Automatic Detection and Segmentation of Robot-Assisted Surgical Motions

Henry C. Lin[1], Izhak Shafran[2], Todd E. Murphy[3], Allison M. Okamura[3], David D. Yuh[4], and Gregory D. Hager[1]

[1] Department of Computer Science
[2] Department of Electrical and Computer Engineering
[3] Department of Mechanical Engineering
The Johns Hopkins University, Baltimore, MD, USA
[4] Division of Cardiac Surgery
Johns Hopkins Medical Institutions, Baltimore, MD, USA
hcl@cs.jhu.edu

Abstract. Robotic surgical systems such as Intuitive Surgical's da Vinci system provide a rich source of motion and video data from surgical procedures. In principle, this data can be used to evaluate surgical skill, provide surgical training feedback, or document essential aspects of a procedure. If processed online, the data can be used to provide context-specific information or motion enhancements to the surgeon. However, in every case, the key step is to relate recorded motion data to a model of the procedure being performed. This paper examines our progress at developing techniques for "parsing" raw motion data from a surgical task into a labelled sequence of surgical gestures. Our current techniques have achieved >90% fully automated recognition rates on 15 datasets.

1 Introduction

Surgical training and evaluation has traditionally been an interactive and slow process in which interns and junior residents perform operations under the supervision of a faculty surgeon. This method of training lacks any objective means of quantifying and assessing surgical skills [1–4]. Economic pressures to reduce the cost of training surgeons and national limitations on resident work hours have created a need for efficient methods to supplement traditional training paradigms. While surgical simulators aim to provide such training, they have limited impact as a training tool since they are generally operation specific and cannot be broadly applied [5–8].

Robot-assisted minimally invasive surgical systems, such as Intuitive Surgical's da Vinci, introduce new challenges to this paradigm due to its steep learning curve. However, their ability to record quantitative motion and video data opens up the possibility of creating descriptive, mathematical models to recognize and analyze surgical training and performance. These models can then be used to help evaluate and train surgeons, produce quantitative measures of surgical proficiency, automatically annotate surgical recordings, and provide data for a variety of other applications in medical informatics.

Recently, several approaches to surgical skill evaluation have had success. In the area of high-level surgical modeling, Rosen et al. [9–11] have shown that statistical models derived from recorded force and motion data can be used to classify surgical skill level (novice or expert) with classification accuracy approaching 90%. However, these results rely on a manual interpretation of recorded video data by an expert physician. In the area of low-level surgical data analysis, the MIST-VR laparoscopic trainer has become a widely used system [12]. These systems perform low-level analysis of the positions, forces, and times recorded during training on simulator systems to assess surgical skill [13–15]. Similar techniques are in a system developed by Darzi et al., the Imperial College Surgical Assessment Device (ICSAD) [16]. ICSAD tracks electromagnetic markers on a trainee's hands and uses the motion data to provide information about the number and speed of hand movements, the distance traveled by the hands, and the overall task time. ICSAD has been validated and used extensively in numerous studies, e.g. [17, 18]. Verner et al. [19] collected da Vinci motion data during performance of a training task by several surgeons. Their analysis also examined tool tip path length, velocities, and time required to complete the task.

It is important to note that ICSAD, MIST-VR, and most other systems mentioned above simply count the number of hand motions, using hand velocity as the segmentation criteria, and do not attempt to identify surgical gestures. In this paper we have developed automatic techniques for not only detecting surgical gestures but also segmenting them. This would allow for the development of automatic methods to evaluate overall proficiency and specific skills.

2 Modeling Robot-Assisted Surgical Motion

Evaluating surgical skill is a complex task, even for a trained faculty surgeon. As a first step, we investigate the problem of recognizing simple elementary motions that occur in a simplified task. Robot motion analysis of users with varying da Vinci experience were studied. Automatic recognition of elementary motion requires complex machine learning algorithms, and, potentially, a large number of parameters. To guide the choice of techniques and to gain useful insight into the problem, we divided the task into functional modules, illustrated in Fig. 2, and akin to other pattern recognition tasks such as automatic speech recognition. In this section, we will describe the data used for this study, the paradigm for training and testing, and a solution for the motion recognition problem.

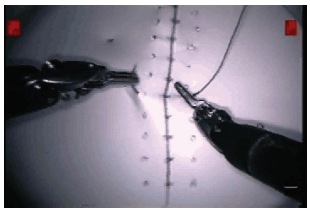

Fig. 1. A video frame of the suture task used for this study.

2.1 Corpus for the Experiments

The da Vinci API data consists of 78 motion variables acquired at 10 Hz during operation. Of these, 25 track each of the master console manipulators, and 14 track each of the patient-side manipulators. We selected the suturing task (Fig. 1) as the model in which our motion vocabulary, $\mathrm{m}(s)$, would be defined.

The eight elementary suturing gestures are:

1. reach for needle (gripper open)
2. position needle (holding needle)
3. insert needle/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 two hands

2.2 Recognizing Surgical Motion

The task of recognizing elementary surgical motions can be viewed as a mapping of temporal signals to a sequence of category labels. The category labels belong to a finite set C, while the temporal signals are real valued stochastic variables, $\mathbf{X}(k)$, tapped from the master and patient-side units. Thus, the task is to map:

$$\mathcal{F} : \mathbf{X}(1:k) \mapsto C(1:n)$$

Our work adopts a statistical framework, where the function $\mathcal{F}$ is learned from the data. The task of learning $\mathcal{F}$ can be simplified by projecting $\mathbf{X}(k)$ into a feature space where the categories are well separated. The sequence of operations is illustrated by the functional block diagram in Fig. 2.

2.3 Feature Processing

The goal of feature processing is to remove redundancy in the input features while retaining the information essential for recognizing the motions with high accuracy. As noted earlier, the input feature vectors consist of 78 position and velocity measurements from the da Vinci manipulators. Feature processing reduces the dimension from 78 to less than 6 features without any loss of performance. In this work, we have found the following feature processing steps to be effective.

1. **Local Temporal Features**: Surgical motion seldom changes from one gesture to another abruptly. Thus information from adjacent input samples can

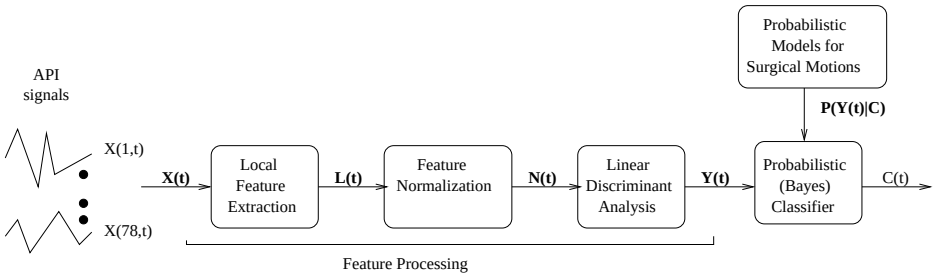

Fig. 2. Functional block diagram of the system used to recognize elementary surgical motions in this study.

be useful in improving the accuracy and robustness of recognizing a surgical motion. As in automatic speech recognition, this information can be incorporated directly by concatenating the feature vector $\mathbf{X}(k_t)$ at time t with those from neighboring samples, $t-m$ to $t+m$, to make it vector of dimension $(2m+1)|\mathbf{X}(k_t)|$.

$$\mathbf{L}(k_t) = [\mathbf{X}(k_{t-m})|\mathbf{X}(k_{t-m+1})|\dots|\mathbf{X}(k_t)|\dots|\mathbf{X}(k_{t+m-1})|\mathbf{X}(k_{t+m})]$$

In addition, derived features such as speed and acceleration were included as a part of each local temporal feature.

2. **Feature Normalization**: Since the units of measurements for position and velocity are different, the range of values that they take are significantly different. This difference in dynamic range often hurts the performance of a classifier or a recognition system. So, the mean and variance of each dimension is normalized by applying a simple transformation,

$$\mathbf{N}_i(k) = \frac{1}{\sigma_i^2}(\mathbf{L}_i(k) - \mu_i),$$

where $\mu_i = \frac{1}{N}\mathbf{L}_i(k)$ and $\sigma_i^2 = \frac{1}{N}(\mathbf{L}_i(k) - \mu_i)^2$.

3. **Linear Discriminant Analysis**: When the features corresponding to different surgical motions are well separated, the accuracy of the recognizer can be considerably improved. One such transformation is the linear discriminant analysis [20].

$$\mathbf{Y}(k) = \mathbf{W}N(k)$$

The linear transformation matrix W is estimated by maximizing the Fisher discriminant, which is the ratio of distance between the classes and the average variance of a class. The transformation that maximizes the ratio projects the features into a space where the classes are compact but away from each other.

2.4 Bayes Classifier

The discriminant function, $\mathcal{F}$, could be of several forms. When all errors are given equal weight, it can be shown that the optimal discriminant function is given by Bayes decision rule.

$$\begin{aligned}\hat{C}(1:n) &= \arg\max_{C(1:n)} P(C(1:n)|\mathbf{Y}(1:k)) \\ &= \arg\max_{C(1:n)} P(\mathbf{Y}(1:k)|C(1:n))P(C(1:n))\end{aligned}$$

In other words, the optimal decision is to pick the sequence whose posterior probability, $P(C(1:n)|\mathbf{Y}(1:k))$, is maximum. Using Bayes chain rule, this can be rewritten as the product of prior probability of the class sequence, $P(C(1:n))$, and the generative probability for the class sequence, $P(\mathbf{Y}(1:k)|C(1:n))$.

As a first step, we make the simplifying assumption that each time frame in the input sequence is independently generated. That is, $P(C(1:k)|\mathbf{Y}(1:k)) = \prod_{i=1}^{k} P(C(i)|\mathbf{Y}(i))$. Thus, the decision is made at each frame independent of its context.

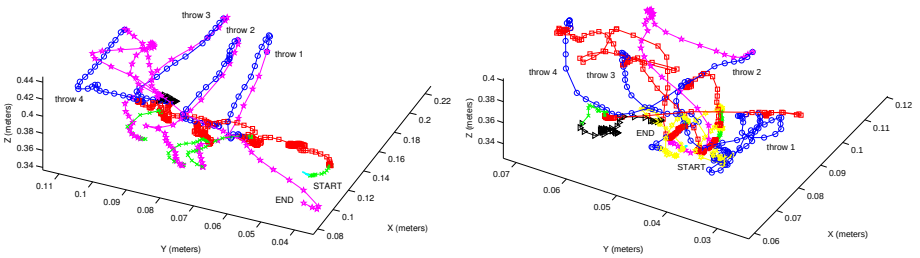

Fig. 3. A plot of the Cartesian positions of the da Vinci left master manipulator, identified by surgical gesture, during performance of a 4-throw suturing task. The left plot is that of an expert surgeon while the right is of a less experienced surgeon.

2.5 Cross-Validation Paradigm

The data used for this study contains 15 expert trials and 12 intermediate trials of performing a suturing task, consisting of 6 to 8 different elementary surgical motions. To improve the statistical significance of the results, we performed a 15-fold cross validation on the expert data. That is, the machine learning algorithm was evaluated by performing 15 different tests. In each test, two trials were held out for testing and the statistical models were trained on the rest of the data. The average across 15 such tests were used to measure the performance of various settings of the parameters.

3 Results

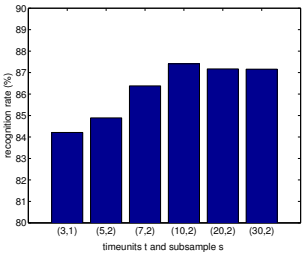

Fig. 4. Results of varying the temporal length t and sampling granularity s

To guide the choice of parameters, our initial experiments were performed on the data collected from 15 trials by an expert da Vinci surgeon, performing a suturing task involving 4 throws (Fig. 1) in each trial. Subsequently, we applied the recognition and segmentation techniques on 12 trials of a surgeon with limited da Vinci experience (intermediate) and compared the results.

After preliminary observation of the data, a few preprocessing steps were carried out before modeling the surgical motions. Of the eight motions defined in Sec. 2.1, the expert surgeon did not utilize motion 5 and 7, so they were not modeled. Each dimension of the feature vector from the expert surgeon contained about 600 samples. For example, Fig. 3 illustrates Cartesian positions of the left master during one of the trials.

3.1 Local Temporal Features

The length and sampling rate of the temporal feature "stacking" was varied to determine the optimal length and granularity of motion to consider. Our results showed, as one would expect, that too little temporal length results in a disappearance of any advantage, whereas too large of a temporal length increased

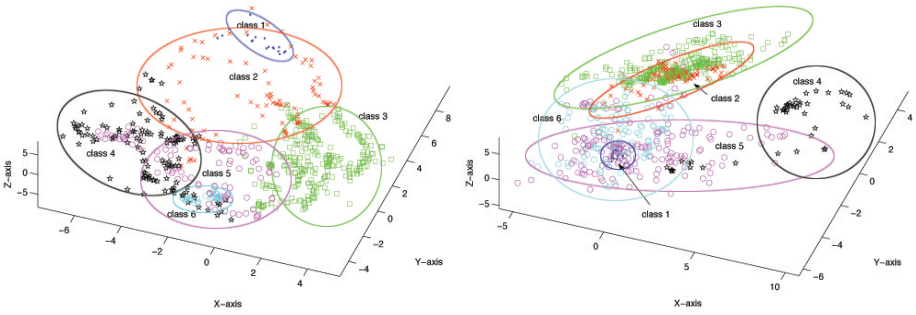

Fig. 5. The result of LDA reduction with m=6 and d=3. The expert surgeon's motions (left) separate more distinctly than the less experienced surgeon's (right).

the chance of blurring the transition between neighboring motions. Fig. 4 shows the results of varying the temporal length (t) and sampling granularity (s). Due to its high recognition rates, we use t=10 and s=2 for the rest of our experiments.

3.2 Linear Discriminant Analysis

Fig. 5 shows the reliability of LDA in separating motion data into 6 distinct regions in a 3-dimensional projection space. An intermediate surgeon's motions tend to not separate as well, indicating less consistent motions.

These initial experiments validated the hypothesis that LDA could be used to simplify the original data into a simpler, low-dimensional data set. A second set of experiments examined the effect of varying the number of motion classes, $C(1{:}\{4,5,6\})$, and the dimensionality of the projection, $d = \{3,4,5\}$. The cross-validation paradigm described in Sec. 2.5 was applied in all experiments to compute a recognition rate. Table. 1 shows the recognition rates of the Bayes classifier after the LDA reduction with varying C and d values.

$\boldsymbol{n}$	$class\_membership$	$LDA\_dimensions$	% correct
1	12345566	3	91.26
2	12345566	4	91.46
3	12345566	5	91.14
4	12345555	3	91.06
5	12345555	4	91.34
6	11234455	3	92.09
7	11234455	4	91.92
8	12234444	3	91.88

Table 1. The results of grouping the motion categories and varying the dimension of the projected space. In the second column, the number of unique integers indicates the number of motion categories, and the position of the integer indicates which motions belong to that category.

Having fine tuned the classifier for surgical motion, we then applied the algorithm to produce segmentations. Fig. 6 shows the comparison of segmentation generated by the algorithm and by a human for a randomly chosen trial of the expert surgeon. In spite of the fact that the model only incorporates weak temporal constraints through the local temporal features described in Sec. 2.3, the segmentation produces surprisingly good results. In most trials, the errors are largely at the transition, as shown in Fig. 6. While using the robotic system, transitions from one motion to the next are often performed without any pause, and as a result it is difficult even for a human to mark a sharp transition boundary. Consequently, we removed a 0.5 second window at each boundary, so as to avoid confidence issues in the manual segmentation. The 0.5 second window is statistically insignificant because an average surgical motion lasts over 3 seconds.

4 Discussion

We have shown that linear discriminant analysis is a robust tool for reducing and separating surgical motions into a space more conducive to gesture recognition. In our highest rated test, we reduced 78 feature vectors into 3 dimensions with 6 classes and still achieved nearly 90% in recognition. With refinement and the combination of other statistical methods, such as Hidden Markov Models (HMMs), we believe mid-90s recognition rates are possible. We have also suggested how this framework can support objective evaluation of surgical skill levels by varying different parameters in our mathematical model. Our experiments have shown that the motions of an expert surgeon are very efficient and thus can be used as a skill evaluator or training model. In ongoing work, we have begun combining the training of expert and intermediate surgeon data to create one model that can distinguish varying skill levels.

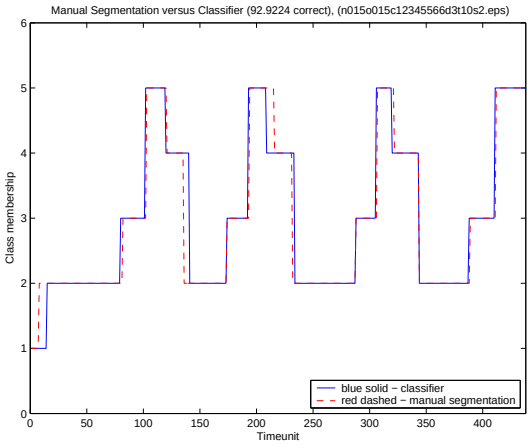

Fig. 6. Comparison of automatic segmentation of robot-assisted surgical motion with manual segmentations. Note, most errors occur at the transitions.

Acknowledgements

This research was supported by NSF and the Division of Cardiac Surgery at the Johns Hopkins Medical Institutions. The authors thank Dr. Randy Brown, Sue Eller, and the staff of Minimally Invasive Surgical Training Center at Johns Hopkins Medical Institutions for access to the da Vinci, and Intuitive Surgical, Inc. for use of the da Vinci API.

References

1. King, R.T.: New keyhole heart surgery arrived with fanfare, but was it premature? Wall Street Journal (1999) 1
2. Haddad, M., et al.: Assessment of surgical residents' competence based on post-operative complications. Int Surg **72** (1987) 230–232
3. Darzi, A., Smith, S., Taffinder, N.: Assessing operative skill needs to become more objective. British Medical Journal **318** (1999) 887–888
4. Barnes, R.W.: But can s/he operate?: Teaching and learning surgical skills. Current Surgery **51(4)** (1994) 256–258
5. Wilhelm, D., et al.: Assessment of basic endoscopic performance using a virtual reality simulator. J Am Coll Surg **195** (2002) 675–681
6. Cowan, C., Lazenby, H., Martin, A., et al.: National health care expenditures, 1998. Health Care Finance Rev **21** (1999) 165–210
7. Acosta, E., Temkin, B.: Dynamic generation of surgery specific simulators - a feasibility study. StudHealth Technology Inform **111** (2005) 1–7
8. R, M., CP, D., SS, M.: The anterior abdominal wall in laparoscopic procedures and limitations of laparoscopic simulators. Surg Endosc **10(4)** (1996) 411–413
9. Rosen, J., et al.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. In: IEEE Trans Biomed Eng. Volume 48(5). (2001) 579–591
10. Rosen, J., et al.: Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. In: Computer Aided Surgery. Volume 7(1). (2002) 49–61
11. Richards, C., Rosen, J., Hannaford, B., Pellegrini, C., Sinanan, M.: Skills evaluation in minimally invasive surgery using force/torque signatures. In: Surgical Endoscopy. Volume 14. (2000) 791–798
12. Gallagher, A.G., Satava, R.M.: Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. In: Surg. Endoscopy. Volume 16(2). (2002) 1746–1752
13. Cotin, S., et al.: Metrics for laparoscopic skills trainers: the weakest link. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Volume 2488. (2002) 35–43
14. O'Toole, R.V., Playter, R.R., Krummel, T.M., Blank, W.C., Cornelius, N.H., Roberts, W.R., Bell, W.J., Raibert, M.R.: Measuring and developing suturing technique with a virtual reality surgical simulator. Journal of the American College of Surgeons **189(1)** (1999) 114–127
15. Yamauchi, Y., et al.: Surgical skill evaluation by force data for endoscopic sinus surgery training system. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Volume 2488. (2002)
16. Darzi, A., Mackay, S.: Skills assessment of surgeons. Surg. **131(2)** (2002) 121–124
17. Datta, V., et al.: The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in laboratory-based model. Journal of the American College of Surgery **193** (2001) 479–485
18. Datta, V., et al.: Relationshop between skill and outcome in the laboratory-based model. Surgery **131(3)** (2001) 318–323
19. Verner, L., Oleynikov, D., Holtman, S., Haider, H., Zhukov, L.: Measurements of the level of expertise using flight path analysis from da vinci robotic surgical system. In: Medicine Meets Virtual Reality II. Volume 94. (2003)
20. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7** (1936) 179–188