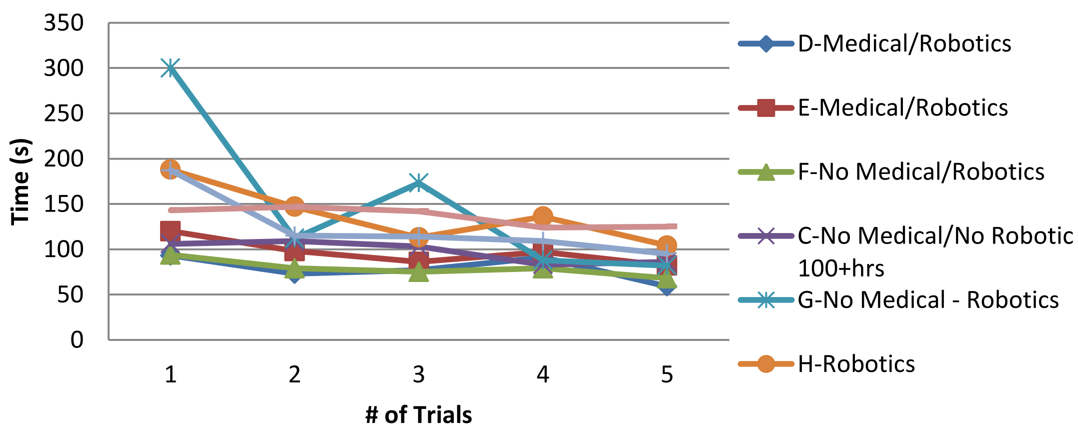


What about HMM state #10?



Skill Assessment: Beyond Time and Motion

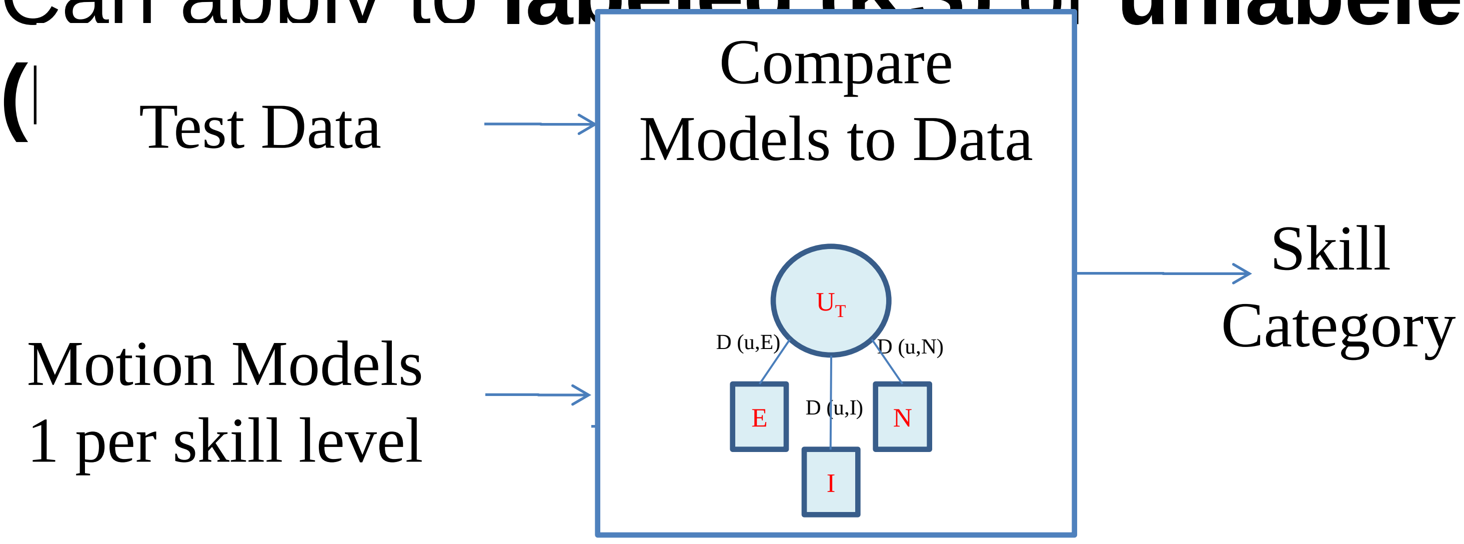
- Can we detect interesting differences in categories of users?
  - Experiment 0: Accuracy of skill classification
- Can we do so at the surgeme level?
  - Experiment 1: Surgeme level vs. Task Level HMM
- Does labeling matter?
  - Experiment 2: Task Level HMM with known states vs. Task Level HMM with unknown states



Task Level Evaluation

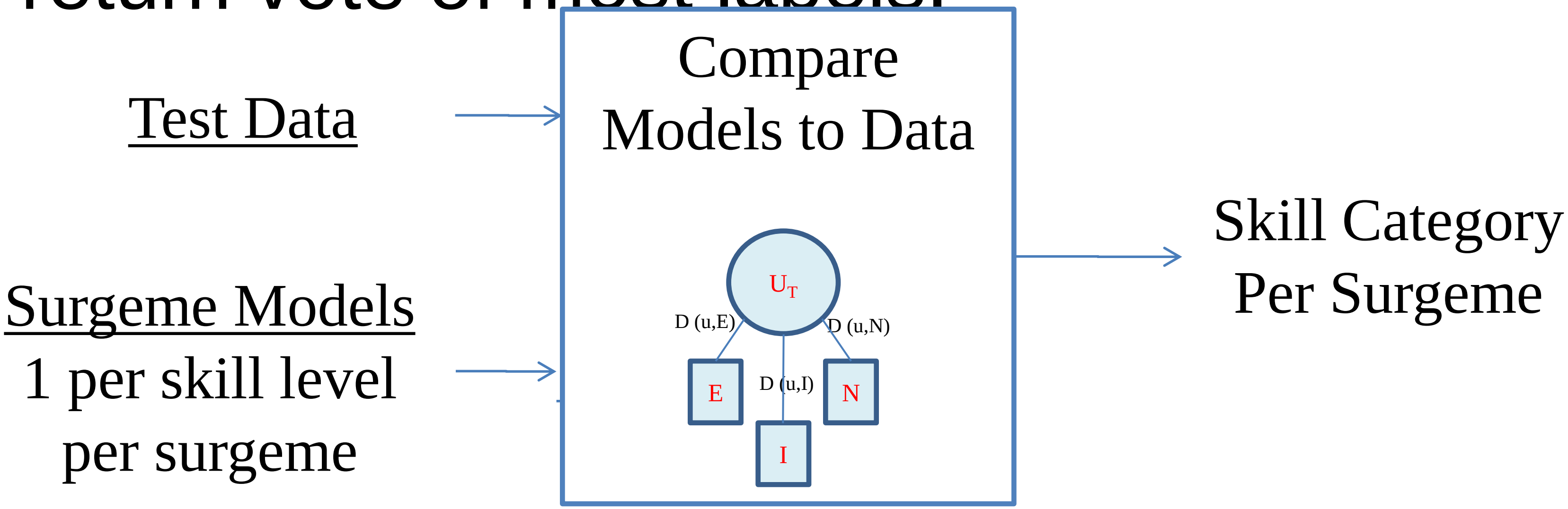
- Build one statistical model of each skill level for each trial using instrument velocities

- Can apply to **labeled (KS)** or **unlabeled** **(**



Task Level Evaluation

- Trained 3 skill level models for **each surgeme**.
- Test each sequence of surgemes and return vote of most labels.



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

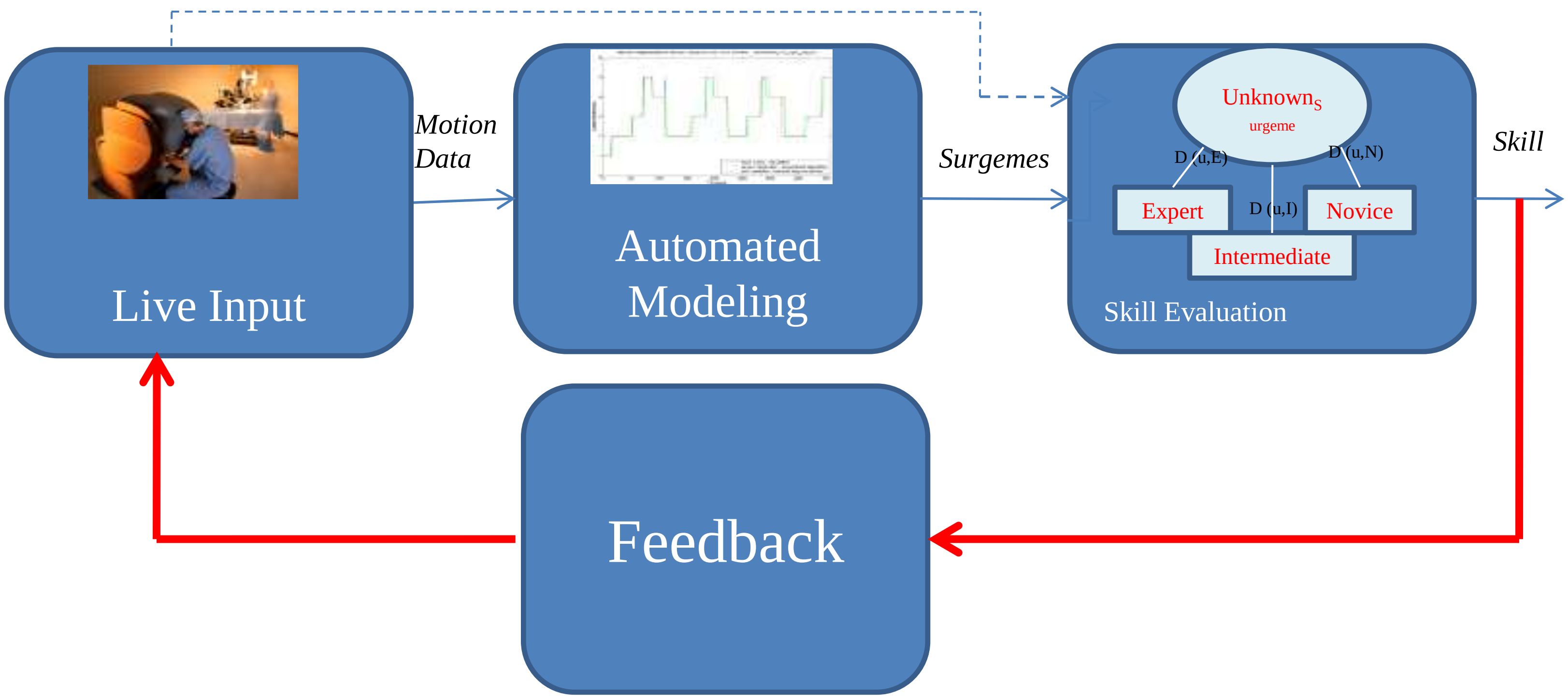
- 100% correct classification on surgeme level
- 100% correct classification on task level, labeled data
- 95% correct classification on task level, unlabeled data
- Certain surgemes more indicative of skill than others

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%

Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery Carol E. Reiley and Gregory D. Hager, Proc. MICCAI 2009

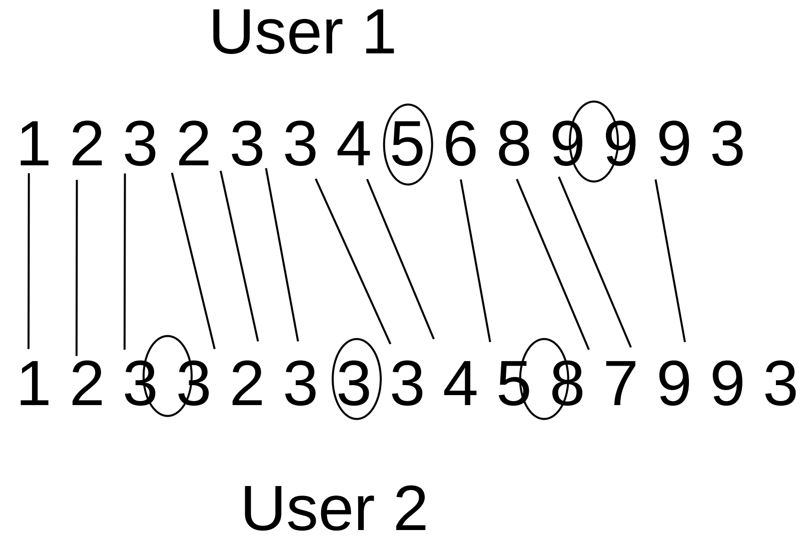
Applications Of Motion Models

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



One Example

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

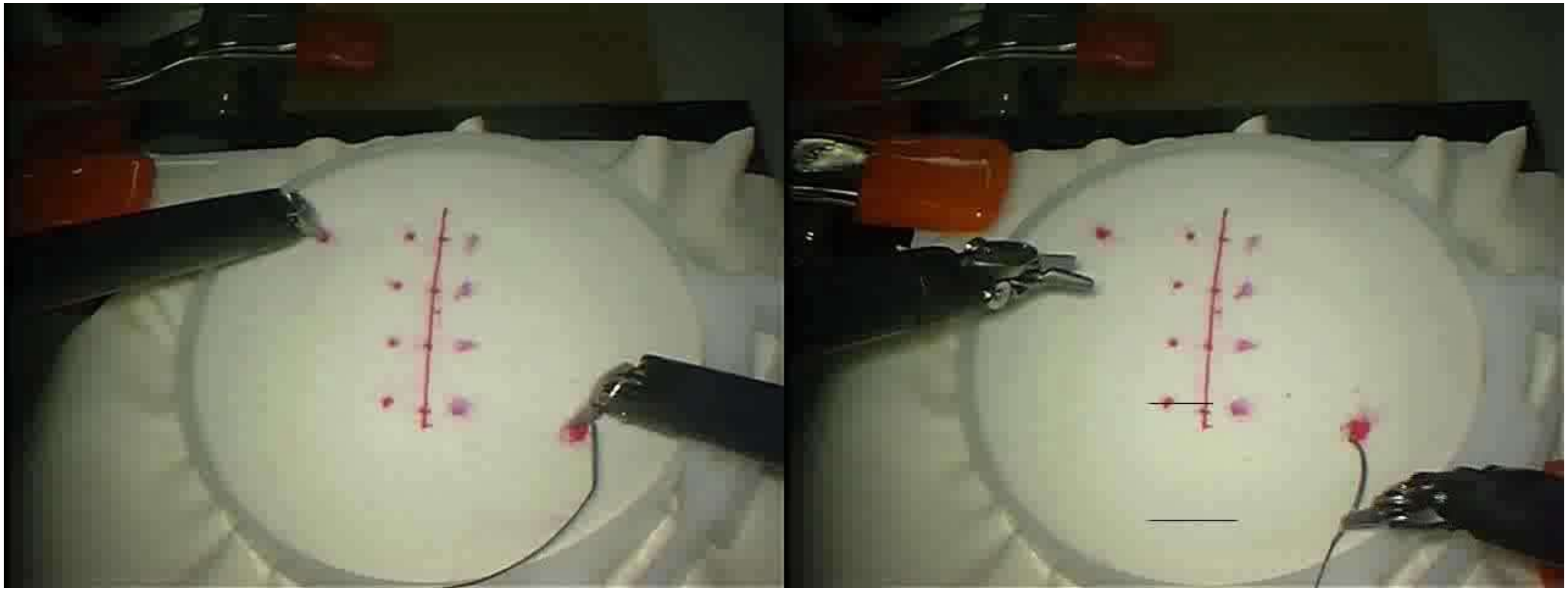


	Expert	Inter.	Novice
Expert	0.38	0.51	0.61
Inter.	0.51	0.42	0.62
Novice	0.61	0.62	0.65

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.

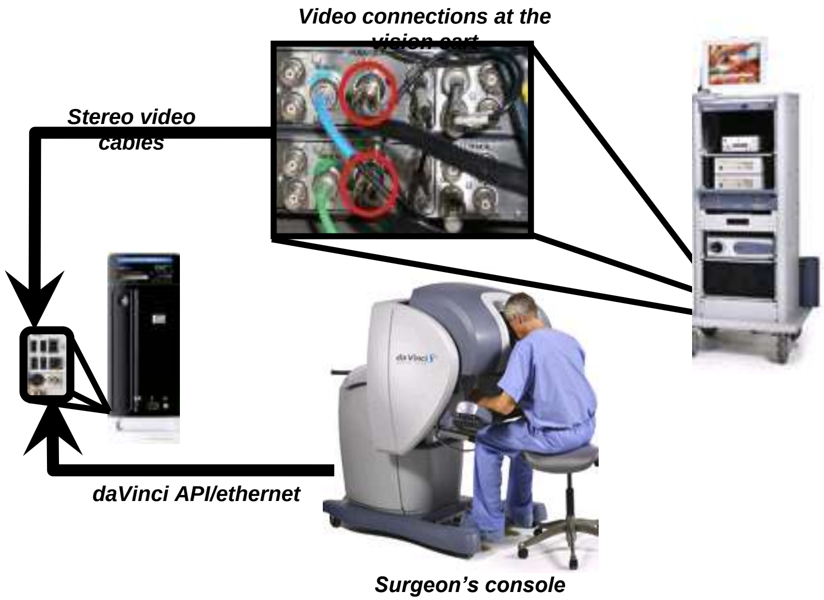
Expert-Expert

Expert-Novice



Validation: Multi-site Data Collection

Rajesh Kumar, JHU



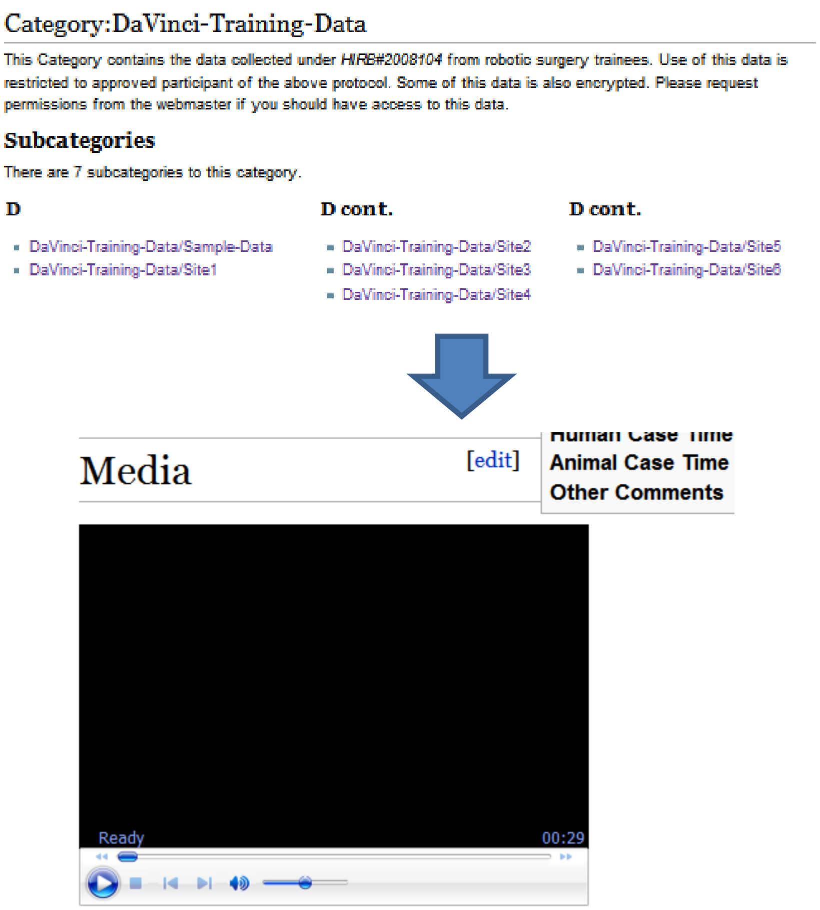
- Secure, anonymized, transparent, and systematic collection of *procedure data* for creation of a longitudinal archive of robotic surgery training with trainees of known surgical and robotic proficiency
- Analysis of system and surgical skill acquisition and identification of key robotic surgery skills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

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

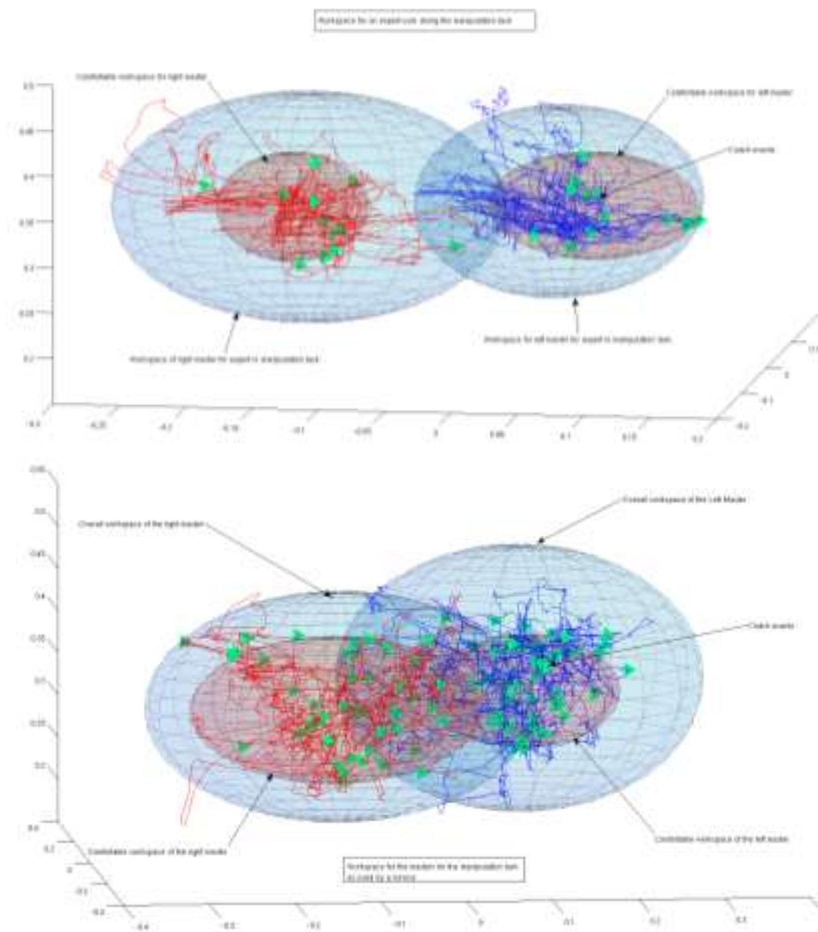
Data Access/Archive

- 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**

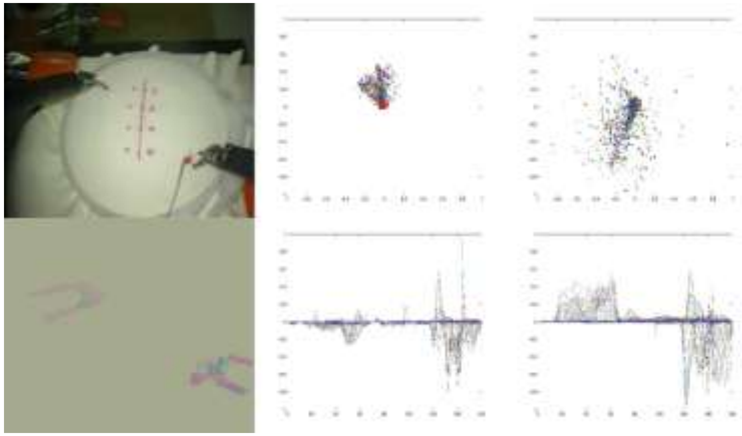


Current Status

- 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



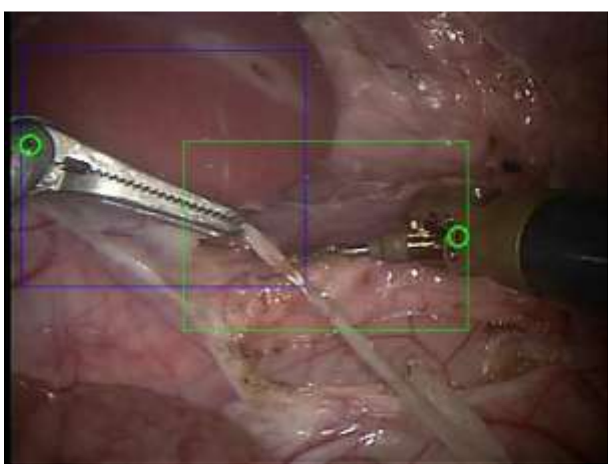
Skill Level	Throws	%
Expert	72/80	90.0
Inter/Novice	68/76	89.5
All	141/156	90.4

Video Analysis

Eye Tracking in Da Vinci Robot

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
Voros, Hager, IEEE BioRob, 2008

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

- C.E. Reiley, H.C. Lin, B. Varadarajan, B. Vagolgyi, S. Khudanpur, D. D. Yuh, and G. D. Hager, “Automatic Recognition of Surgical Motions Using Statistical Modeling for Capturing Variability”, Medicine Meets Virtual Reality, 132:396-401, 2008.
- 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

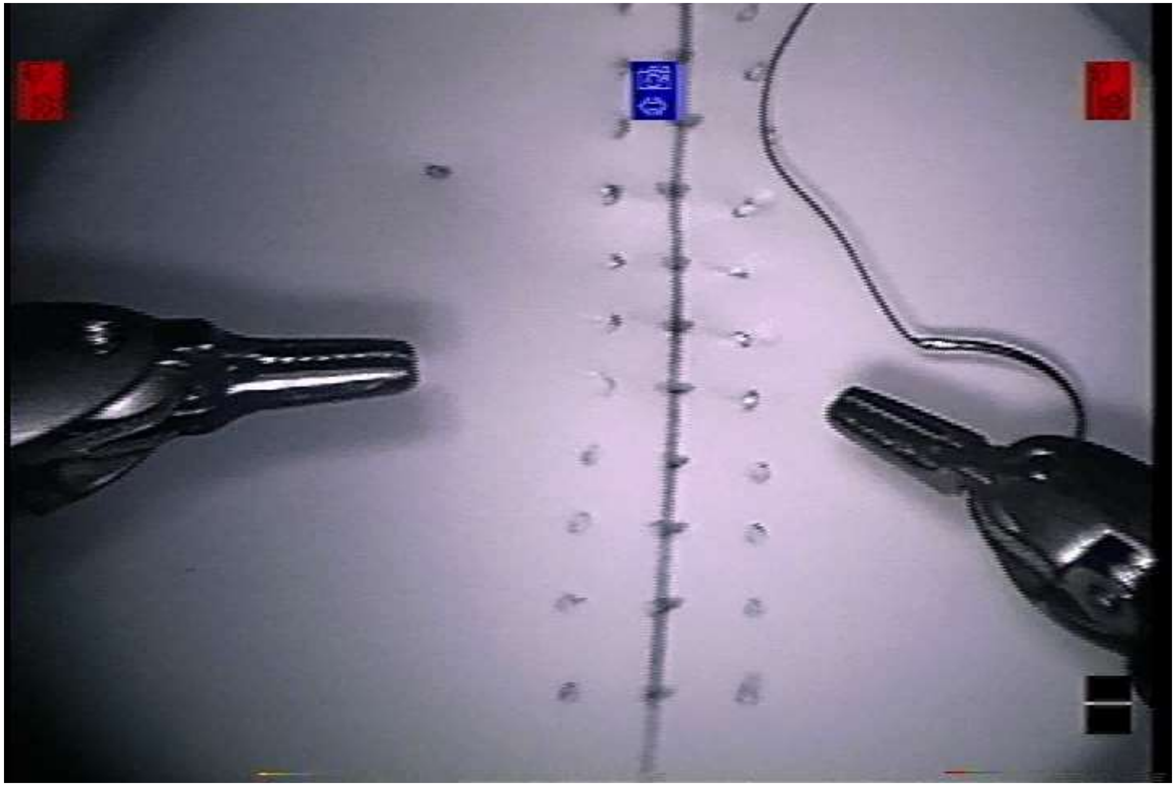
- Intelligent Assistance
  - "I know what you're trying to do"
  - Supervisory interaction rather than "hands on"
- **Fundamental** understanding of human manipulative activity

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

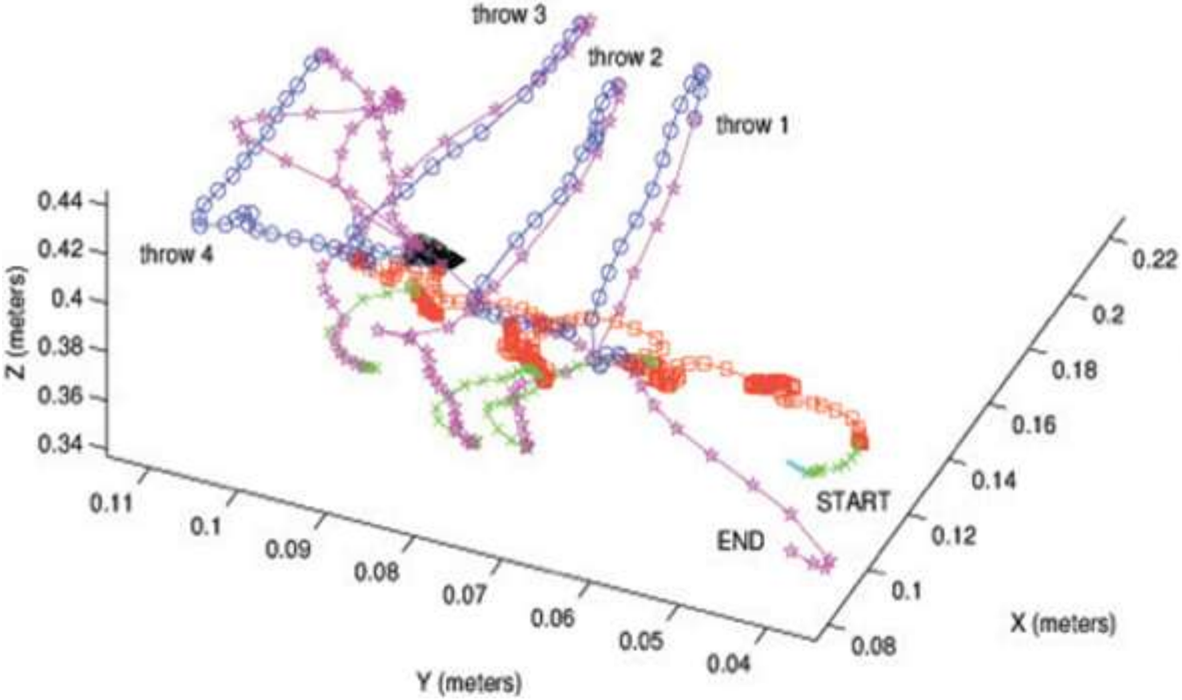
- Funding
  - **NSF IIS 0534359**
  - **NSF CDI 0941362**
  - **NSF CPS 0931805**
  - **NIH R21 EB009143**
  - NSF EEC 9731748
  - NSF EEC 0646678
  - NSF MRI 0722943
  - NIH R42 RR019159
- Intuitive Surgical

Suturing Trial



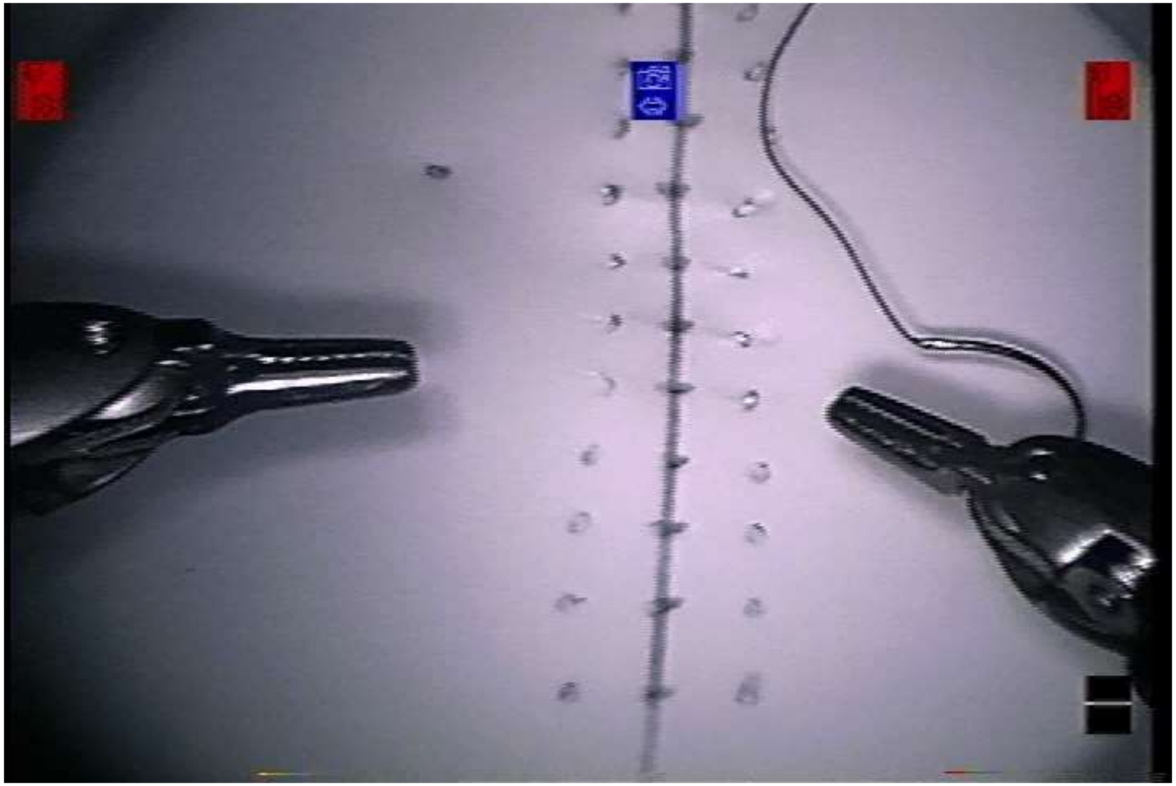
Expert

Suturing Trial



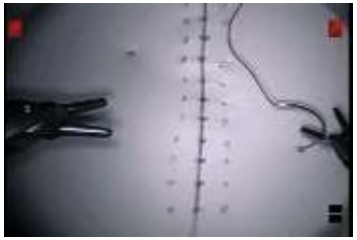
Expert

Suturing Trial



Expert

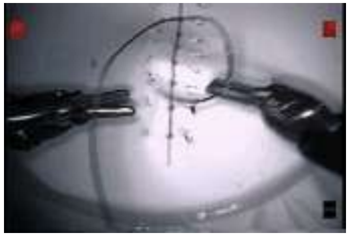
Surgeme Vocabulary


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)

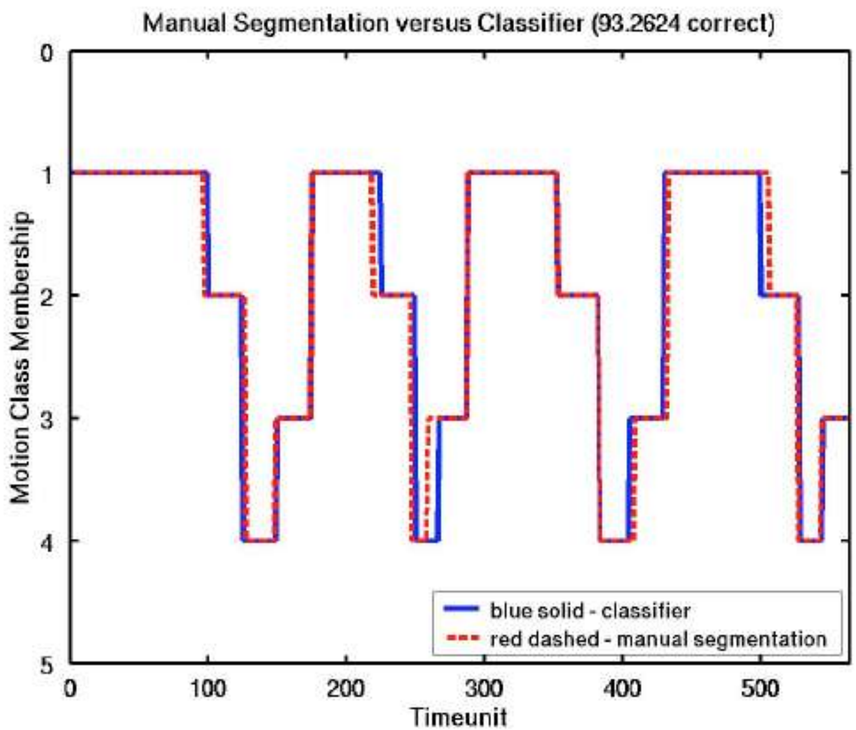

6. Pull suture with left hand

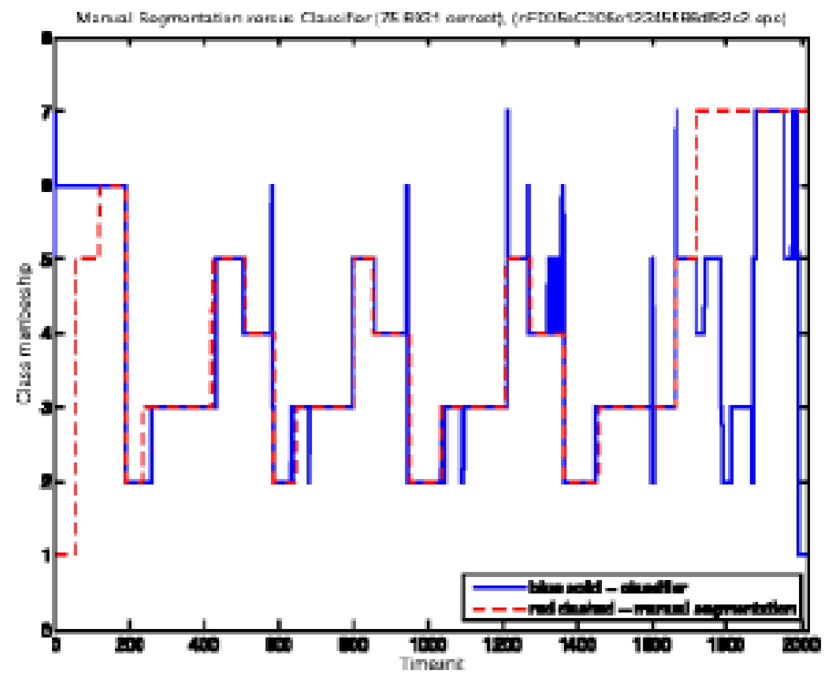

7. Pull suture with right hand


8. Orient needle with both hands

Prior Work on Motion Segmentation

- Example classifier to manual segmentation result



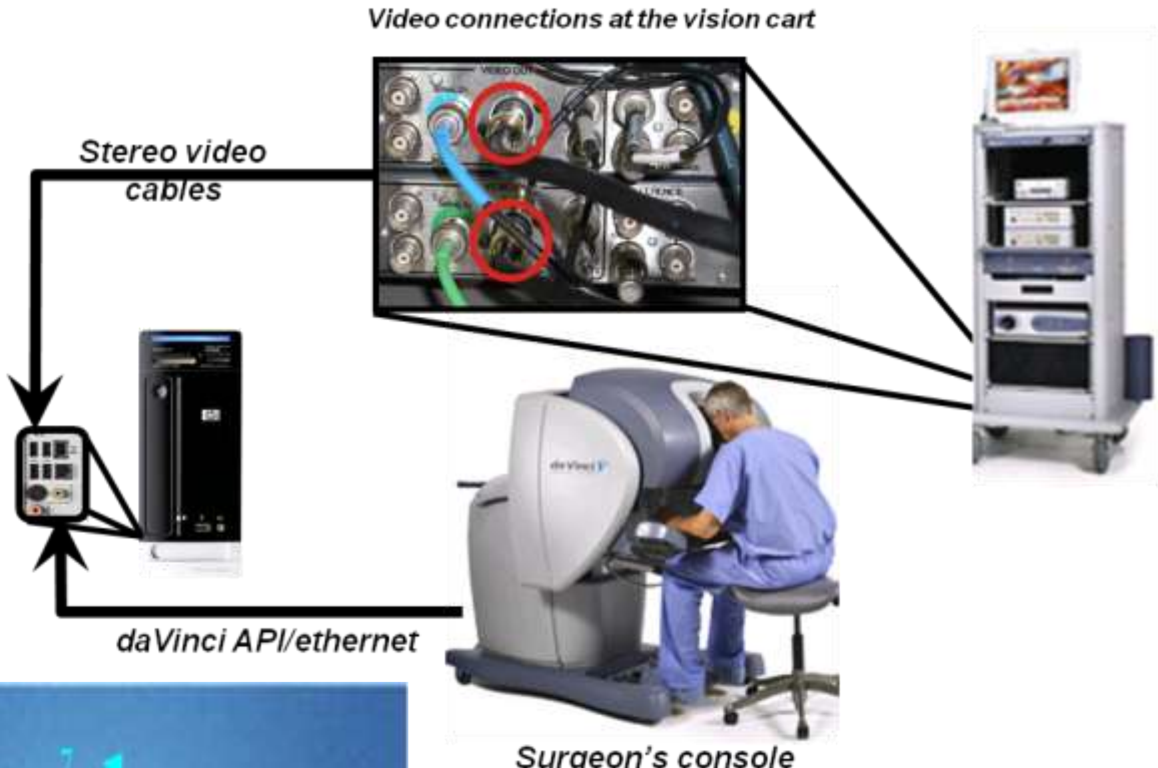


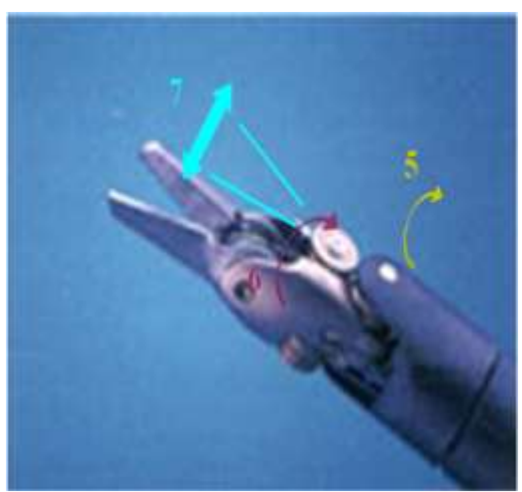
Single User

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



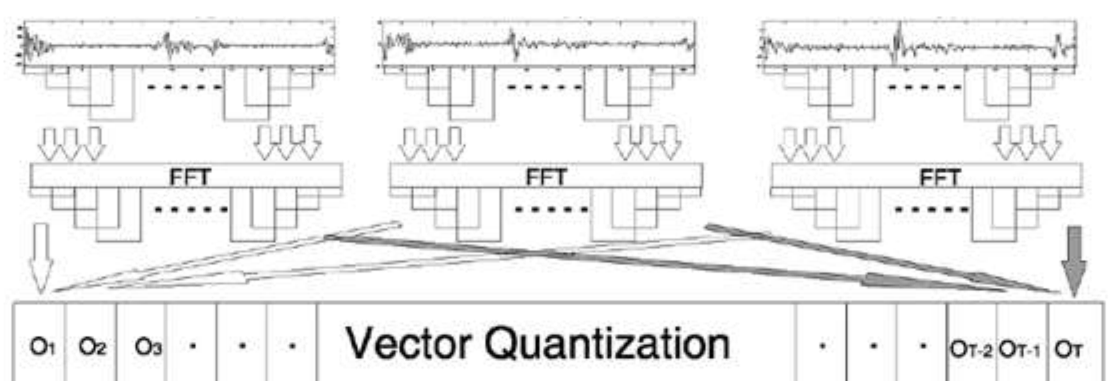


Methods to Assess Skill

- Descriptive Statistics
- Skill Modeling through Language Modeling
  - Task Based
  - Surgeme Based

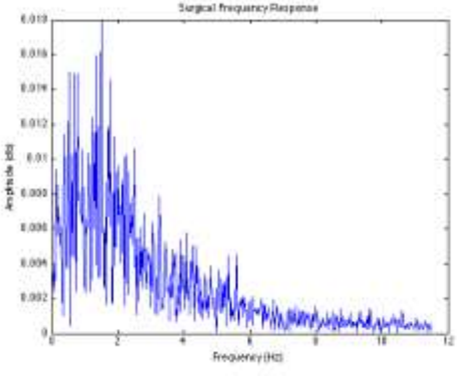
Skill Modeling

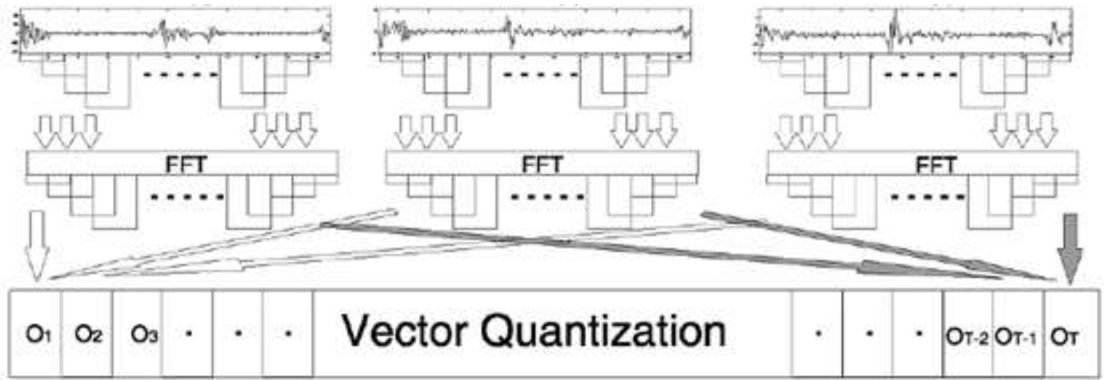
- Data filtering and vector quantization techniques to discretize input data
- Build skill models



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?

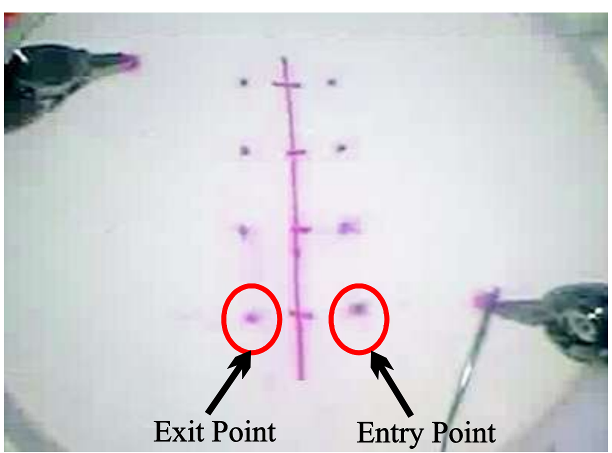
$$\lambda^* = \arg\max[\log P(O_{test}|\lambda_{se}), (\log P(O_{test}|\lambda_{si}), (\log P(O_{test}|\lambda_{sn})]$$

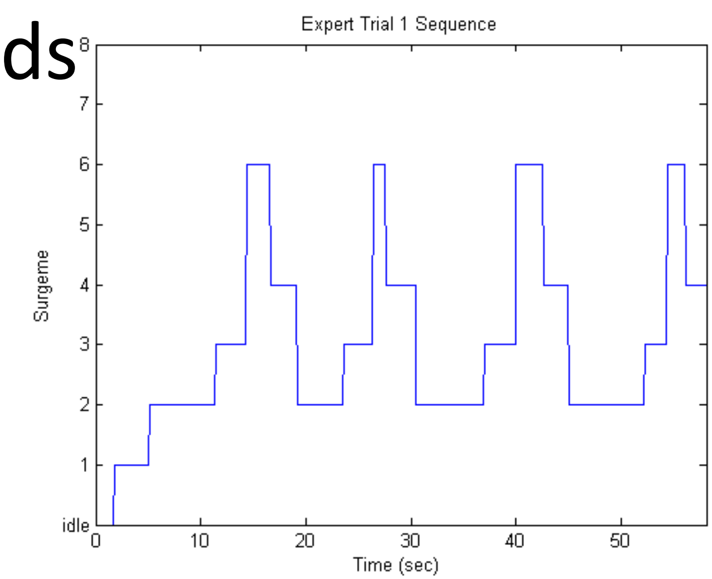
Or similarly…

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

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





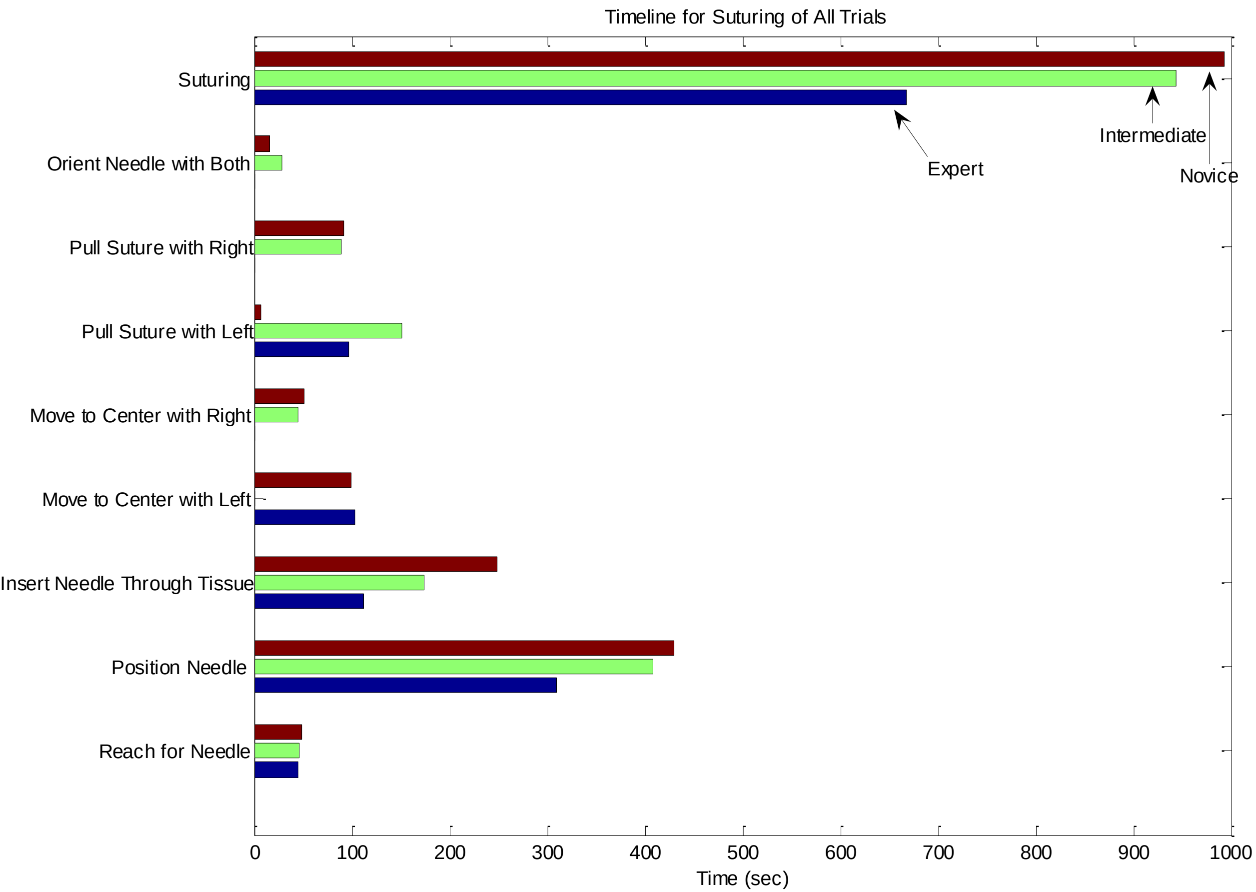
Experiment 1 Results

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

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%

Task decomposition of surgemes according to time



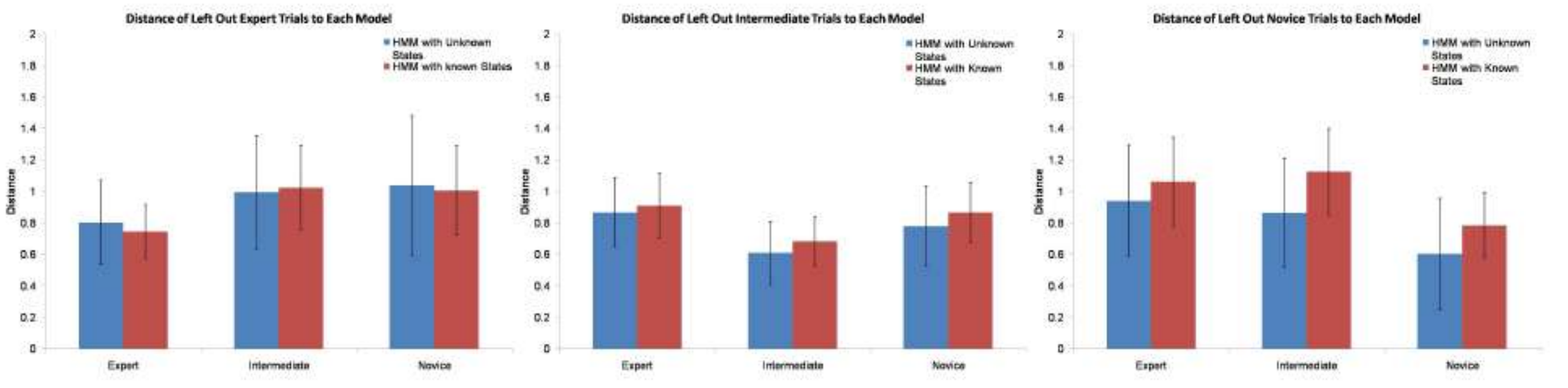
"Confusion Matrix"

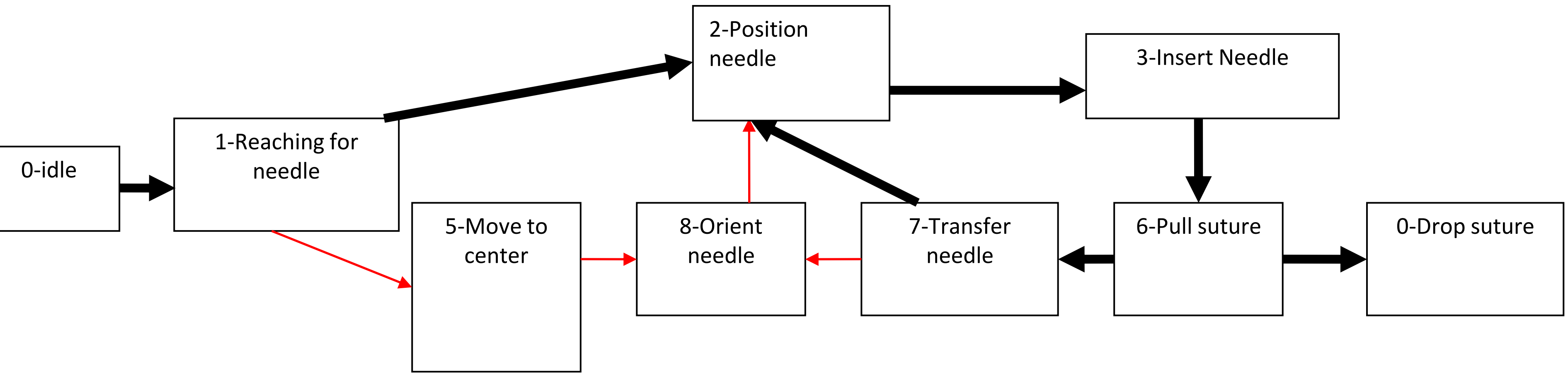
	Exp.	Int.	Nov.	count
Exp. S1	0.50	0.28	0.22	18
Int. S1	0.33	0.67	0	18
Nov. S1	0.31	0	0.69	16
Exp. S2	0.76	0.12	0.12	76
Int. S2	0.78	0.16	0.06	77
Nov. S2	0.16	0.07	0.78	76
Exp. S3	0.79	0.17	0.04	76
Int. S3	0.35	0.53	0.12	75
Nov. S3	0.34	0.12	0.54	74
Exp. S4	0.89	0.02	0.09	57
Int. S4	0.00	0.78	0.22	27
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

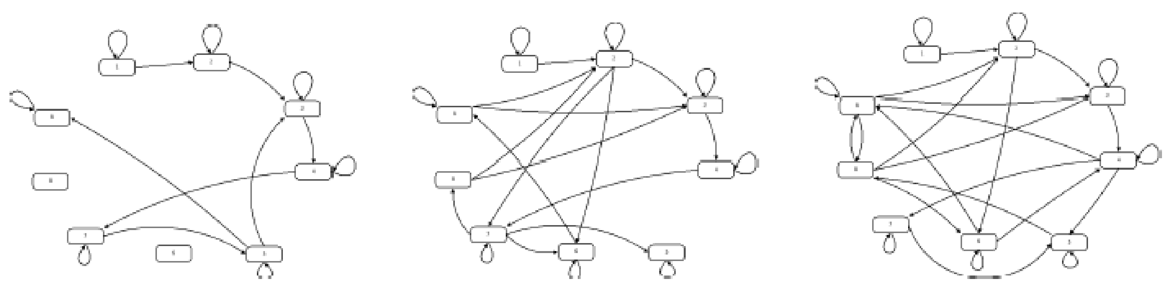
Experiment 2 Results

Task Level HMM with known states (2a) vs. Task Level HMM with unknown states (2c)

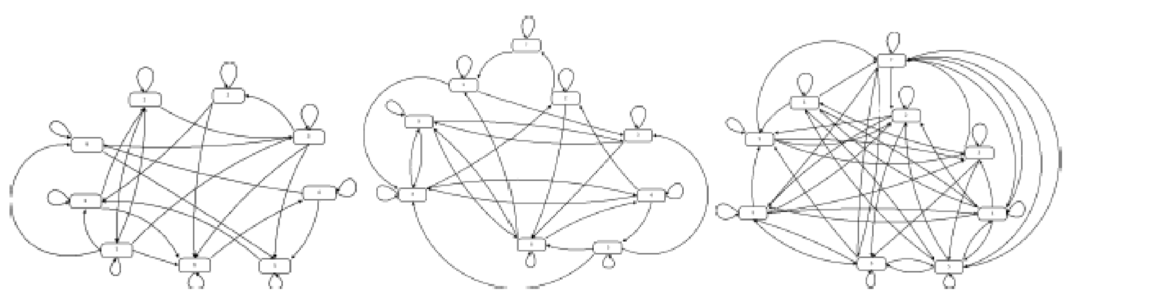
- 100% with known states; 94% correct classification with unknown states
- Unlabeled data does almost as well as labeled



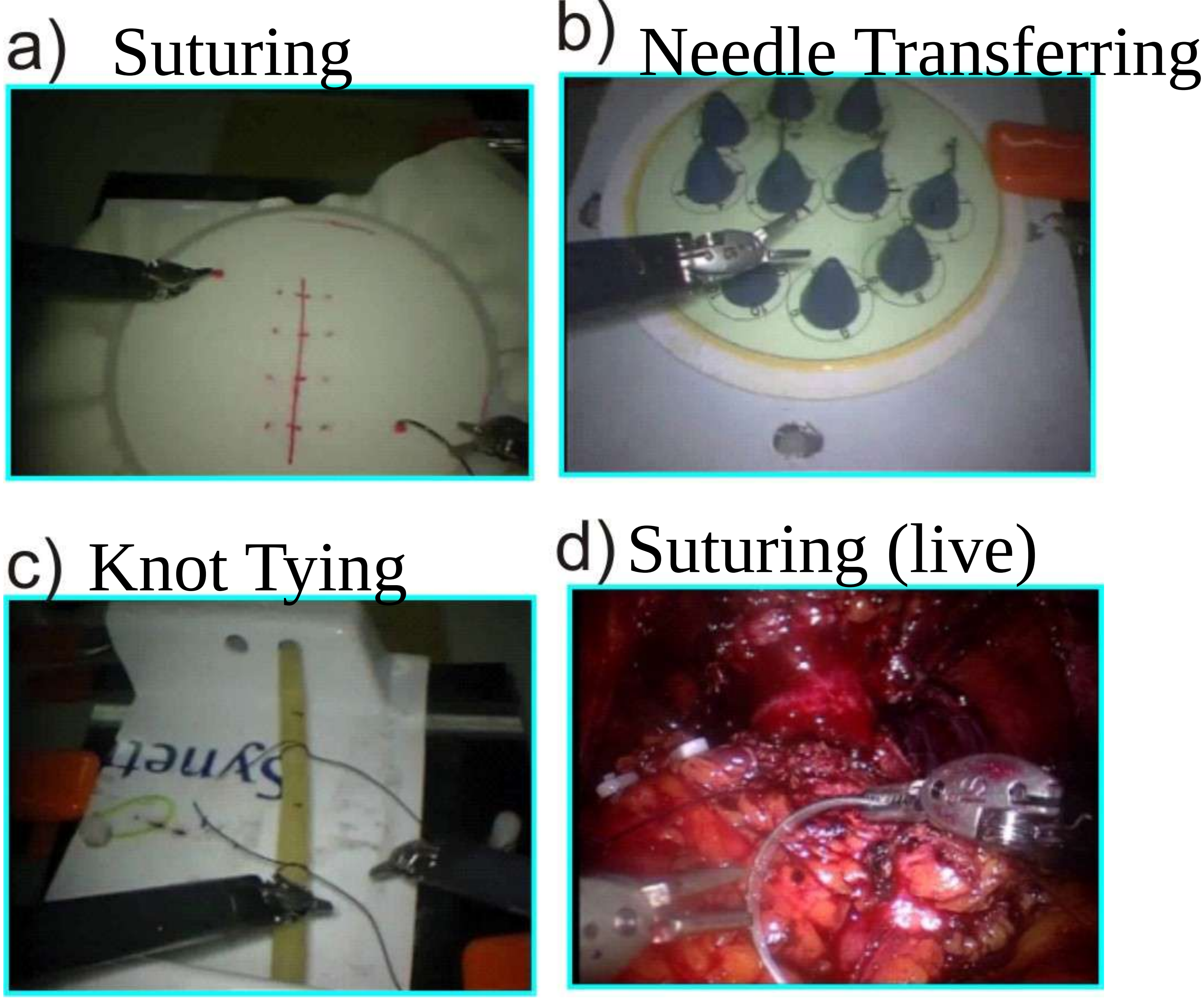




(a) Exp.KS (b) Int.KS (c) Nov.KS



(d) Exp.BW (e) Int.BW (f) Nov.BW

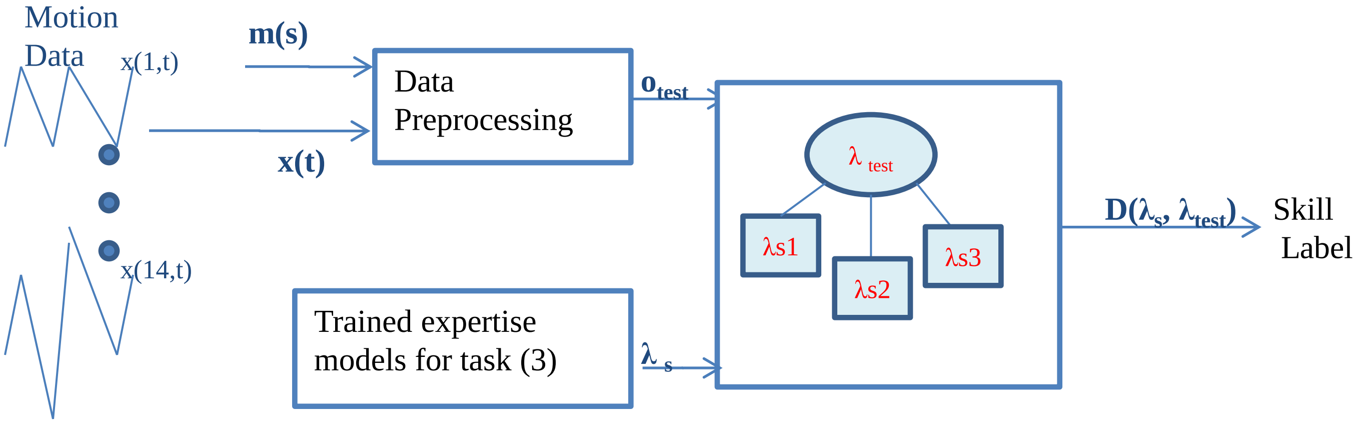

a) Suturing
b) Needle Transferring
c) Knot Tying
d) Suturing (live)

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

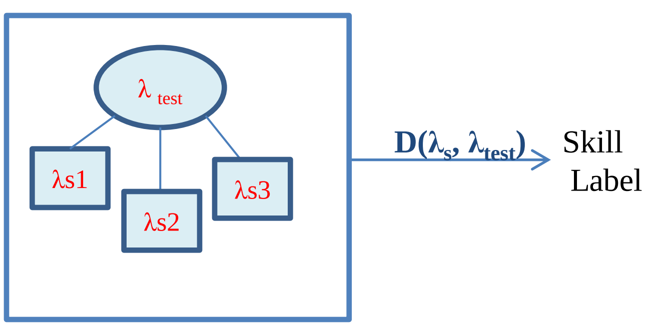


Surgeme Level Methods

- 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

Time/surgeme	
0	1
22	5
50	2
100	1

Transcription file



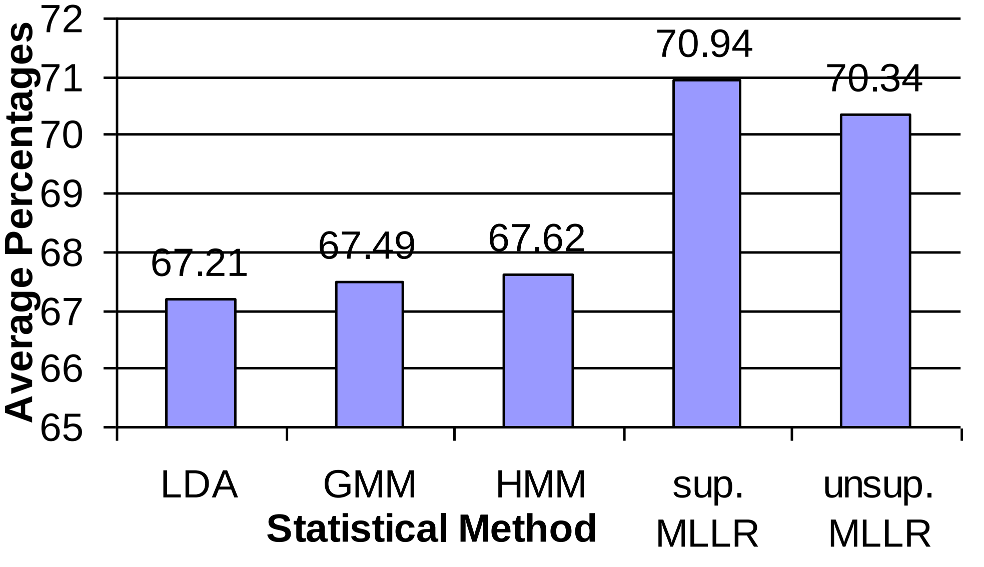
Time/surgeme/skill		
0	1	exp
22	5	int
50	2	exp
100	1	novice

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