International Journal on Computer Assisted Radiology and Surgery manuscript No.
(will be inserted by the editor)

Automated Video-Based Assessment of Surgical Skills for Training and Evaluation in Medical Schools

Aneeq Zia · Yachna Sharma · Vinay Bettadapura · Eric L. Sarin · Thomas Ploetz · Mark A.Clements · Irfan Essa

Received: date / Accepted: date

Abstract *Purpose*: Routine evaluation of basic surgical skills in medical schools requires considerable time and effort from supervising faculty. For each surgical trainee, a supervisor has to observe the trainees in-person. Alternatively, supervisors may use training videos, which reduces some of the logistical overhead. All these approaches however are still incredibly time consuming and involve human bias. In this paper, we present an automated system for surgical skills assessment by analyzing video data of surgical activities.
Method: We compare different techniques for video-based surgical skill evaluation. We use techniques that capture the motion information at a coarser granularity using symbols or words, extract motion dynamics using textural patterns in a frame kernel matrix, and analyze fine-grained motion information using frequency analysis.
Results: We were successfully able to classify surgeons into different skill levels with high accuracy. Our results indicate that fine-grained analysis of motion dynamics via frequency analysis is most effective in capturing the skill relevant information in surgical videos.
Conclusion: Our evaluations show that frequency features perform better than motion texture features, which in-turn perform better than symbol/word based features. Put succinctly, skill classification accuracy is positively correlated with motion granularity as demonstrated by our results on two challenging video datasets.

Keywords Surgical skill · Classification · Feature modeling

Aneeq Zia
Georgia Tech, College of Computing
801, Atlantic Drive
Atlanta, GA 30332
Tel.: +1-404-924-0313
E-mail: aneeqzia@gmail.com

1 Introduction

Surgical skill development, i.e., the process of gaining expertise in procedures and techniques required for professional surgery, represents an essential part of medical training. Acquiring high quality surgical skills is a time-consuming process that demands expert supervision and evaluation throughout all stages of the training procedure. However, the manual assessment of surgical skills poses a significant resource problem to medical schools and teaching hospitals and results in complications in executing and scheduling their day-to-day activities [1]. In addition to the extensive time requirements, manual assessments are often subjective and domain experts do not always agree on the assessment scores. This is evidenced by studies that show poor correlations between subjective evaluations and objective evaluations through standardized written and oral exam [2].

Surgery is a complex task and even basic surgical skills such as suturing and knot tying (that involve hand movements in a repetitive manner) require every surgical resident to go through training in order to master these basic skills before moving on to more 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 evaluation, automated assessment of these basic surgical skills can be of tremendous benefit to medical schools and teaching hospitals.

Medical literature recognizes the need for objective surgical skill assessment in surgical training [4]. Yu et al. [5] have suggested evaluations from residents and interns who frequently supervise the students instead of the consultant surgeons who do not have the opportunity to directly observe the medical students. However,

Table 1 Summary of the OSATS scoring system [3]. The score is a Likert scale from levels 1-5 but the guidelines are provided only for levels 1, 3, and 5. The diversity of the criteria, lack of guidelines for all levels, and the need to manually observe each surgeon, makes the manual OSATS scoring a time consuming and challenging process.

Score	Respect for tissue (RT)	Time and motion (TM)	Instrument handling (IH)	Suture handling (SH)	Flow of operation (FO)	Knowledge of procedure (KP)	Overall performance (OP)
1	Unnecessary force on tissue, caused damage	Unnecessary moves	Inappropriate instrument use	Repeated entanglement, poor knot tying	Seemed unsure of next move	Insufficient knowledge	Very poor
2	–	–	–	–	–	–	–
3	Occasionally caused damage	Some unnecessary moves	Occasionally stiff or awkward	Majority of knots placed correctly	Some forward planning	Knew all important steps	Competent
4	–	–	–	–	–	–	–
5	Minimal tissue damage	Economy of movement	Fluid movements	Excellent suture control	Planned operation	Familiarity with all steps	Clearly superior

the subjectivity and time-consuming nature of these evaluations still cannot be ruled out.

Structured grading systems such as the Objective Structured Assessment of Technical Skills (OSATS) [3] have been developed to reduce the subjectivity. Table 1 summarizes the OSATS scoring system. OSATS consists of seven generic components of operative skill that are marked on a 5 point Likert scale. OSATS criteria are diverse and depend on different aspects of motion. For instance, qualitative criteria such as "respect for tissue" depend on overall motion quality while sequential criteria such as "time and motion" and "knowledge of procedure" depend on motion execution order.

A major drawback of manual OSATS assessment is the substantial requirements on time and resources involved in getting several staff surgeons to observe the performance of trainees. However, only few research efforts have addressed automated OSATS assessments for surgical teaching evaluations. For instance, Datta et al. [6] defined surgical efficiency score as the ratio of OSATS "end product quality score" and the number of detected hand movements. Their results indicate significant correlations between the overall OSATS rating and the surgical efficiency. However, they did not correlate the hand movements to individual OSATS criteria. It is important to provide automated assessment on individual OSATS criteria since several studies have demonstrated its efficacy for objective assessment of surgical skills [7].

In this work, we analyze different features and classification back-ends that have been used for automated classification of surgical skills using video data. We note that most of the features are built upon basic spatio-temporal motion attributes such as Histogram of Gradients (HoG) and Histogram of Flow (HoF) features. These basic motion features in videos can be represented by a time series of symbols (or words) as in Hidden-Markov-Models (HMMs), Bag-of-Words (BoW) and Augmented-BoW (ABoW) techniques. The motion dynamics can also be represented as textural variations in a frame kernel matrix representing the similarity between two frames using a kernel function. Furthermore, since surgical motion for basic surgical skills (suturing and knot tying) is inherently repetitive, the periodicity of motion can be captured by frequency based features such as Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT).

We note that classification accuracy increases progressively as we move from coarse word-based (symbolic) features to fine grained frequency based features. Our results on two independently acquired and challenging data-sets demonstrate that frequency based features are well suited for automated video-based assessment of surgical skills.

Contributions: (1) Comparison of state-of-the-art techniques for video-based automated assessments of OSATS; (2) Analysis of three different types of features (symbolic, texture based, and frequency based) within an automated generalized video based assessment framework; and (3) Evaluation of the various techniques on two independently acquired challenging data-sets.

2 Background

Automated analysis of surgical motion has gained attention in recent years [8–20]. Pioneering works addressed skill assessment in robotic minimally invasive surgery (RMIS) and proposed techniques for automatic detection and segmentation of surgical motions assisted by robots [15–20]. However, the techniques described in these works are specifically for RMIS and laparoscopic surgeries and, to the best of our knowledge, have not

Table 2 Related works on surgical video analysis. CNN: Convolutional Neural Network, DCT: Discrete Cosine Transform, DFT: Discrete Fourier Transform, MT: Motion Texture, SMT: Sequential Motion Texture, CRF: Conditional Random Field, BoW: Bag-of-Words, ABoW: Augmented Bag-of-Words, LDS: Linear Dynamical Systems, DTW: Dynamic Time Warping, CCA: Canonical Correlation Analysis, HMM: Hidden Markov Model

Reference	Technique	Gesture	Analysis goal	Data
Twinanda (2016) [8]	CNN	Yes	Surgical tool detection and phase recognition	Laparoscopic cholecystectomy (endoscopic video), 13 subjects
Lea C. (2015) [9]	CRF	Yes	Surgical action segmentation and recognition	RMIS (both kinematic and video data from robotic surgery), 8 subjects
Zia (2015) [10]	DCT,DFT	No	OSATS classification	General suturing task (only video data), 16 subjects
Sharma (2014) [11, 12]	MT,SMT	No	OSATS prediction, classification	General suturing task (only video data), 16 subjects
Tao (2013) [13]	CRF	Yes	Surgical gesture segmentation and recognition	RMIS (both kinematic and video data from robotic surgery), 8 subjects
Bettadapura (2013) [14]	ABoW	No	OSATS classification	General suturing task (only video data), 16 subjects
Haro (2012), Zapella (2013) [15,16]	BoW, LDS	Yes	Surgical gesture recognition	RMIS (both kinematic and video data from robotic surgery), 8 subjects
Padoy (2012) [17]	DTW, HMM	Yes	Surgical phase recognition	Laparoscopic cholecystectomy (endoscopic video), 4 subjects
Lalys (2011) [18]	DTW	Yes	Surgical phase recognition	Cataract surgery, 20 videos
Blum (2010) [19]	CCA, HMM	Yes	Surgical phase recognition	Laparoscopic surgery, 10 videos
Lin (2009) [20]	HMM	Yes	Skill classification but not on individual OSATS criteria	RMIS (both kinematic and video data from robotic surgery), 6 subjects

addressed the traditional OSATS based trainee evaluation.

Automated assessment of basic surgical skills for both RMIS and conventional medical teaching can be categorized based on the approaches used for time series analysis. The local approaches model specific surgical tasks and model the task as a sequence of manually defined surgical gestures [15,16]. On the other hand, the global approaches involve the analysis of the whole motion trajectory without segmentation into surgical gestures [21,6].

Several RMIS works have used Hidden Markov Models (HMMs) 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). Tao et al. [13] proposed a combined Markov/semi-Markov conditional random field (MsM-CRF) model for gesture segmentation and recognition for RMIS.

With advances in video data acquisition, the attention has shifted towards video based analysis in both RMIS and teaching domains. Table 2 summarizes recent work on surgical video data. Most of these classify different surgemes or surgical phases and the data from different types of surgeries are used. Haro et al [15] and Zapella et al. [16] employed both kinematic and video data for RMIS surgery. They used linear dynamical systems (LDS) and bag-of-features (BoF) for surgical gesture (surgeme) classification in RMIS surgery. Twinanda et al. [8] proposed a CNN architecture, called *EndoNet*, for phase recognition and tool presence detection in Laparoscopic cholecystectomy. Lea et al. [9] developed a method to capture long-range state transitions between actions by using higher-order temporal relationships using a variation of the Skip-Chain Conditional Random Field. These works have mainly focused on RMIS and do not address assessment of OSATS criteria as done in general surgical training.

Some works based on automated assessment of the OSATS criteria for general surgical training have also been proposed recently. In [14], the authors introduced Augmented BoW (ABoW), 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 [11], the authors proposed Motion Texture (MT) analysis technique in which each video is represented as a multidimensional 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 MT technique provided good OSATS prediction, it is computationally intensive ($N \times N$ sized frame kernel matrix for a video with

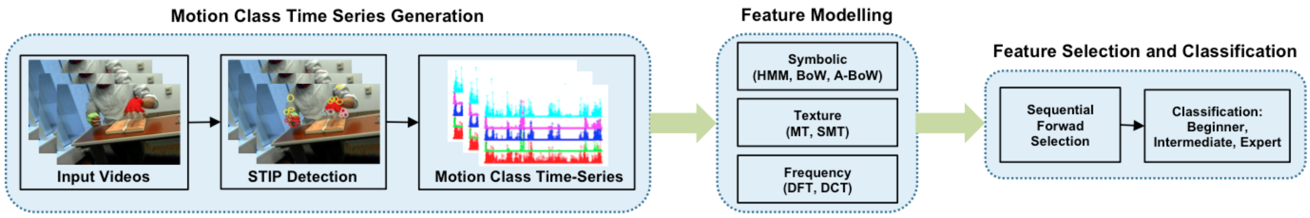

Fig. 1 Overview of the system used for skill assessment.

N frames) and does not account for the sequential motion aspects in surgical tasks. A variant of MT, called Sequential Motion Texture (SMT) [12], encoded both the qualitative and sequential motion aspects.

Some recent skill assessment works in other domains such as competitive sports [22] have used frequency analysis techniques such as Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) to assess the quality of sporting actions. OSATS skill criteria depend on the different characteristics of the motion performed by the surgeon (Table 1). For instance, an expert surgeon's movements are smooth with no unnecessary moves as compared to stiff movements of a novice surgeon. Thus, we need to analyze the changing motion characteristics (motion dynamics) in the surgical video. In addition, suturing and knot tying are inherently repetitive tasks. Inspired by these advances, a recent work used DFT and DCT features for automated video based skill assessment [10].

Our goal is to develop an automated, portable and cost effective assessment system that replicates the traditional OSATS assessment without any manual intervention. The 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. We note that video based OSATS assessment techniques mainly use three types of features (1) Symbolic: HMM, BoW and ABoW; (2) Texture: MT and SMT; and (3) Frequency: DCT and DFT. In this work, we build upon the work in [10] and provide a comparative analysis of these features in a generalized framework for video-based skill assessment. We test the different feature performances on two independently acquired and diverse data-sets collected in a general surgical lab setting. Our results show that frequency features outperform other feature types previously reported in literature indicating its skill assessment potential for medical schools and teaching hospitals.

3 Methodology

We use video based processing for evaluating the skill level of each surgeon. The videos are initially preprocessed and converted into a multi-dimensional time-series which is then used to extract different types of features which are used for skill classification. Figure 1 shows the proposed pipeline for the system. We have divided the flow into three steps: (1) Motion class time-series generation; (2) Feature modeling; (3)Feature selection and classification. We will now discuss these stages in detail.

3.1 Motion Class Time Series Generation

The first stage in our approach is to encode the motion in the videos and generating a motion class time-series representation of each video. Many different types of motion features have been proposed in the literature for extracting relevant information from video data [23–25]. For our purpose, we use Spatio-Temporal Interest Points (STIPs) [26] proposed by Laptev in order to encode the motion from the videos. Let V be the set containing all the videos in our dataset. Then, for all $v \in V$, a Harris3D detector is used to compute the spatio-temporal second-moment matrix μ at each video point given by

$$\mu = g(.;\sigma^2,\tau^2) * \begin{pmatrix} L_x^2 & L_x & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \quad (1)$$

where $g(.;\sigma^2,\tau^2)$ is a 3D Gaussian smoothing kernel with a spatial scale σ and a temporal scale τ. $L_{x,y,t}$ are

Fig. 2 Clustering STIPs into motion classes.

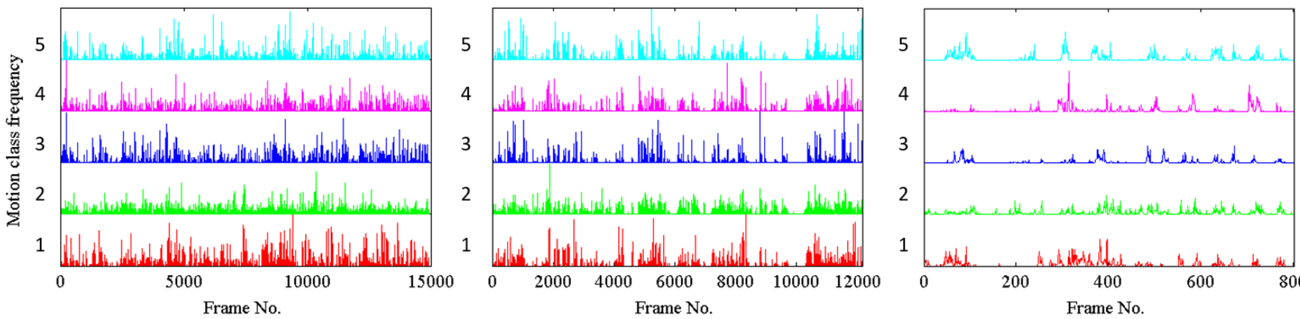

Fig. 3 Motion class time series samples using $K = 5$ for a novice (left), an intermediate (center) and an expert (right) surgeon. Note that the beginner motion is more frequent and exists in almost all frames for all motion classes as compared to fewer motion for intermediate and expert surgeons. These sample plots were obtained from dataset-B (see the "Data Collection" section for description of the dataset), represented by varied length of the time series.

gradient functions along the x, y and t domains. The final position of the STIPs is then calculated by finding the local maxima of the Harris corner function given by

$$H = \det(\mu) - \omega(\text{trace}(\mu))^3 \tag{2}$$

We use Laptev's STIP implementation [27] with default parameters and sparse feature detection mode for different spatio-temporal scales with ω set to be 0.005. We then compute Histogram of of Optical Flow (HOF) and Histogram of Oriented Gradients (HOG) on a three-dimensional video patch in the neighborhood of each detected STIP. A 4-bin HOG and a 5-bin HOF descriptor is calculated resulting in 72-dimensional HOG vector and a 90-dimensional HOF vector. The final feature vector for each STIP is obtained by concatenating HOG and HOF vectors resulting in a 162-dimensional vector.

Once the STIPs for all videos are extracted, we learn motion classes by using k-means clustering on STIPs from two expert videos. Expert STIPs are used since they are more distinct and uncluttered as compared to non-experts. Therefore, expert motions provide exemplary templates for the surgical task to be evaluated. The STIPs from experts are clustered using k-means for different number of clusters 'c'. Figure 2 shows a sample frame with STIPs extracted and the cluster assignment of each STIP. The different colors in the right image correspond to different clusters. The learned clusters can be thought of as representing of the number of moving parts in the video as evident in Figure 2 where you can see different colored STIPS for the different moving parts like hands, arms, instrument etc. The expert clusters are then used to transform the remaining videos in the data set into a multi-dimensional time series. This is done by assigning each STIP in every frame of the video to one of the 'c' learned clusters using minimum Mahalanobis distance from the cluster distribution. This results in a time series $T \in \Re^{K \times N}$ representing each video, where K represents the dimension of the time series (equivalent to the number of clusters used in k-means) and N is the number of frames of the video. Figure 3 shows some sample motion class time series for a beginner, intermediate and an expert using $K = 5$.

3.2 Feature Modeling

The features we use for our analysis are divided into three categories: (1) Symbolic features; (2) Texture features; and (3) Frequency features. The different type of features in each category is described below. Note that for description of each technique, we will use $X \in \Re^{K \times N}$ to denote a time series where K is the dimension of the time series and N being the number of frames of the video

3.2.1 Symbolic Features

Previous state-of-the art has mostly focused on words-based/symbolic methods for describing video and time-series data for a variety of application like activity recognition and skill categorization. In this category, we use HMMs [28, 29], Bag-of-Words (BoW) and Augmented Bag-of-Words (ABoW) models [14, 15].

HMM: We implemented HMM using semi-continuous modeling with Gaussian mixture models (GMM) representing the feature space [29]. We used k-means clustering using different number of clusters to convert the multi-dimensional time series data into a set of discrete symbols n. The GMMs were obtained using an unsupervised density learning procedure. The HMM was trained using the classical Baum-Welch training for different number of states s and classification was done using Viterbi-decoding.

BoW: BoW techniques represent the state-of-the-art for video-based activity recognition. The BoW model is typically constructed using visual codebooks derived from local spatio-temporal features. The clusters obtained by clustering the HOG-HOF STIP feature vectors form the vocabulary for our BoW codebook [15]. The STIPs are then mapped to the words in our vocabulary which results in each video being represented by a histogram of words. With this feature representation, we then use a k-Nearest Neighbor (kNN) classification back-end to categorize the videos into the various OSATS skill categories.

ABoW: While BoW models are better than HMMs, standard BoW techniques do not capture the underlying structural information, neither of causal nor of sequential type, that is inherent by the ordering of the words. To solve this problem, [14] introduced the Augmented Bag-of-Words (ABoW) model that represents temporal information by quantizing time and defining new temporal words in a data-driven manner. Furthermore, the model uses n-grams to augment the BoW with the discovered temporal events in a way that preserves the local structural information (relative word positions) of the activity. In addition, to discover the global patterns in the data, the ABoW model uses randomly sampled Regular-Expressions to find patterns across the words within the activities. We built ABoW models by augmenting our BoW models and, like before, used a kNN classification back-end to categorize the skill levels.

3.2.2 Texture Features

Textural features have been shown to give good accuracy for skill classification of surgical skills [12]. We will now describe the computation of texture features for classification.

Motion Texture: Motion Texture (MT) encodes the motion dynamics in a frame kernel matrix which is then used to calculate texture features [12]. The time series $X \in \Re^{K \times N}$, the frame kernel matrix $M \in \Re^{N \times N}$ is calculated using

$$M = \phi(X)'\phi(X) \tag{3}$$

A Gaussian kernel function is used as a kernel function and each element in the kernel matrix M, $m_{i,j}$ denotes the similarity between the frame number i and j and is given by

$$m_{i,j} = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{4}$$

The matrix M is then used to derive textural statistics using Gray-Level Co-occurrence Matrix (GLCM). GLCM is obtained by calculating how often a pixel with a certain intensity level occurs in a specific spatial relationship to a pixel with different intensity level. The final feature vector obtained is 20-dimensional.

Sequential Motion Texture: Sequential Motion Texture (SMT) extends MT by incorporating temporal information into the features [12]. The time series $X \in \Re^{K \times N}$ is first divided into equally sized temporal windows W such that each window contains equal proportion of the STIPs corresponding to largest motion class in a given video. Frame kernel matrices are calculated for each time window using equation 3. The final GLCM features are then calculated for each time window resulting in a $20W$-dimensional feature vector.

3.2.3 Frequency Features

Frequency based features have been widely used in various application exploiting the periodic nature of data. Recently, works of [22] and [10] have shown that frequency features work extremely well for assessing quality of actions like sports and basic surgical tasks. The two types of frequency features used for our evaluation are described below.

Discrete Fourier Transform: Discrete Fourier Transform (DFT) is used to convert data from time domain into frequency domain and has been extensively used for many application across several domains. For our time series $X \in \Re^{K \times N}$, we calculate the frequency coefficients for each dimension independently and concatenate them to form the frequency matrix $Q \in \Re^{K \times N}$ [10]. The i^{th} row in the frequency matrix Q, $Q(i)$ is calculated by

$$Q(i) = \theta X(i)' \tag{5}$$

where $X(i)$ is the i^{th} dimension of the time series X. θ is an $N \times N$ matrix and $\theta(m, n)$ is given by

$$\theta(m, n) = \exp(-j2\pi\frac{mn}{N}), \tag{6}$$

where $\{m, n\} \in [0, 1, \dots, N-1]$. Once the matrix Q is calculated, the higher frequency terms are removed in order to eliminate noise. This results in a reduced matrix $\hat{Q} \in \Re^{K \times F}$ where F denotes the highest frequency component used from each dimension of the time series X. This can also be thought of as low-pass filtering of the time series. The elements of $\hat{Q}$ are then concatenated to form a final feature vector of KF dimensions.

Discrete Cosine Transform: Discrete Cosine Transform (DCT) is also a transformation of data from time domain to frequency just like DFT. However, DCT only uses cosine functions instead of both sines and cosines. This results in the DCT coefficients being real as opposed to DFT where the coefficients can be complex.

Similar to DFT, the i^{th} row of the frequency matrix $Q \in \Re^{K \times N}$ is also calculated using equation 5 [10] but the θ matrix is given by

$$\theta(0, n) = \sqrt{\frac{1}{N}}, \tag{7}$$

$$\theta(m, n) = \sqrt{\frac{2}{N}} cos(\frac{\pi(2n+1)m}{2N}), \tag{8}$$

where $\{m, n\} \in [0, 1, \ldots, N-1]$. Similar to DFT, the matrix Q is reduced to $\hat{Q} \in \Re^{K \times F}$ and a final KF-dimensional feature vector is obtained.

3.3 Feature Selection and Classification

The final feature vector obtained from the previous step may contain many elements that may be redundant (provide no more information) or irrelevant (contain no useful information) to the skill level. In order to tackle this, we reduce the number of elements in the final feature vector by using feature selection. For our experiments, we use Sequential Forward Selection (SFS) to have a fair comparison between different techniques since it has been used before in similar works [12,10].

Given a feature set $\Phi = \{\phi_i | i = [1, \ldots, Z]\}$, SFS aims to find a subset of features $\hat{\Phi} = \{\hat{\phi}_i | i = [1, \ldots, U]\}$, with $U < Z$ by starting with an empty set and sequentially adding the features that maximize the objective function when combined with the features that have already been selected. We use a Nearest-Neighbor (NN) classifier with cosine distance metric as a wrapper function for SFS.

4 Experimental evaluation

4.1 Data Collection

In order to test the performance of the various skill assessment techniques, we collected two datasets in different settings. We will refer to them as "dataset-A" and "dataset-B". In dataset-A, each video was captured for a specified time and there was minimal involvement of any other human, other than the participant. In dataset-B, there were large variations in the length of the video being captured along with delays in the middle of the tasks and people were moving around within the participant's environment adding to the noise in the motion captured. The suturing type performed by participants in both datasets was a '*running suture*' and there were variations in the number of sutures performed by each participant. All the participants in dataset-A were right-handed except for 2,

Fig. 4 Sample frames from the datasets. The top 4 images are from dataset-A and the bottom 2 images are from dataset-B.

whereas information regarding dominant hand for dataset-B was not available. More specific details of data capture for both datasets are given below.

Dataset-A: This dataset contains videos captured from 18 recruited participants (surgical residents and nurse practitioners). A standard camera was used for capturing the videos while the participants performed the surgical tasks wearing colored finger-less gloves. Each participant performs two attempts of suturing and knot tying each, resulting in 36 videos for knot tying and 35 videos for suturing (one video not used due to data corruption). We collected 4000 and 1000 frames for suturing and knot tying respectively, at a resolution of 640×480 pixels and 30 frames per second. The camera was placed at different angles in each attempt and the data was captured in multiple rooms in order to make the dataset invariant to view and illumination changes.

Dataset-B: This data set was collected by recruiting 16 new participants (medical students). Each participant performed suturing activity using a needle-holder, forceps and the tissue suture pads. The session were recorded using a standard camera with 1280×720 pixels and 50 frames per second. Each session was recorded in a separate video. An expert surgeon performed three sessions giving a total of 33 videos. The number of frames for each recording varied largely with the average duration of the videos being 18 minutes each.

Figure 4 shows some of the sample frames from both data sets for Suturing and Knot Tying tasks. Ground truth for the OSATS score for both data sets were obtained by showing the videos to an expert. Two independant experts graded the two datasets respectively. The training data was grouped into three skill levels: beginner ($OSATS \leq 2$) was given a score of 1, an intermediate ($2 \leq OSATS \leq 3.5$) was given a score of 2 and an expert ($3.5 \leq OSATS \leq 5$) was given a score of 3. Table 3 gives the distribution of the different skill levels for each class for the two datasets.

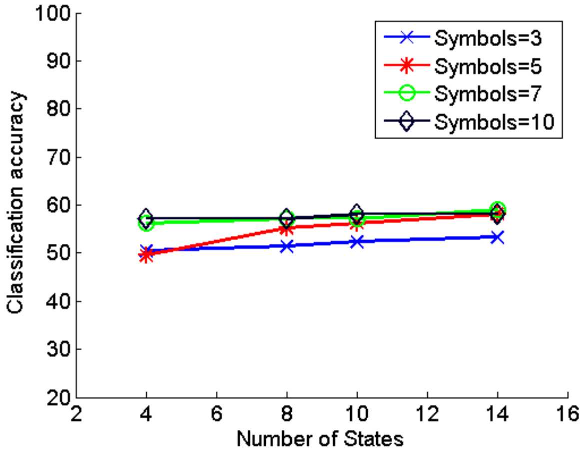

Fig. 5 Plots of average classification accuracy versus number of states with varying number of discrete observation symbols.

4.2 Parameter Estimation

The performance of each of the techniques described in Section 3 are dependent on the values of parameters that we need to learn. We select each of these parameters empirically. The following describes how each parameter (for the different proposed techniques) was selected. All the experiments were performed using leave-one-out cross-validation (LOOCV), where one video was left out for testing in each experiment. Moreover, we use 5-dimensional time series ($K = 5$) for estimating parameters in this section. The optimum parameters are selected based on average classification accuracy $C^K_{avg}(P)$, over all OSATS criteria for a specific parameter set P. This is calculated by $C^K_{avg}(P) = \frac{1}{O}\sum_{o=1}^{O} C^K_o(P)$, where $C^K_o(P)$ represents the classification accuracy for a respective OSATS criteria o and parameter set P using K-dimensional time series, while O denotes the total number of applicable OSATS criteria. The parameter set $\hat{P}$ achieving highest C_{avg} is then used to run experiments for all values of K in the next section.

4.2.1 Symbolic Features

We described three techniques in Section 3 under symbol based feature representation. For BoW and ABoW, the parameters proposed in [14] were used wherein the BoW model was built using 50 clusters and augmented using interspersed encoding with 3-grams, 5 time bins and 20 random regular expressions. For HMM, we learned the optimum value for the number of symbols n and number of states s. We evaluate the classification rate for all combinations of n and s, where $n = [3, 4, ..., 10]$ and $s = [4, 8, 10, 12, 14]$. Figure 5 show a plot showing the variation in the average classification accuracy with respect to varying n and s. The average classification accuracy was calculated by taking the mean of the individual classification percentages achieved. Each plot for a specific number of symbols was achieved by averaging the classification accuracies over all the OSATS criteria for the respective number of states. It can be seen that using $n = 7$ and $n = 10$ seem to work best and equally good and the classification rate stays constant across varying s. However, the training time increases significantly using higher number of states. Therefore, we selected $n = 7$ and $s = 4$ to achieve best possible accuracy while saving computation time.

4.2.2 Texture Features

For both MT and SMT, we use the standard Gray Level Co-Occurrence Matrices (GLCM) with 8 gray levels. However, for SMT, the performance is dependent on the number of time windows W. In order to find the optimum value for W, we calculate the classification rates of varying the number of windows for $W \in [6, 8, 10, 12, 14]$

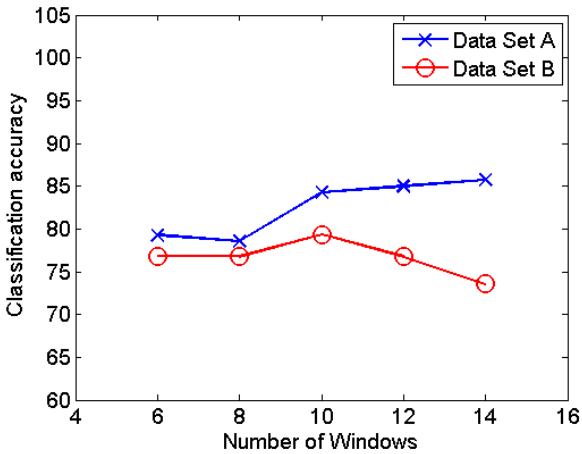

Fig. 6 Plots of classification accuracy versus number of windows.

Table 3 No. of samples for different expertise levels for dataset-A and dataset-B for each of the OSATS criteria (RT: Respect for Tissue, TM: Time and Motion, IH: Instrument Handling, SH: Suture Handling, FO: Flow of Operation, OP: Overall Performance). Within each cell, "S" refers to Suturing and "KT" refers to Knot Tying and "NA" corresponds to either samples not available or the respective OSATS criteria being not applicable for the task.

	Dataset-A (S: Suturing, KT: Knot Tying)						Dataset-B (S: Suturing)				
	RT	TM	IH	SH	FO	OP	RT	TM	IH	SH	FO
Beginner	S: 5 KT: NA	S: 13 KT: 6	S: 13 KT: NA	S: 14 KT: 5	S:12 KT: 2	S: NA KT: 2	S: 2	S: 9	S: 8	S: 10	S: 3
Intermediate	S: 20 KT: NA	S: 11 KT: 12	S: 10 KT: NA	S: 13 KT: 17	S: 14 KT: 19	S: NA KT: 17	S: 14	S: 15	S: 16	S: 15	S: 16
Expert	S: 10 KT: NA	S: 11 KT: 18	S: 12 KT: NA	S: 8 KT: 14	S: 9 KT: 15	S: NA KT: 17	S: 15	S: 7	S: 7	S: 6	S: 12

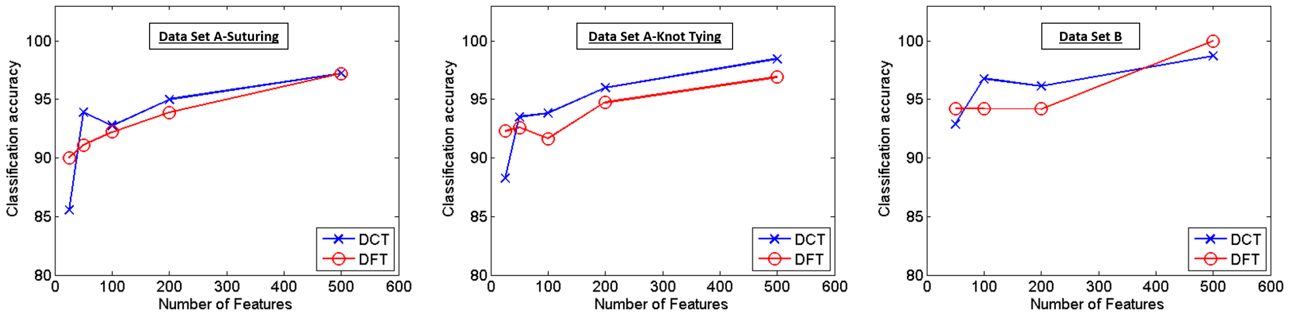

Fig. 7 Plots of average classification accuracy versus highest frequency component used from each dimension of the time series. The left two plots are for dataset-A and the right most for dataset-B

on both data sets. Figure 6 shows a graph for classification rate vs number of windows (W). Again, we average the classification accuracy over all the OSATS criteria applicable for each value of W. As evident from the plots, $W = 10$ seems to work best for both data sets. For dataset-A, the accuracy seem to stay constant after further increasing W whereas for dataset-B, the accuracy deteriorates after 10 time windows. Therefore, we select $W = 10$ for our evaluation and result comparison for SMT.

4.2.3 Frequency Features

As described in Section 3, DCT coefficients are always real values whereas, DFT can have complex coefficients as well. Therefore, the DCT coefficients are used as it is whereas the absolute value of the DFT coefficients is used to make sure they are real valued. For frequency based methods described, the only parameter that needs to be selected empirically is F which is the highest frequency component selected from each dimension of the time series (or the cutoff frequency in the low pass filter). Therefore, we calculate the classification accuracy for $F \in [25, 50, 100, 200, 500]$. Figure 7 shows the plots obtained for classification rate vs number of frequency features used per dimension of the time series. The accuracies were averaged over all OSATS criteria for each value of F. The graphs depict a correlation between average accuracy and number of features (F). We select a value of 500 for both datasets as it embodies a good tradeoff between accuracy and computational time. We maintain $F = 500$ for our evaluation and results comparison.

5 Results

We evaluate the techniques described above on two diverse datasets and report the classification accuracy for the different applicable OSATS criteria. For dataset-A, there are two surgical tasks being assessed: Suturing and Knot Tying. Therefore, we report the classification results attained from the techniques described before on both of them. However, dataset-B only has Suturing task so the results are presented for just that.

Dataset-A: Figure 8 shows the heat maps for the applicable OSATS criteria using the different type of methods described in Section 3. We implement each method for $K \in [2, 3, \ldots, 10]$, where K is the dimension of time series used. It is evident that there is an improvement in the classification as we move from Words/ Symbol based methods to Texture based to Frequency based. SMT, DCT and DFT seem to be the top performing features. Figure 9 shows some more detailed plots of the classification accuracies for a better comparison between the top three methods. Frequency based features perform better than SMT for almost all the

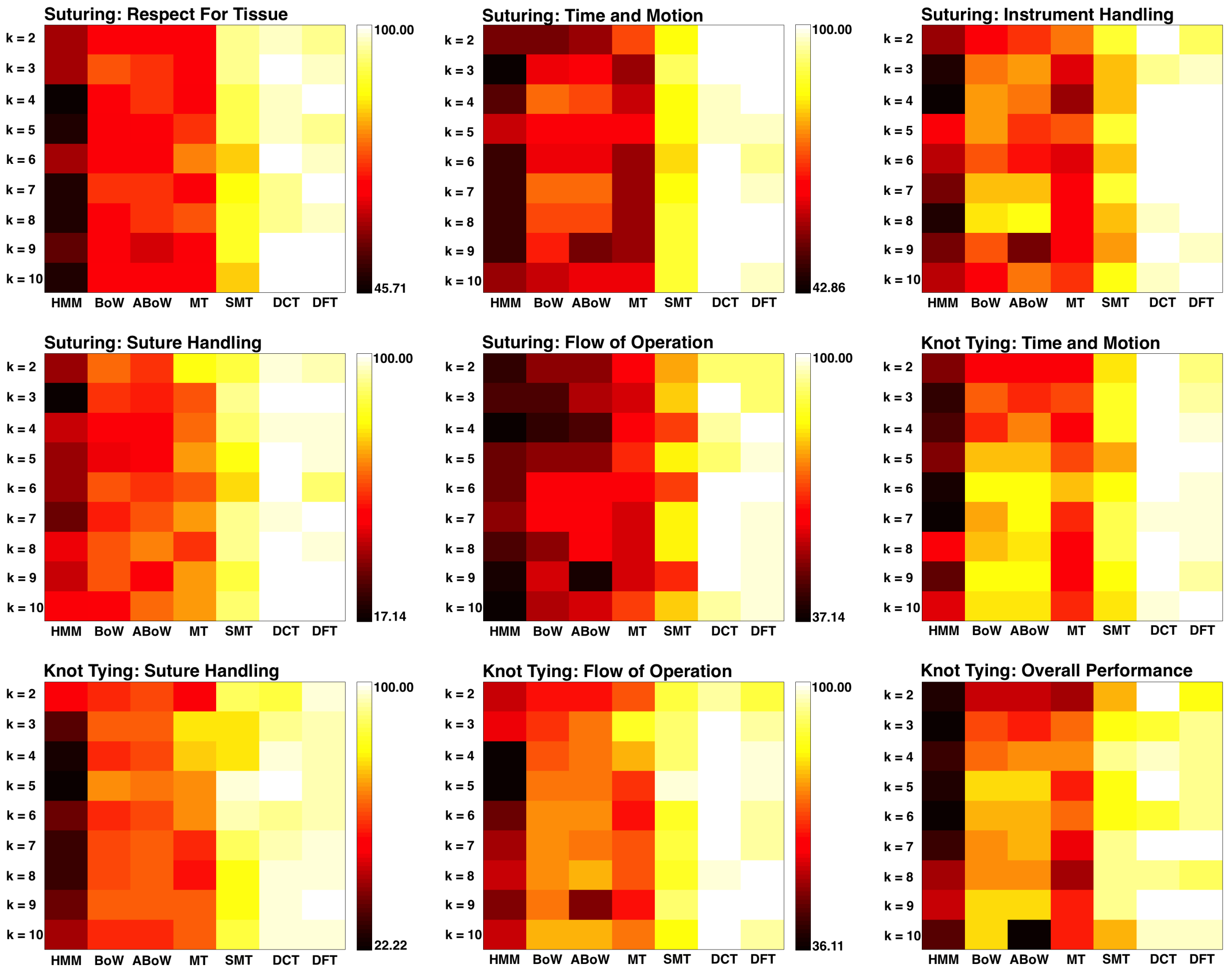

Fig. 8 Heatmaps showing the classification accuracies for the different OSATS criterion for Suturing and Knot Tying. The columns of each heatmap show the different methods. We can see a clear improvement in accuracies from left to right (symbolic features to frequency features).

OSATS criteria and for almost all the values of K. This shows that frequency based features are more robust across the different OSATS criteria and don't seem to depend too much on the dimension of the time series used.

Dataset-B: The results from dataset-A clearly show that words/symbol based method don't seem to capture the information relevant to the skill level of the surgeons performing the basic surgical tasks. Moreover, texture based feature without temporal information perform poorly as well. Since this data set seem more tough due to the variation in the length of the videos and the noisy motion, we only evaluate and compare the features which perform best on dataset-A i.e. SMT, DCT and DFT. Figure 10 shows the classification results obtained using these 3 features. It is clearly evident from the graphs that frequency based features DCT and DFT outperform the best performing texture based feature SMT by a good margin for almost all OSATS criteria and for all values of K.

Table 4 gives the average classification rates for the different techniques on both data sets. Each classification is averaged over all OSATS and over all values of K and is given by the equation

$$C'_{avg} = \frac{1}{9}\sum_{K=2}^{10}\frac{1}{O}\sum_{o=1}^{O} C_o^K(\hat{P}) \tag{9}$$

where $\hat{P}$ was the optimum parameter set found in the previous section. It is clear from the averaged results that frequency based features out perform all other features compared in this paper. DCT seem to be working slightly better than DFT on average.

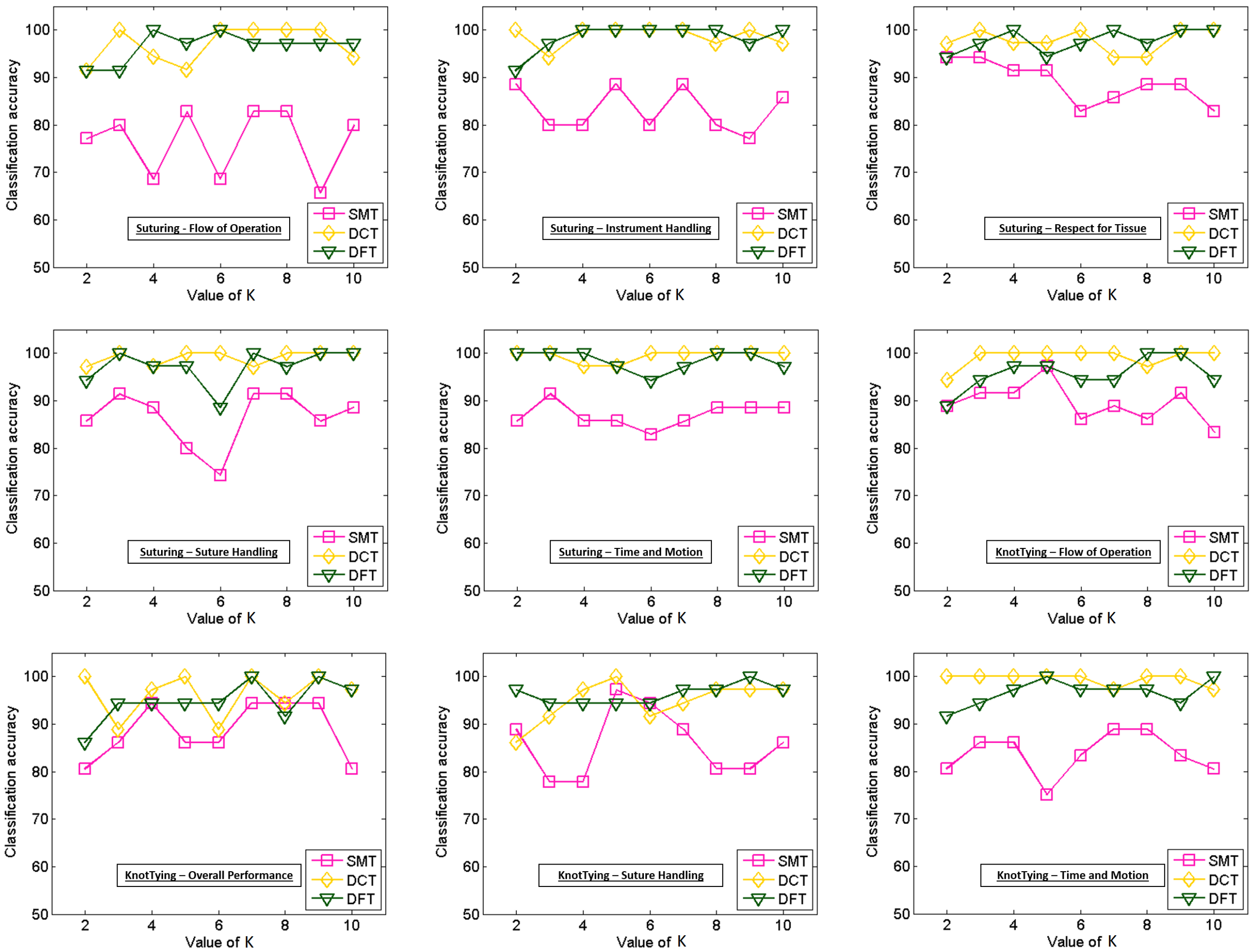

Fig. 9 Plots showing classification rates for various OSATS criteria for dataset-A. The corresponding task (suturing or knot tying) and the OSATS criteria for each plot are mentioned in the boxes.

Table 4 Classification accuracies for different features on both data sets. The classification rates were averaged over all OSATS criteria and over all values of K (different number of dimensions of time series used for the evaluation) for each technique.

	HMM	BoW	ABoW	MT	SMT	DCT	DFT
Dataset-A Suturing	47.4	63.3	63.1	64.3	84.4	**98.4**	97.7
Dataset-A Knot Tying	44.8	71.2	70.5	67.3	86.9	**97.4**	95.8
Dataset-B	-	-	-	-	78.1	**98.1**	97.6

6 Discussion

The results described above clearly show an increasing trend in classification accuracies going from using symbolic features to frequency features. Symbolic features like BoW and ABoW are useful in classifying human activities in general. Sufficient literature has shown their efficacy in predicting what is being done in the video. For example, RMIS works on gesture recognition [16], and [15] reported good results for surgical gesture recognition using BoW model. However, in their work, the goal was to classify what (or which) gesture is the test sample. Whereas, in skill assessment, it is essential to assess the motion quality i.e. how competent the subject is in performing the given activity. Therefore, symbolic features performed poorly on evaluating skill for both the data sets described in this paper.

A better representation for skill assessment was to encode motion dynamics of the surgeons using texture features. However, its important to note that texture features without temporal information performed poorly (this is also noted by [12]). SMT performed quite well for skill classification for both data sets and is able to capture the sequential information important for skill differentiation. However, SMT is quite computationally expensive due to the calculation of frame kernel matrices and the corresponding textural features. Moreover, SMT also seem to be prone to noisy movements in the video as their is a significant decrease in the average classification accuracy for dataset-B (which had significant movements of people other than the performing surgeon). That noted, SMT does give reasonably high accuracy for skill classification.

The best features to encode the skill level of the surgeons performing basic surgical tasks were frequency based i.e. DCT and DFT. The data sets used in this paper for evaluations only had basic surgical tasks of suturing and knot tying. Both of these activities contain sequential periodic motion of the hands and arms of the surgeon. Keeping this in mind, one could expect that frequency based features might be able to extract the relevant information for skill classification from the time series data. And the results presented in this paper do in-fact conform with this. Moreover, these frequency based skill classification does not require the time series to be divided into different windows nor does it require any manually defined surgical gestures. Also, DCT and DFT both are extremely robust to noisy movements in the videos as evident from the average classification rates given for both data sets in Table 4. This is mainly because low pass filtering of the time series removes such noise in the data, thus making them more robust as compared to SMT. Another thing to note here is that from Table 4, we see that DCT perform slightly better than DFT on average. This can possible be because of not using DFT coefficients as is (since they are complex). We used DCT coefficients in its original form while taking the absolute for DFT. This results in loss of some information which can cause a slightly lower average classification accuracy for DFT.

In order to better understand the difference in the top performing features quantitatively, we need to visualize the feature in their spaces. However, since the dimension of the final feature vector is always much greater than 3, it is very hard to visualize them as is. Therefore, we used linear discriminant analysis (LDA) to project the higher dimensional features onto a 2-dimensional space. LDA was used for dimensionality reduction here since it tries to model the difference between the classes and that would potentially result in distinct class clusters in projected space if the data in higher dimension also forms separated clusters. Figure 11 shows sample scatter plots for SMT, DCT and DFT (from left most column to right most, respectively) features after projecting them using LDA. It is interesting to see that even after significant information loss caused by dimensionality reduction, DCT and DFT form pretty distinct clusters for each skill class whereas there is significant overlap between skill classes clusters for SMT. This shows that the selected frequency features for each class in a higher dimension would be sufficiently distinct, hence achieving classification accuracies upto 100%.

Our experiments in this paper showcase a promising method that uses videos for skill assessment for traditional surgical tasks of suturing and knot tying. We believe that the proposed technique can be used for motion quality assessment in other types of data that have repetitive motion patterns. For example, in RMIS, the same pipeline of video processing could be used for skill assessment involving tasks like suturing and knot tying. Furthermore, the proposed features for time series analysis could be used for skill assessment using kinematic data in RMIS. However, in surgical tasks such as cutting and dissection that do not involve repetitive motions, frequency based features would probably be unable to model the skill level of the surgeons.

7 Conclusion

In this paper, we presented a system for automated assessment of basic surgical skill using video data. Videos of surgical residents and nurse practitioners were classified into different OSATS skill groups. We implemented and compared three different feature types for skill assessment: Symbolic, Texture and Frequency. These feature types were evaluated on two diverse data sets. The results presented in this paper clearly show that frequency features (DCT and DFT) outperform the both symbolic and texture features used on both data sets with average classification accuracy reaching as high 98.7%.

Conflict of Interest: The authors declare that they have no conflict of interest.
Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Informed consent: Informed consent was obtained from all individual participants included in the study.

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. Awad, S., Liscum, K., Aoki, N., Awad, S., Berger, D.: Does the subjective evaluation of medical student surgical knowledge correlate with written and oral exam performance? Journal of Surgical Research **104**(1) (2002) 36–39
3. 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
4. Reznick, R., MacRae, H.: Teaching surgical skills–changes in the wind. The New England journal of medicine **355**(25) (2006) 2664
5. Yu, T., Wheeler, B., Hill, A.: Clinical supervisor evaluations during general surgery clerkships. Medical Teacher **33**(9) (2011) 479–484
6. Datta, V., Bann, S., Mandalia, M., Darzi, A.: The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. The American journal of surgery **192**(3) (2006) 372–378
7. Moorthy, K., Munz, Y., Sarker, S.K., Darzi, A.: Objective assessment of technical skills in surgery. BMJ: British Medical Journal **327**(7422) (2003) 1032
8. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. arXiv preprint arXiv:1602.03012 (2016)
9. Lea, C., Hager, G.D., Vidal, R.: An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE (2015) 1123–1129
10. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer (2015) 430–438
11. 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)
12. 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)
13. Tao, L., Zappella, L., Hager, G.D., Vidal, R.: Surgical gesture segmentation and recognition. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer (2013) 339–346
14. 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)
15. Haro, B.B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: MICCAI 2012. Springer (2012) 34–41
16. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical Image Analysis **17**(7) (2013) 732–745
17. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis **16**(3) (2012) 632–641
18. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: An application-dependent framework for the recognition of high-level surgical tasks in the or. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer (2011) 331–338
19. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer (2010) 400–407
20. Lin, H., Hager, G.: User-independent models of manipulation using video contextual cues. In: International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-Workshop. (2009)
21. Judkins, T.N., Oleynikov, D., Stergiou, N.: Objective evaluation of expert and novice performance during robotic surgical training tasks. Surgical endoscopy **23**(3) (2009) 590–597
22. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: ECCV. Springer (2014) 556–571
23. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision **103**(1) (2013) 60–79
24. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3337–3344
25. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Computer Vision–ECCV 2010. Springer (2010) 392–405
26. Laptev, I., Lindeberg, T.: Space-time interest points. In: IN ICCV. (2003) 432–439
27. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
28. Reiley, C., Lin, H., Yuh, D., Hager, G.: Review of methods for objective surgical skill evaluation. Surgical endoscopy **25**(2) (2011) 356–366
29. Reiley, C.E., Hager, G.D.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: M2CAI workshop. MICCAI, London. (2009)

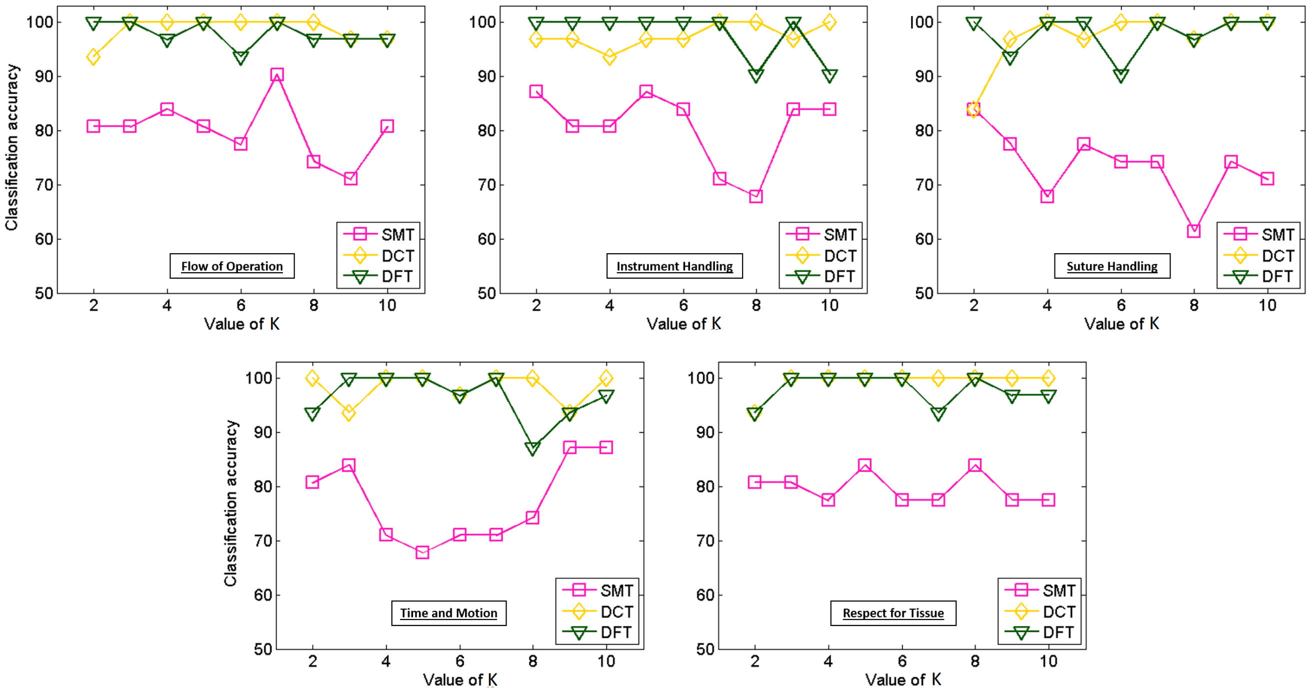

Fig. 10 Plots showing classification rates for various OSATS criteria for dataset-B. The corresponding OSATS criteria for each plot are mentioned in the boxes.

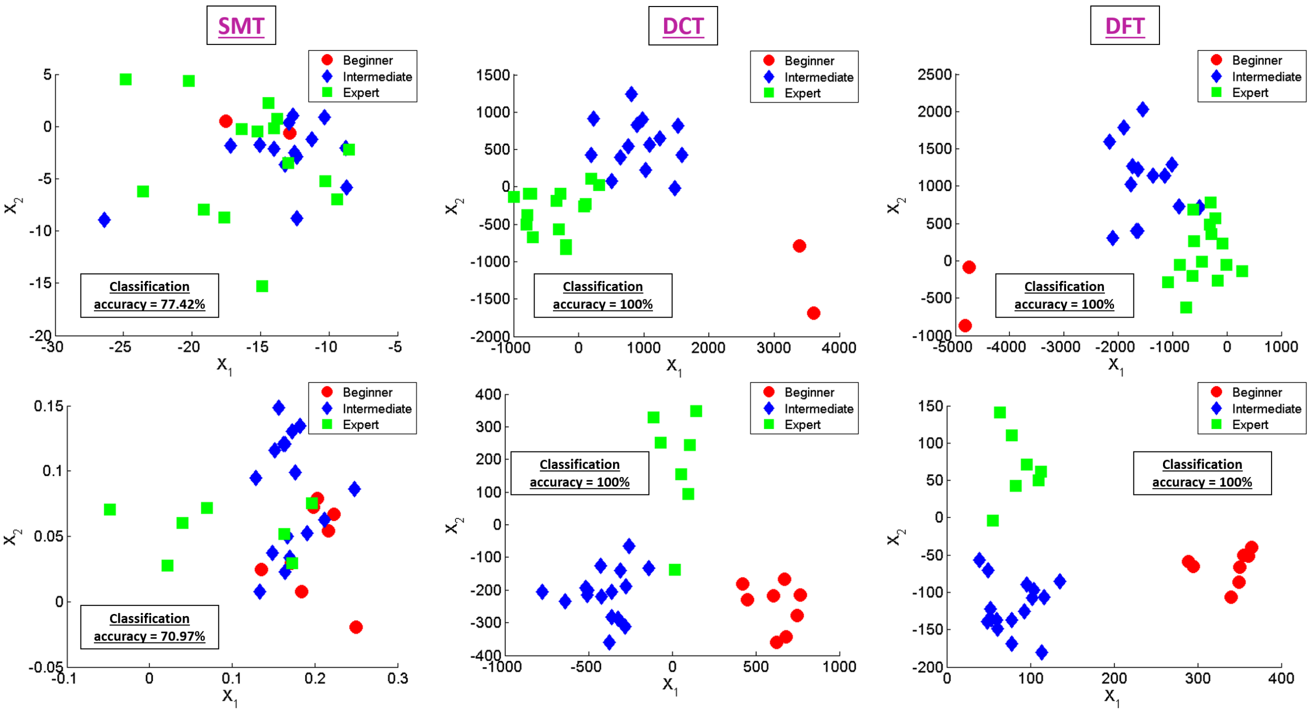

Fig. 11 Sample scatter plots showing the distribution of the 3 skill classes after projecting the selected features onto a two dimensional space using linear discriminant analysis (LDA). Left to right columns show scatter plots for SMT to DFT, respectively. The top row plots were obtained using $k = 4$ for Respect for Tissue OSATS criteria. Whereas, the bottom row plots were obtained using $k = 7$ for Instrument Handling OSATS criteria. All plots shown here were obtained from dataset-B. The classification accuracy achieved in each case using all the selected features is also given in the boxes within each plot.