Optimization of a Vector Quantization Codebook for Objective Evaluation of Surgical Skill

Timothy M. KOWALEWSKI BS (1), Jacob ROSEN Ph.D. (1,2), Lily CHANG, MD (2), Mika N. SINANAN MD Ph.D. (2,1), Blake HANNAFORD, Ph.D. (1,2)

(1) Department of Electrical Engineering, (2) Department of Surgery, University of Washington, Seattle, WA 98195-2500, USA

<rosen, lchang, mssurg, blake > @u.washington.edu
Biorobotics Lab: http://brl.ee.washington.edu
Center of Videoendoscopic Surgery: http://depts.washington.edu/cves/

ABSTRACT

Surgical robotic systems and virtual reality simulators have introduced an unprecedented precision of measurement for both tool-tissue and tool-surgeon interaction; thus holding promise for more objective analyses of surgical skill. Integrative or averaged metrics such as path length, time-to-task, success/failure percentages, etc., have often been employed towards this end but these fail to address the processes associated with a surgical task as a dynamic phenomena. Stochastic tools such as Markov modeling using a 'white-box' approach have proven amenable to this type of analysis. While such an approach reveals the internal structure of the of the surgical task as a process, it requires a task decomposition based on expert knowledge, which may result in a relatively large/complex model. In this work, a 'black box' approach is developed with generalized cross-procedural applications., the model is characterized by a compact topology, abstract state definitions, and optimized codebook size. Data sets of isolated tasks were extracted from the Blue DRAGON database consisting of 30 surgical subjects stratified into six training levels. Vector quantization (VQ) was employed on the entire database, thus synthesizing a lexicon of discrete, task-independent surgical tool/tissue interactions. VQ has successfully established a dictionary of 63 surgical code words and displayed non-temporal skill discrimination. VQ allows for a more cross-procedural analysis without relying on a thorough study of the procedure, links the results of the black-box approach to observable phenomena, and reduces the computational cost of the analysis by discretizing a complex, continuous data space.

1. INTRODUCTION

Modern surgical trainees and preceptors face decreasing time and resources for training and evaluation as well as pressure for consistent self-accreditation directed towards the general surgical community. This in turn motivates the development of more consistent, objective, and computerized tools for surgical skill evaluation. The implementation of virtual reality (VR) simulation and robotic interfaces into surgical tasks has allowed the extraction of quantitative metrics. Much work has been devoted to translating such data into a scoring of surgical skill. While many of these findings reveal generally significant performance metrics such as completion time, path length, and

success/failure percentages, they often neglect other physical variables inherent in surgical tasks [1,2,3].

A complete, quantitative characterization of surgical skill is a very ambitious undertaking. Such a system must include all spatial, temporal, and energetic aspects of the surgeon/patient interaction, the interactions between those factors as well as the context of the specific organ/tissue/procedure, and the consequences of each manipulation in terms of patient outcome. The eventual goal is an algorithm which processes surgeon actions, measured by a variety of sensors simultaneously, in the context of a specific task or surgical procedure.

Stochastic techniques such as hidden Markov modeling (HMM), taking advantage of their success in speech recognition [4], hold much promise to robustly capture the relevant temporal information of surgical tasks. It was shown that HMM is amenable to the analysis of telemanipulation [5] but that it is more successful when developed with human knowledge of the manipulation task. Moreover, HMMs were applied in manipulation data with particular emphasis on context dependence [6]. This 'white-box' approach of pre-programmed, built-in human knowledge is amenable to surgical analysis and utilizes dynamic surgical characteristics [7,8].

A discrete HMM can be utilized instead of its continuous counterpart through the use of Vector Quantization (VQ). This discretization strategy has been used in speech processing on HMM's [9,10]. For a surgical application, VQ allows the compression of a high-dimensional input data vector, having both continuous and discrete components, into a single "codeword" for each unit of time. The traditional implementation has been the *k-means* algorithm [11,12,13]. Rosen, et al., have shown that this approach was successful when aided by built-in expert human knowledge. While this VQ approach proved successful in processing surgical data, it remains dependent on a task decomposition based on highly specialized expert knowledge (a "white-box" approach) and may result in an unnecessarily complex model. Such requirements may prove undesirable or forbidding for efficiently generalizing the application of this VQ-HMM approach across different surgical procedures. Image processing applications catalyzed much of the development and optimization of VQ algorithms themselves [14] and a variety of more specialized VQ algorithms exists. These are surveyed in [15]. For surgical applications, a category of 'greedy' VQ algorithms appears particularly relevant [16,17]. In this paper, we compare the performance of four different VQ algorithms with surgical data and investigate the ability for VQ alone to differentiate surgical skill.

2. METHOD/TOOLS

Data sets of isolated surgical tasks were extracted from a database acquired with the Blue DRAGON consisting of 30 surgical subjects stratified into six training levels, each completing a laparoscopic porcine task of bowel suturing [18]. The input vector included forces, torques, and velocities in the xy plane and along the z axis (aligned with the tool' shaft) expressed with respect to a coordinate system located at the port of each tool, aligning with grasping force, angular velocity of the handle and binary contact information. The data streams of all 30 participants were concatenated into a single sequence and each channel was normalized via a linear scaling to an interval of [-1 1]. Binary/discrete data were scaled directly while continuous data channels were normalized based on their 97^{th} percentile.

Vector quantization was employed on this scaled, concatenated database, thus synthesizing a lexicon of discrete tool/tissue interactions. Towards this end, three variant VQ training algorithms were utilized and compared: (1) the generalized Lloyd algorithm (GLA, Method I), (2) a variant of GLA that increments codebook size by 1 instead of

doubling it, and a modified GLA which also increments its codebook size by 1 but has a 'greedy' criterion for best word-choice. The traditional k-means algorithm, a well-known VQ method, could not be implemented due to the large size of the database. Each VQ algorithm is briefly described in Table I. The algorithm exhibiting best performance based on lowest final distortion was chosen. Codebook size 'M' was established by examining a relative 'knee' characteristic in the distortion vs. codebook-size curve, as well as some less subjective sudden-drop-in-distortion artifacts. Once chosen, this codebook encoded each subject's scaled data. Hence for every unit of time, a multi dimensional vector of continuous data is mapped to the closest discrete codeword, taken from a codebook of established size M.

	GLA (Method I)	*Modified GLA (Iterated Method I)*	*Full Search "Greedy" VQ Algorithm*	*Traditional K-means Algorithm*
Step 1	Place first word at mean vector of data	Place first word at mean vector of data	Place first word at mean vector of data	Initially choose code-book size N and randomly distribute N points in the data space
Step 2	Increase the size of the codebook by 2^n, splitting all code-words	Increase the size of the codebook by 1, splitting the most populous codeword(s)	Find the codeword giving largest distortion drop when split (full search through codebook, using step 4 each time) and split it	Iteratively migrate each codeword towards local point clusters to min-imize global distortion
Step 3	Relocate each word until a (local) minimum of distortion is reached	Relocate each word until a (local) minimum of distortion is reached	Relocate each word until a (local) minimum of distortion is reached (using GLA techniques)	Continue step 2 until the (percent) change in distortion is less than threshold
Step 4	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Continue steps 2-3 until the (percent) change in distortion is less than threshold	Repeat steps 1-3 with different random initializations to approach global extrema
Normalized Execution Time*	1	72	1440 (degrades greatly w/ larger book sizes)	Forbiddingly Slow (sensitive to initial conditions)
Relative Distortion	High	High	Low	Lowest (theoretically) (highly sensitive to initial conditions)

Table I: *Overview of different VQ algorithms- (*)The execution time factor is normalized with respect to GLA method. Given the currently available computational power, the execution time for identifying 250 codewords using the GLA method on a SUN Ultra Enterprise 450 Platform with dual UltraSPARC™ II processors is 5 min.*

3. RESULTS

Figure 1 depicts three VQ codebook training sessions using the selected algorithms and initializations. Each curve begins at a codebook size of M=1, having a distortion equal to the variance of the entire training data set. Codebook distortion drops with increase in codebook size until it reaches zero when the number of codewords equals the number or samples in the training data. Figure 1 illustrates this curve in the domain of 3 to 250 codewords. Both the Standard GLA (Method I) and iterated GLA yield the highest distortion values. The two runs (different random seed initializations) of the 'greedy' word-choice algorithm gave lowest overall distortion and exhibited more convergent behavior. While training time for the 'greedy' trials increased rapidly with higher codebook size, it was acceptable for codebooks smaller than 250.

Both a 'knee' characteristic and occasional sudden drops in distortion appear in the curves. Selecting Run #2 of the 'greedy' VQ algorithm for lowest distortion, a small but discrete drop at size M=63 suggests that it is a good codebook size for the training set. Each subject's data was grouped into appropriate experience levels of *experts* and 1st through 5th year *residents* (R1…R5) then encoded with this codebook. Histograms of the percent of

time each codeword was in use, normalized to the average task completion time of that group were plotted in Fig. 1 B-F. First-year residents (R1) were compared with experts to illuminate the differences between the skill levels. Figure 1B shows the percentage of time each codeword was in use as taken from the Expert or R1 group, whichever was larger. The shading scheme corresponds to this value and identifies the codewords in all subsequent figures. Figure 1C illustrates the logarithmic ratios of each codeword frequency between the R1s and experts. For example, R1s used codeword 26 approximately 20 times more than experts, both groups used codeword 54 approximately equally, and experts never used codeword 39 (Fig 1C) while R1s used it more than two percent of their task time (Fig 1A). Considering that average R1 task time was 12 minutes, this percentage appears to be significant. Because the darker shading corresponds to more time-in-use, the darker ratio bars might be given greater weight in assessment of skill.

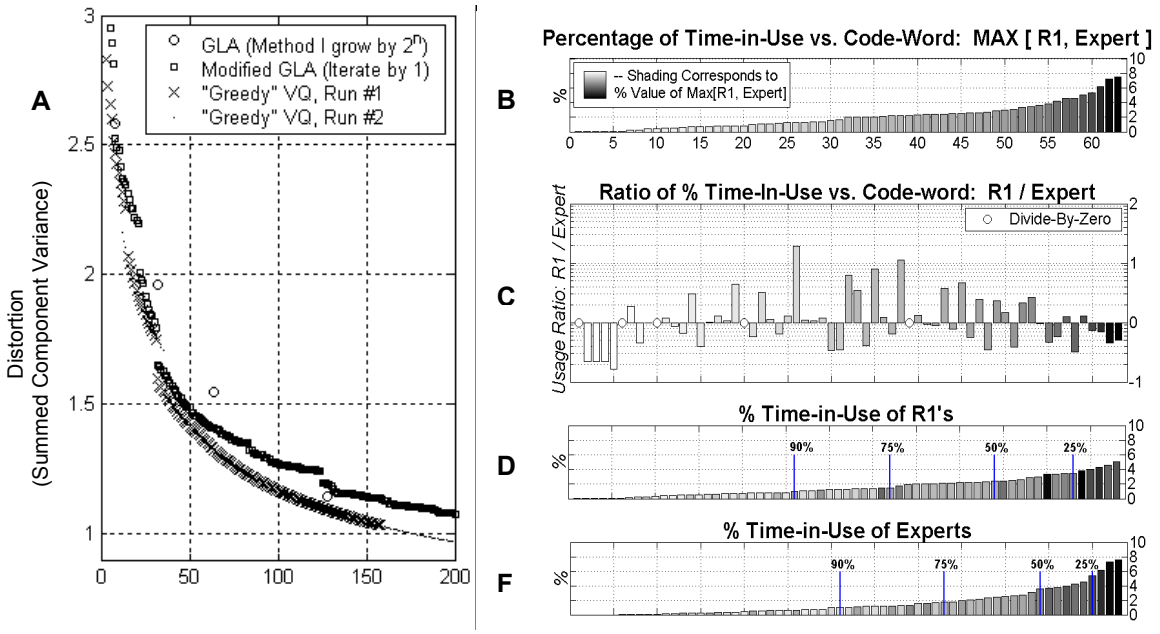
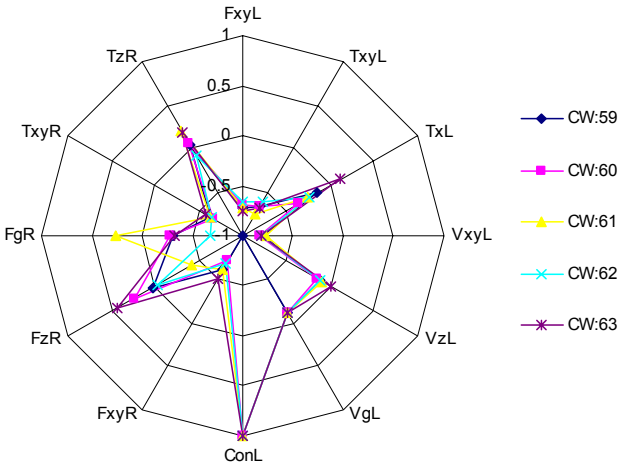


Figure 1: *Percentage of time experts and 1st-year residents (R1) spent using each codeword. Max[R1, Expert] refers to the larger of the value from Expert or R1 levels. (A) and (B) are sorted by the value of Max[R1, Expert], (C) and (D) are sorted by their title value. The shading scheme is consistent for all axes and acts as an identifier.*

Figure 1 D and 2F shows the expert and beginner codeword histograms sorted by frequency, but separately for each group. Beginners appear to use a larger variety of codewords. Experts spend 25% of their task time using only four words which in fact are the darkest, as opposed to the 6 words used in the top 25% of the R1's task time. These differences are also evident at the 50%, 75%, and 90% levels. It should also be noted that code-words used most often by experts (the three darkest bars) are used less often by beginners. Figure 2 summarizes the 5 normalized most frequently used code-words (code word 59-63) in the database.



Range	Max	Min
Norm	1	-1
FxyL [N]	4.745	0.059
Txy [Nm]	1.326	0.046
Tz [Nm]	0.023	–0.016
Vxy [Rad/s]	0.015	0
Vz [m/s]	0.029	–0.026
Vg [Rad/s]	0.0480	-0.0390
ConL	1.000	0.000
FxyR	4.863	0.0798
FzR [N]	36.002	-41.959
FgR [N]	29.069	-5.419
Txy [Nm]	1.021	0.0211
TzR [Nm]	0.0190	-0.020

Figure 2: *The Five most commonly used normalized code-words marked as 59-63 with the appropriate correspondence to Fig. 1 (Definitions: Fxy, Txy- force and torque in the xy plane; Fz, Tx - force and torque along the long shaft of the endoscopic instrument; Vxy/g- Angular velocity; Vz –Linear Velocity ; ConL – Tool/Tissue contact; L – left Tool; R- right tool). Each signal was normalized into a range of [1 -1] which is translated into the Max/Min values appeared in the table.*

4. DISCUSSION AND CONCLUSIONS

Analyzing the data with four VQ methods indicates that a 'greedy' VQ algorithm provided a codebook with the lowest distortion while adequately coped with the complexity level of the surgical database. A code-book with 63 code-words performed well with the bowel suturing data. The VQ successfully established a dictionary of surgical vocabulary and even displayed an ability to objectively differentiate surgical skill level. It should be stressed that this was an entirely 'black-box' approach that did not rely on pre-existing human knowledge of the task to achieve its capability to distinguish surgical skill level. Moreover, this approach affords a significant data reduction method that not only allows a transfer from a high-dimensional continuous space to a finite set of discrete values but also allows the mixing of data types such as binary tissue contact information (context) with continuous force-torque signatures. While in itself this technique shows potential to differentiate experience levels, it does so without any temporal considerations. Future improvement in skill discrimination should be possible once these code-words are processed by a temporal process such as Hidden Markov Model (HMM).

The approach of choosing optimal codebook size presented in this paper is undesirably subjective. While some analytical methods do exist for quantifying optimality, they were not yet explored. For significantly large data sets of high complexity, training the codebook can be time consuming. However, this step needs to be completed only once (offline) and so long as the training data are sufficiently characteristic of subsequently encoded surgical behavior, this issue does not hinder VQ implementation into surgical skill evaluation. Once an optimal codebook is established, input data can be processed (VQ encoded) in real-time [19].

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation (ITR) via collaboration with Greg Hager, Allison Okamura and Russel Taylor from Johns Hopkins University. The authors would like to thank Professor Eve Riskin of the University of Washington for her contributions to this VQ work.

REFERENCES

[1] Verner L., Oleynikov D., Holtmann S., and Zhukov L., Measurements of Level of Surgical Expertise Using Flight Path Analysis from *Da Vinci*™ Robotic Surgical System. Studies in Health Technology and Informatics - Medicine Meets Virtual Reality, Vol. 94, pp.373-378, IOS Press, January 2003.
[2] Moody L., Baber C., and Arvanitis T. N., Objectice Surgical Performance Evaluation on Haptic Feedback. Studies in Health Technology and Informatics - Medicine Meets Virtual Reality, Vol. 85, pp.304-310, IOS Press, January 2002.
[3] Payandeh S., Lomax A., Dill J., Mackenzie C. and Cao C. G. L., On Defining Metrics for Assessing Laparoscopic Surgical Skills in a Training Envirnoment. Studies in Health Technology and Informatics - Medicine Meets Virtual Reality, Vol. 85, pp.334-340, IOS Press, January 2002.
[4] Rabiner L., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, Vol. 77, No. 2, February 1989.
[5] Hannaford B. and Lee P., Multi-Dimensional Hidden Markov Model of Telemanipulation Tasks with Varying Outcomes. Proceedings IEEE Intl. Conf. Systems Man and Cybernetics, Los Angeles, CA, Nov. 1990.
[6] Murphy T., Vignes C., Yuh D. and Okamura A., Automatic Motion Recognition and Skill Evaluation for Dynamic Tasks. Accepted for Eurohaptics 2003, *see http://www.mle.ie/palpable/eurohaptics2003/*.
[7] Rosen J., Hannaford B., Richards C., and Sinanan M., Markov Modeling of Minimally Invasive Surgery Based on Tool/Tissue interaction and Force/Torque Signatures for Evaluating Surgical Skills, IEEE Transactions on Biomedical Engineering Vol. 48. No. 5, pp. 579-591 May 2001.
[8] Rosen J., Solazzo M., Hannaford B., and Sinanan M., Objective Evaluation of Laparoscopic Skills Based on Haptic Information and Tool/Tissue Interactions, Computer Aided Surgery, Volume 7, Issue 1, pp. 49-61 July 2002.
[9] Makhoul J., Roucos S., and Gish H, Vector Quantization In Speech Coding. Proc. IEEE, Vol. 73, No. 11, pp. 1551-1587, November 1985.
[10] Rabiner L. and Juang B. *Fundamentals of Speech Recognition*. (Prentice Hall Signal Processing Series) Englewood Cliffs, NJ: Prentice-Hall Inc, 1993.
[11] S.-T. Bow. *Pattern Recognition: applications to large data set problems*. (Electrical Engineering and Electronics; 23). Mercer Dekker, Inc. New York and Basel, pp. 110-114, 1984.
[12] MacQueen J, Some methods for classification and analysis of multivariate observations. Proc. of the Fifth Berkely Symposium on Math. Stat. and Prob., Vol 1, pp. 281-296, 1967.
[13] Seber, G.A.F., *Multivariate Observations*, Wiley, New York, 1984.
[14] Cosman P., Oehler K., Riskin E. and Gray R.M, Using Vector Quantization for Image Processing. Proc. IEEE, Vol. 91, No. 9, pp. 1326-1341, September 1993.
[15] Gersho A. and Gray R.M., *Vector Quantization and Signal Compression.* (The Kluwer International Series in Engineering and Computer Science) Boston, Dordrecht, London: Kluwer Academic Publishers, 1992.
[16] Riskin E. A. and Gray R.M., A Greedy Tree Growing Algorithm for the Design of Variable Rate Vector Quantizers. IEEE Transactions on Image Proc, Vol. 39, No. 11, pp. 2500-2507, November 1991.
[17] Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. *Classification and Regression Trees* (The Wadsworth Statistics/Probability Series) Belmont, CA: Wadsworth, 1984.
[18] Rosen J., Brown J. D., Barreca M., Chang L, Hannaford B, and Sinanan M., 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 in Health Technology and Informatics - Medicine Meets Virtual Reality, Vol. 85, pp.412-418, IOS Press, January 2002.
[19] Huang C.M., Bi Q., Stiles G. S. and Harris R. W, Fast Full Search Equivalent Encoding Algorithms for Image Compression Using Vector Quantization. IEEE Transactions on Image Processing, Vol. 1, No. 3, July 1992.