Automated Assessment of Surgical Skills Using Frequency Analysis

Aneeq Zia[1], Yachna Sharma[1], Vinay Bettadapura[1], Eric L. Sarin[2], Mark A. Clements[1], and Irfan Essa[1]

[1]Georgia Institute of Technology, Atlanta, GA, USA
[2]Emory University, Atlanta, GA, USA

Abstract. We present an automated framework for visual assessment of the expertise level of surgeons using the OSATS (Objective Structured Assessment of Technical Skills) criteria. Video analysis techniques for extracting motion quality via frequency coefficients are introduced. The framework is tested on videos of medical students with different expertise levels performing basic surgical tasks in a surgical training lab setting. We demonstrate that transforming the sequential time data into frequency components effectively extracts the useful information differentiating between different skill levels of the surgeons. The results show significant performance improvements using DFT and DCT coefficients over known state-of-the-art techniques.

1 Introduction

Timely evaluation and feedback is essential in surgical training. In medical schools, surgical skills are traditionally assessed manually by a supervising surgeon who observes a trainee surgeon performing a procedure. Although, supervision is necessary for resident training, manual evaluation and observing each individual trainee is time consuming and subjective, with known complications [1]. Structured manual grading systems, such as the Objective Structured Assessment of Technical Skills (OSATS) [2] are used in medical schools to alleviate the problem of subjectivity in manual assessments. OSATS covers several assessment criteria such as respect for tissue (RT), time and motion (TM), instrument handling (IH), suture handling (SH), flow of operation (FO), knowledge of procedure (KP) and overall performance (OP). However, manual assessment of each trainee surgeon on a variety of OSATS criteria is still time consuming besides being inherently subjective due to the manual nature of the assessment.

Overall, surgery is a complex task, including, basic surgical skills such as suturing and knot tying that involve hand movements in a repetitive manner. Every surgical resident masters these basic skills before moving on to complicated procedures. Considering the volume of trainees that need to go through basic surgical skills training along with the time consuming and subjective nature of manual OSATS evaluations. Automated assessment of these basic surgical skills can be of a huge benefit to medical schools and teaching hospitals.

In this work, we propose a frequency based video analysis system for OSATS assessment of basic surgical skills such as suturing and knot tying. First, we compute motion features from video data of surgeons performing basic surgical skills to obtain multi-dimensional time series representation of their motions. Then, frequency coefficients of the time series are calculated and used to predict the proficiency level of the surgeon. Our system requires minimal setup, is inexpensive, and is portable for ubiquitous data collection and analysis.
Our contributions are, (1) a novel analysis method of surgical motion without any a priori (and manual) segmentation of the movements; (2) a framework that leverages simple inexpensive equipment with potential for ubiquitous surgical assessment, with easy setup, relieving the time and resource requirements for surgical training in medical schools and; (3) a frequency based analysis method applied to video motion data, which provides better OSATS skill assessment as compared to state-of-the-art techniques.

2 Background

Automated analysis of surgical motion has received attention in recent years [3–7]. The pioneering works addressed skill assessment in robotic minimally invasive surgery (RMIS) [3, 4] and proposed techniques for automatic detection and segmentation of robot-assisted surgical motions. The techniques described in these works are specifically for RMIS and laparoscopic surgeries and have not, to our knowledge, addressed the traditional OSATS based trainee evaluation.

Several RMIS works have used Hidden Markov Modeling (HMM) to represent the surgical motion flow. The motivation for HMMs and gesture based analysis is derived from speech recognition techniques and the goal is to develop *a language of surgery* where a surgical task can be modeled as a sequence of predefined gestures (also known as *surgemes* analogous to phonemes in speech recognition). Some recent works such as [3, 4] have also used linear dynamical systems (LDS) and bag-of-features (BoF) (or Bag-of-Words (BoW)) for surgical gesture classification.These RMIS works provide background and motivation for our work on surgical skill assessment. However, in this work our focus is on OSATS based skill assessment in traditional setting with trainee surgeons practicing basic surgical skills such as suturing and knot tying.

Our goal is to develop an automated, portable and cost effective assessment system that replicates the traditional OSATS assessment without any manual intervention. Some works based on automated assessment of the OSATS criteria for general surgical training have been proposed recently. In [5], the authors introduced Augmented BoW (A-BoW), in which time and motion are modeled as short sequences of events and the underlying local and global structural information is automatically discovered and encoded into BoW models. They classified surgeons into different skill levels based on the holistic analysis of time series data. In [6], the authors proposed Motion Texture (MT) analysis technique in which each video is represented as a multi-dimensional sequence of motion class counts to obtain a frame kernel matrix. The textural features derived from the frame kernel matrix are used for prediction of OSATS criteria. Although,

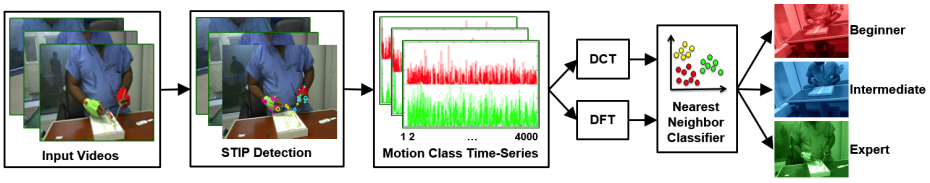

Fig. 1. An overview of our OSATS skill assessment system for suturing and knot tying.

MT technique provided good OSATS prediction, it is computationally intensive ($N \times N$ sized frame kernel matrix for a video with N frames) and does not account for the sequential motion aspects in surgical tasks. A variant of MT, called Sequential Motion Texture (SMT) [7], encoded both the qualitative and sequential motion aspects.

However, None of these past works represent periodic motion elements inherent in basic surgical tasks such as suturing and knot tying. This limitation is addressed in our representation and we show that our proposed method outperforms both traditional techniques like HMMs and the more modern techniques like BoW, A-BoW, MT and SMT.

Some recent skill assessment works in other domains such as competitive sports [8] have used frequency analysis techniques such as Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) to assess the quality of sporting actions. We hypothesize that frequency analysis via DFT and DCT could provide similar results in the assessment of basic surgical skills such as suturing and knot tying due to inherent repetitive motion involved in these tasks.

3 Framework for Skill assessment

The sequential and repetitive nature of basic surgical tasks such as suturing and knot tying results in an inherent dependency on consecutive time series samples. To reveal this time dependency and repetitiveness, we encode the video motion information into a time series, which is then analyzed by using two different frequency analysis methods (DCT and DFT). Our technique involves the following steps: (1) Computing motion features from video data; (2) Generating a time-series from motion features; (3) Computing frequency coefficients of the extracted time series; (4) Selecting the optimum frequencies distinguishing the three skill levels and classifying the skill level of the surgeon. Figure 1 shows the flow diagram of our system. We describe each of these steps in detail:

Step 1: Extraction of motion information from video data: We use computer vision based local features such as Spatio-Temporal Interest Points (STIPs) that have been shown to work well in action classification [9, 10] tasks. Let V be the set containing all videos in our dataset. For all $v \in V$, we use a Harris3D detector to compute the spatio-temporal second-moment matrix μ at each video point (using independent spatial and temporal scales, a separable Gaussian smoothing function and space-time gradients). The final location of the STIPs are given by the local maxima of $H = \det(\mu)$ - $k\text{trace}^3(\mu)$ with a standard parameter setting of $k = 0.0005$ as per the original implementation [10]. Then

we compute 162-element HoG-HoF (histogram of oriented gradients-histogram of optical flow) descriptors $\forall v \in V$ as described in [10].

Step 2: Transformation of motion data into time-series: We use two expert videos for motion class learning (via K-means clustering) since expert motions provide exemplary templates of the surgical task to be evaluated. The learned clusters essentially represent the different moving parts in the video which include arms and tool of the surgeon. For each remaining video, we assign its STIPs to one of the K motion classes learnt by clustering expert videos using minimum Mahalanobis distance. This gives a time series $S \in \Re^{K \times N}$, where N is the number of frames in the video and each element $S(k,n)$ represents the number of STIPs belonging to the n^{th} frame and the k^{th} cluster.

Step 3: Frequency analysis: Recent works such as [8] have used frequency coefficients in assessing the quality of competitive sports, like diving and ice-skating. Since the basic surgical tasks of suturing and knot tying are inherently repetitive in nature, the use of frequency coefficients would essentially be effective in extracting the useful information differentiating between the different skill levels of the surgeons. We use DCT and DFT to obtain frequency coefficients as follows: Let $s_k \in \Re^N$ be the k^{th} dimension of the time series S. We calculate the frequency coefficient vector $\Theta_k \in \Re^N$ by evaluating the expression $\Theta_k = (F s_k^T)^T$ for $k \in [1, 2, \ldots, K]$, where $F \in \Re^{N \times N}$ representing the discrete Cosine/Fourier transformation matrix. For DFT, each element of the transformation matrix is given by $F(m,n) = e^{-j2\pi \frac{mn}{N}}$ for $m \in [0, 1, \ldots, N-1]$ whereas, for DCT, $F(0,n) = \sqrt{\frac{1}{N}}$ and $F(m,n) = \sqrt{\frac{2}{N}} cos(\frac{\pi(2n+1)m}{2N})$ for $m \in [1, 2, \ldots, N-1]$. All the Θ_k's are then concatenated vertically to produce a feature matrix $\Theta \in \Re^{K \times N}$, where each entry in the matrix $\Theta(k,n)$ represents the n^{th} frequency coefficient of the k^{th} dimension of the time series S. Since higher frequencies are likely a result of noisy or abrupt movements, we use the lowest D frequencies producing a reduced sized feature matrix $\bar{\Theta} \in \Re^{K \times D}$. The value of D was then selected empirically depending on the classification accuracy and computation time. The rows of the reduced feature matrix $\bar{\Theta}$ were then concatenated horizontally giving a feature vector $\phi \in \Re^{KD}$ representing the video.

Step 4: Feature selection and skill classification: Since all DCT and DFT frequency coefficients may not be relevant for skill assessment, we perform feature selection to determine a subset of skill defining frequency coefficients. We use Sequential Forward Feature Selection (SFFS) [11] to select a subset of relevant features for each OSATS criteria giving a final feature vector $\hat{\phi} \in \Re^P$, where P is the number of features selected by SFFS. We use a Nearest-Neighbor (NN) classifier with cosine distance metric as a wrapper function for SFFS and select a subset of features with minimum classification error in leave-one-out cross-validation (LOOCV) as reported in [5, 7, 6].

4 Experimental evaluation

Data Acquisition: We recruited 18 participants (surgical residents and nurse practitioners) to collect data for skill evaluation using a standard off-the-shelf

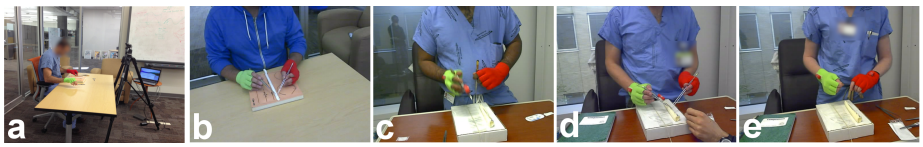

Fig. 2. (a) Setup for data acquisition, (b-e) Sample frames for suturing and knot tying

camera The camera was mounted on a tripod and the participants performed the surgical task wearing colored finger-less gloves. Figure 2 shows the set up for data acquisition along with a few sample frames with participants performing the two surgical tasks. The videos were acquired at 30 frames per second.

We collected two instances for the suturing and knot tying task from each participant, resulting in 36 videos for knot tying and 35 videos for suturing (one video discarded due to data corruption). The number of frames for each video was 4000 for suturing and 1000 for knot tying with the RGB resolution of 640×480 pixels. The videos were captured with varying camera positions and in different rooms to make the dataset invariant to view and illumination changes.

The videos were viewed by an expert (Professor and MD of Surgery and Surgical Translational Studies) who provided the ground-truth OSATS scores. A beginner was given a score of 1, an intermediate was given a score of 2 and an expert was given a score of 3. The distribution of the different skill levels for each OSATS criteria is given in Table 1.

Comparison with State-of-the-Art: All the experiments were performed using leave-one-out cross-validation (LOOCV). We evaluated our proposed technique against five previously published methods: traditional Hidden Markov Models (HMM) and more modern approaches such as Bag of Words (BoW), Augmented-BoW (A-BoW), Motion Textures (MT) and Sequential Motion Texture (SMT).

HMM: Our HMM employed semi-continuous modeling with Gaussian mixture models (GMM) as feature space representations. The $K \times N$ time series was converted into a vector of discrete symbols using k-means with $k \in [3, \ldots, 10]$ and the GMMs were derived by means of an unsupervised density learning procedure.The HMM model is based on linear left-right topologies and are trained using classical Baum-Welch training. Classification is pursued using Viterbi-decoding. HMMs with 4, 8, 10, 12 and 14 states were trained and used in the

Table 1. Distribution of skill levels for different OSATS criteria (Suturing|Knot Tying). NA corresponds to either samples not available or the respective OSATS criteria being not applicable for the task (explained in results section)

	RT	TM	IH	SH	FO	OP
Beginner	5 | NA	13 | 6	13 | NA	14 | 5	12 | 2	NA | 2
Intermediate	20 | NA	11 | 12	10 | NA	13 | 17	14 | 19	NA | 17
Expert	10 | NA	11 | 18	12 | NA	8 | 14	9 | 15	NA | 17

Abbreviations: RT: Respect for Tissue, TM: Time and Motion, IH: Instrument Handling, SH: Suture Handling, FO: Flow of Operation, OP: Overall Performance.

experiments. Classification results were obtained for all possible combinations of states and symbols. A final model with 4 states and 7 symbols was empirically selected. We use a discrete HMM instead of a continuous one in order to have a fair comparison with some of the state-of-the-art methods used in surgical skill evaluation. However, a continuous HMM could also be used in which the observation sequence would simply be the $K \times N$ time series matrix.

BoW/A-BoW: While BoW approaches are good at building powerful and sparser representations of the data, they ignore the ordering information of the particular words and disregard the underlying temporal information. The A-BoW approach [5] attempts to solve this by quantizing time into N bins and encoding them into bag-of-words models using n-grams. We used the BoW and A-BoW code publicly available using the values of $n = 3$ and $N = 5$.

MT/SMT: Motion Texture [6] and Sequential Motion Texture [7] were implemented as proposed in the original papers. As described for OSATS assessment and for computational simplicity, Gray Level Co-Occurrence Matrices (GLCM) texture features with 8 gray levels were used for both MT and SMT. For SMT, the number of windows was set to 10 (determined empirically).

DCT/DFT: We use the DCT coefficients in their original form. However, since the DFT coefficients are complex numbers, we only use their magnitudes in the feature matrix. Once the feature matrix $\Theta \in \Re^{K \times N}$ of the frequency coefficients is obtained, we reduce the number of features by using only the lowest 50 frequency coefficients ($D = 50$) in each dimension of the time series giving the $50K$-dimensional feature vector. Due to a relatively small dataset and to avoid over-fitting, we further reduce the number of features using SFFS feature selection. The number of selected features was empirically set to 30. It was observed that including frequency coefficients greater than 50 (for each dimension) and number of final features greater than 30 did not result in any substantial improvement in the classification accuracy.

5 Results and Discussion

We report the skill classification accuracies of surgeons performing basic surgical tasks of suturing and knot tying for the OSATS criteria applicable and used for each task, except Knowledge of Procedure (KP). KP was excluded since suturing and knot tying are both repetitive tasks and are not procedures with technical progression. Hence, there really is no knowledge component to assess in these tasks. Figure 3 shows the heat maps generated for the classification accuracies for different OSATS criteria using all 7 techniques described and for all the values of K in the motion count time-series. We can see a clear improvement when moving from left to right (HMM to DCT) for all the OSATS criteria and for both the surgical tasks. But overall, all the techniques seem to work well for $K = 6$ and hence those results are presented in Table 1.

In Table 2, the first number corresponds to suturing and the second to knot tying. Some OSATS are not applicable to the knot tying tasks (shown as NA in Table 2) e.g. IH (third column) since there was no instrument used. Suturing scores for OP were not available. For almost all the OSATS criteria, the use of

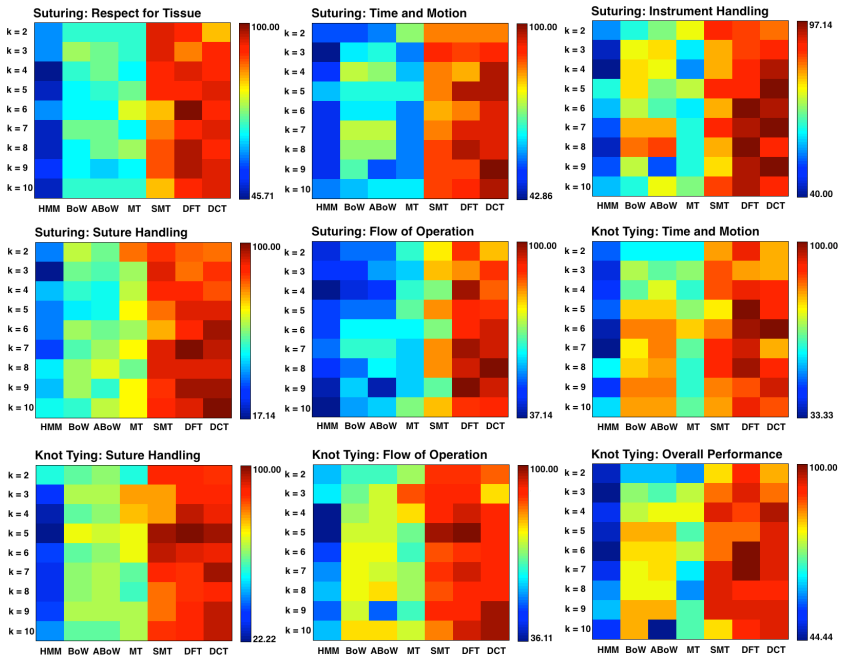

Fig. 3. Heatmaps showing the classification accuracies for the different OSATS criterion for Suturing and Knot Tying. The columns show the proposed DFT and DCT methods and the 5 other method that they are compared against (for the various K in the motion count time-series). We can see a clear improvement in accuracies from left to right (from HMM to DCT).

DCT and DFT significantly improves the classification accuracy for both the surgical tasks as compared to the state-of-the-art methods.

The basic surgical tasks of suturing and knot tying are both sequential periodic activities. An expert surgeon seems to do much more of a task in a given time as compared to a beginner. Thinking in terms of a signal, the expert 'signal' will have a different frequency as compared to a beginner or intermediate. Our technique exploits the periodicity of the two surgical tasks and successfully extracts the frequency coefficients to distinguish the three skill levels of the surgeons. Our technique is also computationally less expensive as compared to SMT which gives reasonably good results too. SMT divides the time series into windows and uses texture analysis on a self-similarity matrix which is computationally expensive. Our technique does not require complex texture analysis and dividing the time-series into windows. One should note that our method is designed for basic repetitive type of surgical motions. For other non-repetitive tasks, frequency based features could be used along with other feature types that are not dependent upon periodicity of the task.

In Summary, we present an automated framework for surgical skill assessment. Using our technique, we classified surgical residents and nurse practitioners into different OSATS skill groups. Our method enables skill assessment without using

Table 2. Percentage of correctly classified subjects (suturing | knot tying) for $K = 6$

Method	RT	TM	IH	SH	FO	OP
HMM	60.0 | NA	48.5 | 36.1	57.1 | NA	37.1 | 36.1	48.5 | 47.2	NA | 44.4
BoW	65.7 | NA	62.8 | 83.3	71.4 | NA	60.0 | 58.3	60.0 | 75.0	NA | 80.5
A-BoW	65.7 | NA	62.8 | 83.3	65.7 | NA	57.1 | 61.1	60.0 | 75.0	NA | 80.5
MT	77.1 | NA	57.1 | 77.7	60.0 | NA	60.0 | 69.4	60.0 | 63.8	NA | 75.0
SMT	82.8 | NA	82.8 | 83.3	80.0 | NA	74.2 | **94.4**	68.5 | 86.1	NA | 86.1
DFT	**100** | NA	85.7 | 97.2	**97.1** | NA	88.5 | 91.6	91.4 | 88.8	NA | **100**
DCT	91.4 | NA	**94.2** | **100**	94.2 | NA	**97.1** | 91.4	**94.2** | **91.4**	NA | 91.4

Abbreviations: RT: Respect for Tissue, TM: Time and Motion, IH: Instrument Handling, SH: Suture Handling, FO: Flow of Operation, OP: Overall Performance.

time windowing, texture analysis or manually defined surgical gestures. The proposed system is simple, easy to setup and is cost effective to deploy.

References

1. Dennis, B.M., Long, E.L., Zamperini, K.M., Nakayama, D.K.: The effect of the 16-hour intern workday restriction on surgical residents' in-hospital activities. Journal of Surgical Education **70**(6) (2013) 800–805
2. Martin, J., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British Journal of Surgery **84**(2) (1997) 273–278
3. Haro, B.B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: MICCAI 2012. Springer (2012) 34–41
4. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical Image Analysis **17**(7) (2013) 732–745
5. Bettadapura, V., Schindler, G., Plötz, T., Essa, I.: Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In: CVPR, IEEE (2013)
6. Sharma, Y., Plötz, T., Hammerla, N., Mellor, S., Roisin, M., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Automated surgical OSATS prediction from videos. In: ISBI, IEEE (2014)
7. Sharma, Y., Bettadapura, V., Plötz, T., Hammerla, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Video based assessment of OSATS using sequential motion textures. In: International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-Workshop. (2014)
8. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: ECCV. Springer (2014) 556–571
9. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, IEEE (2008) 1–8
10. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
11. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern recognition letters **15**(11) (1994) 1119–1125