Noname manuscript No.
(will be inserted by the editor)

Video and Accelerometer-Based Motion Analysis for Automated Surgical Skills Assessment

Aneeq Zia · Yachna Sharma · Vinay Bettadapura · Eric L. Sarin · Irfan Essa

Received: date / Accepted: date

Abstract *Purpose*: Basic surgical skills of suturing and knot tying are an essential part of medical training. Having an automated system for surgical skills assessment could help save experts time and improve training efficiency. There have been some recent attempts at automated surgical skills assessment using either video analysis or acceleration data. In this paper, we present a novel approach for automated assessment of *OSATS-like* surgical skills and provide an analysis of different features on multi-modal data (video and accelerometer data).

Methods: We conduct a large study for basic surgical skill assessment on a dataset that contained video and accelerometer data for suturing and knot-tying tasks. We introduce *"entropy based"* features – *Approximate Entropy (ApEn)* and *Cross-Approximate Entropy (XApEn)*, which quantify the amount of predictability and regularity of fluctuations in time-series data. The proposed features are compared to existing methods of *Sequential Motion Texture (SMT), Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT),* for surgical skills assessment.

Results: We report average performance of different features across all applicable OSATS-like criteria for suturing and knot tying tasks. Our analysis shows that the proposed entropy-based features outperform previous state-of-the-art methods using video data, achieving average classification accuracies of 95.1% and 92.2% for suturing and knot tying, respectively. For accelerometer data, our method performs better for suturing achieving 86.8% average accuracy. We also show that fusion of video and acceleration features can improve overall performance for skill assessment.

Conclusions: Automated surgical skills assessment can be achieved with high accuracy using the proposed entropy features. Such a system can significantly improve the efficiency of surgical training in medical schools and teaching hospitals.

Keywords Surgical skills assessment · Computer vision · Machine learning · Multi-modal data

Aneeq Zia
College of Computing, Georgia Tech
Tel.: +1-404-924-0313
E-mail: aneeqzia@gmail.com

1 Introduction

Surgical trainees are required to acquire specific skills during the course of their residency before performing real surgeries. Surgical training involves constant practice of skills and seeking feedback from supervising surgeons, who generally have a packed schedule. Furthermore, manual assessments, even by experts are subjective and prone to errors. *Objective Structured Assessment of Technical Skills (OSATS)* is adopted in most medical schools as a standard to assess surgical residents [1]. The OSATS grading scheme includes specific criteria like *Respect for Tissue (RT), Time and Motion (TM), Instrument Handling (IH), Flow of Operation (FO), Knowledge of Procedure (KP),* and *Overall Performance (OP)*. While adopting OSATS grading system reduces the subjectivity of assessment to some extent, the grading itself can take up lot of time of the generally few expert surgeons available. The original OSATS grading is done on a scale of 1 to 5 for each criteria except for Overall Performance (OP) which has a Pass/Fail grading. Please note that some recent works, along with ours, have used a modified version of OSATS where OP is also graded on a scale of 1 to 5 along with an additional criteria *"Suture Handling" (SH)* [2,

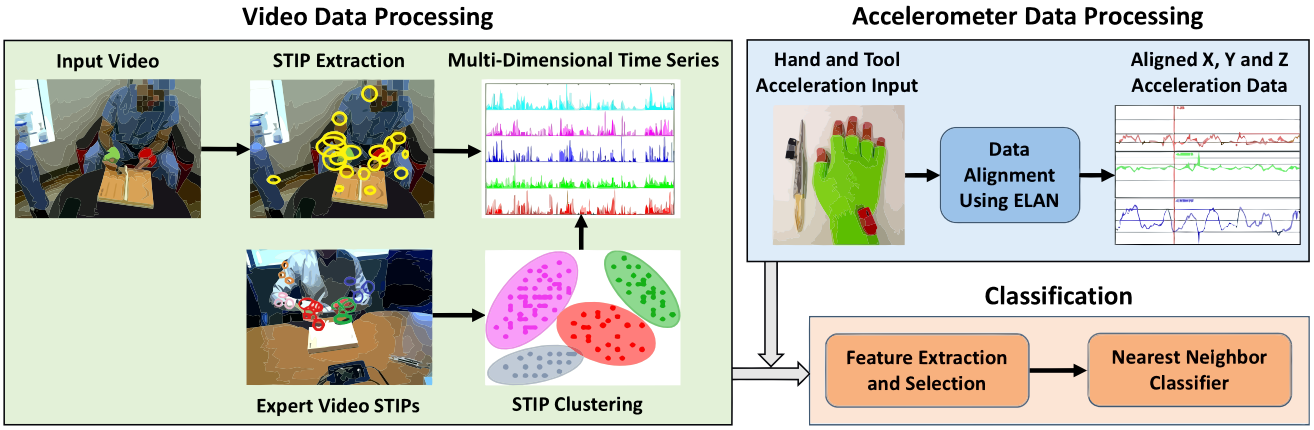

Fig. 1 Flow diagram for processing the video and accelerometer data.

3]. Therefore, we will use the term *"OSATS-like"* to denote this modified grading scheme in this paper where appropriate.

To address the time consuming and subjective nature of manual assessments, recent works have proposed techniques that analyze motion from videos [4,5,2,6] and wearable sensors to assess surgical skills [7–10]. These approaches showcase different feature types to perform OSATS-like assessments. In this paper, we present entropy based features for automated surgical skills assessment. Our main contributions are as follows:

Contributions: (1) We propose a novel way of leveraging the predictability and irregularity of fluctuations in surgical motions to assess OSATS-like skills using entropy features. (2) We provide a comparison of existing techniques on both video and acceleration data. (3) We conduct a large study (containing 41 participants) on assessing basic surgical tasks like suturing and knot tying.

2 Background

The problem of automated surgical skills assessment and gesture recognition has recently seen some good progress. Pioneering efforts were based on robotic minimally invasive surgery (RMIS) and focused on gesture recognition and skill assessment using Hidden Markov Models [11,12]. These initial endeavors attempted to identify gestures or motion sequences for a specific surgical task. These gesture based methods were mostly used for surgical activity recognition and in some cases for surgical skill assessment. Surgical activity recognition works have used methods like linear dynamical systems (LDS) and bag of words (BoW) models [13,14] and more recently, deep neural networks and recurrent neural network based approaches have been reported [15, 16] for surgical tool detection and gesture recognition. Some works have also proposed unsupervised methods for clustering surgical activities [17,18].

Unlike surgical gesture recognition or tool detection, assessment of surgical skills attempts to classify a motion sequence into different expertise levels. Different techniques and data modalities have been used for this purpose. For example, [8] studied robotic surgical movements and reported significant difference in the needle-driving movements of experienced surgeons and novices. GEARS(Global Evaluative Assessment of Robotic Skills) is an assessment tool specifically developed to assess levels of robotic surgical expertise and is known to be consistent and reliable as reported in [19]. Rising interests in assessment of robotic surgical skills have also led to crowd sourcing techniques to derive surgical skill assessment metrics [9].

Despite advances in robotic surgical procedures, assessment of conventional surgical skills such as suturing and knot tying is done using OSATS in medical schools and teaching hospitals. Several works have recently addressed automated assessment based on OSATS-like scores. For example, Augmented BoW (A-BoW) features introduced in [4], modeled motion as short sequences of events and the underlying temporal and structural information is automatically discovered and encoded into BoW models. Other techniques based on the holistic analysis of time series data include Motion Texture(MT) [5] for prediction of surgical skill scores by encoding video motion dynamics into frame kernel matrices followed by texture analysis. Sequential Motion Textures (SMT) was proposed in [2] which included the sequential information into MT technique by dividing the time series into sequential time windows. More recently frequency based features (DFT and DCT) [6,3] have also been used for surgical skill classification. An exhaustive analysis of video based OSATS-like assess-

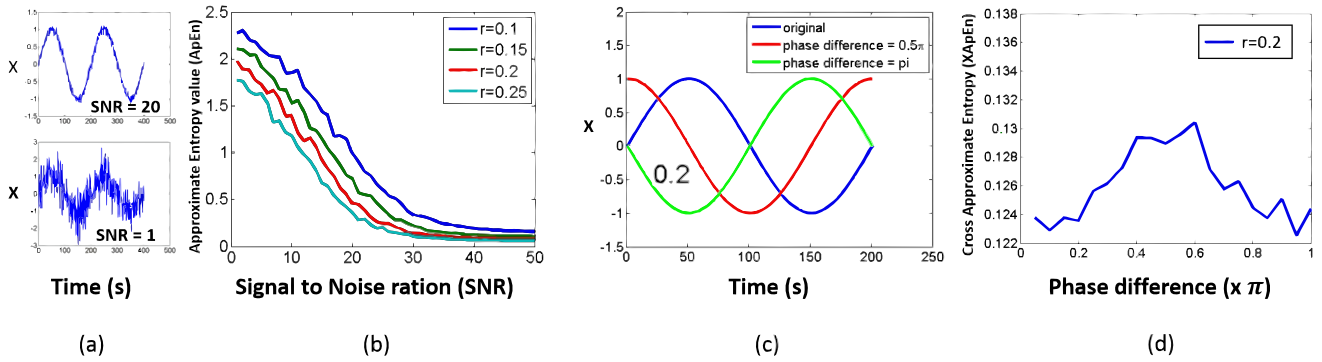

Fig. 2 (a) Sample sine waves with different SNR. (b) Variation of approximate entropy *(ApEn)* with respect to SNR (c) Sample sine waves with different phases (d) Variation of cross approximate entropy *(XApEn)* with respect to phase difference between signals

ments is presented in [3], however, results for only video data are presented.

The techniques mentioned above do provide encouraging results for video based OSATS-like surgical skill assessment. However, these studies use very few participants which limits their ability to capture the wide variation in surgical skills. An expert surgeon's hand motion might be more clean, distinct, ordered and sequential as compared to a non-expert and having more samples helps capture skills of varying levels. Most of the works mentioned above have focused on granularity (MT, SMT) and repetitiveness (DFT, DCT) of motion, however, disorder in motion has not been addressed. Also, they do not include studies on wearable motion sensing devices such as accelerometers that may provide precise motion information for surgical skills assessment.

In the computer vision literature, there has been some recent progress in assessing quality of actions, especially in the sports domain. In [20], the authors presented an approach of using pose with frequency features to predict sports scores. More recently, [21] used entropy features with pose to predict scores for Olympic diving videos. We take inspiration from these works and propose to encode predictability in surgical motions via entropy based features for skills assessment.

In this work, we provide comparative analysis of several features using video and acceleration data on a large group of participants. We also propose entropy based features (encoding orderliness in motion) and demonstrate their efficacy as they outperform other types of features for both acceleration and video data.

3 Methodology

We believe that difference in the motion predictability of surgeons with varying skills levels can be used to assess the basic surgical skills, for specific tasks like suturing and knot tying. An expert will have more predictable hand motion while a beginner will exhibit erratic and irregular patterns. We propose to measure this difference in predictability of motions using entropy based features *Approximate Entropy (ApEn)* and *Cross Approximate Entropy (XApEn)*.

Figure 1 shows the flow diagram for processing video and accelerometer data. For videos, we follow the standard approach, as used by [2,6,3], for encoding motion information from video data into a multi-dimensional time series using Spatio-Temporal Interest Points (STIPS) [22]. As presented in these previous works, we use expert videos to learn motion classes via k-means clustering for different number of clusters K. These motion classes are then used to convert each video into a multi-dimensional time series $T_v \in \Re^{K \times N}$, where K represents the number of motion classes learnt (number of clusters used in k-means clustering) and N is the number of frames in the video. For the three-dimensional accelerometer data, the x, y, and z acceleration time series captured from two accelerometers for each surgical task are concatenated to produce a time series $T_a \in \Re^{6 \times Q}$, where Q is the number of samples captured. We also use individual accelerometer time series data for our analysis as discussed in Section 5. The time series data obtained for both modalities is then used for feature extraction and skill prediction. Sequential forward selection (SFS) [23] is used to reduce the dimensionality of the features and a Nearest-Neighbor (NN) classifier is used for classification.

3.1 Entropy Features for Skill Assessment

Approximate Entropy: Approximate entropy is a measure of regularity in time series data initially proposed in [24]. A more predictable time series would have a

low approximate entropy value whereas an irregular time series would have a higher entropy. For a one-dimensional time series, the approximate entropy *ApEn* is dependent on three parameters: embedding dimension (m), radius (r) and time delay (τ). The embedding dimension (m) represents the length of the series which is being checked for repeatability, the radius (r) is used for local probabilities estimation and time delay (τ) is selected in order to make the components of the embedding vector independent. For a given time series $T \in \Re^N$, we form a sequence of embedding vectors $x(1), x(2), \ldots, x(N-m+1)$, where $x(i)$ is given by $x(i) = [T_i, T_{i+\tau}, \ldots, T_{i+(m-1)\tau}]$, for $1 \leq i \leq N-(m-1)\tau$. Then, for each embedding vector $x(i)$, the frequency of repeatable patterns $C_i^m(r)$ is calculated by

$$C_i^m(r) = \frac{1}{N-(m-1)\tau} \sum_j H(r - dist(x(i), x(j))) \quad (1)$$

where H is a Heaviside step functions and $dist(x(i), x(j)) = \max(|T(i+(k-1)\tau) - T(j+(k-1)\tau)|)$ for $k \in [1, 2, \ldots, m]$. The conditional frequency estimates are calculated by

$$\Omega^m(r) = \frac{1}{N-(m-1)\tau} \sum_{i=1}^{N-(m-1)\tau} ln(C_i^m(r) \quad (2)$$

$\Omega(r)$ is then used to calculate the approximate entropy for the time series $T \in \Re^N$ as $ApEn(m, r, \tau) = \Omega^m(r) - \Omega^{m+1}(r)$.

In order to show how *ApEn* varies for signals with different predictability, we generate a set of sinusoids V. A pure sine wave without any noise can be considered as completely predictable since it has a fixed repeating pattern. However, adding noise to the same function would make it less predictable. We induce white Gaussian noise into our set of sinusoids V to vary the signal-to-noise (SNR) of the set of signals. The range of SNR in the set V was kept from 1 to 50. Figure 2(a) shows some sample sinusoidal waves in the set V with different SNR. Figure 2(b) shows the variation of *ApEn* with varying SNR and radius. As expected, we can see that the higher the SNR (lesser noise), the lower the value of *ApEn* gets for any value of r.

Cross Approximate Entropy: Cross approximate entropy *(XApEn)* is a measure of asynchrony between two time series [25]. For two given time series $[T, S] \in \Re^N$, the embedding vectors are defined as $x_1(i) = [T_i, T_{i+\tau}, \ldots, T_{i+(m-1)\tau}]$ and $x_2(i) = [S_i, S_{i+\tau}, \ldots, S_{i+(m-1)\tau}]$, for $1 \leq i \leq N-(m-1)\tau$.

The frequency of repeatable patterns $C_i^m(r)(T||S)$ for the embedding vectors $x_1(i)$ and $x_2(i)$ is then calculated by

$$C_i^m(r)(T||S) = \frac{1}{N-(m-1)\tau} \sum_j H(r - dist(x_1(i), x_2(j))) \quad (3)$$

$\Omega^m(r)$ is then calculated using

$$\Omega^m(r) = \frac{1}{N-(m-1)\tau} \sum_{i=1}^{N-(m-1)\tau} ln(C_i^m(r)(T||S)) \quad (4)$$

This is then used to finally calculate the cross approximate entropy between the two time series by $XApEn(m, r, \tau) = \Omega^m(r)(T||S) - \Omega^{m+1}(r)(T||S)$.

Similar to *ApEn*, we generate a set of sinusoids W to show the variation of *XApEn* for varying synchrony between different signals. The set W consists of sinusoids with the same SNR but with phase varying from 0 to π. Figure 2(c) shows some sample of sinusoids in this set. Figure 2(d) shows how the value of *XApEn* varies when the phase difference between the signals varies. We can see that the value of *XApEn* reaches a max at about 0.5π and then reduces back to 0 at π phase difference. It is important to note that two sinusoids with a phase difference of π are completely out of phase but in perfect synchrony. This is because if one increases the other decreases with the same rate. This should result in a very low *XApEn* value which we observe in Figure 2(d) as well.

Surgical motions in suturing and knot tying tasks are inherently repetitive in nature. The repetitiveness of motion can be encoded using frequency features. However, frequency features would not be able to capture the sudden movements or jerks in motion that define the competency of a surgeon. They do not quantify the orderliness or predictability of patterns. On the other hand, approximate entropy represents the likelihood of occurrence of similar patterns of observations. A time series containing many repetitive patterns has lower approximate entropy and is more predictable.Therefore, using *ApEn* features can potentially capture repetitiveness along with more finer details crucial for skills assessment. Moreover, in surgical motions, it is also important for surgeons to move their hands and tools in a smooth motion together. We think that *XApEn* features can potentially capture information on how synchronized the surgeon's hands and tools are with each other. We use both the entropy based features described above to encode surgical motion predictability for our analysis.

Table 1 Skill class distribution for each of the OSATS-like criteria (RT: Respect for Tissue, TM: Time and Motion, IH: Instrument Handling, SH: Suture Handling, FO: Flow of Operation, OP: Overall Performance). Each cell contains two values $V : A$, where V = No. of samples for video data, A = No. of samples for acceleration data.

	Suturing					**Knot Tying**			
	RT	**TM**	**IH**	**SH**	**FO**	**TM**	**SH**	**FO**	**OP**
Beginner	38 : 28	46 : 34	47 : 35	47 : 35	45 : 33	27 : 18	27 : 19	22 : 15	23 : 15
Intermediate	22 : 20	15 : 15	13 : 13	17 : 17	18 : 18	22 : 17	28 : 21	28 : 22	28 : 22
Expert	14 : 14	13 : 13	14 : 14	10 : 10	11 : 11	25 : 19	19 : 14	24 : 17	23 : 17

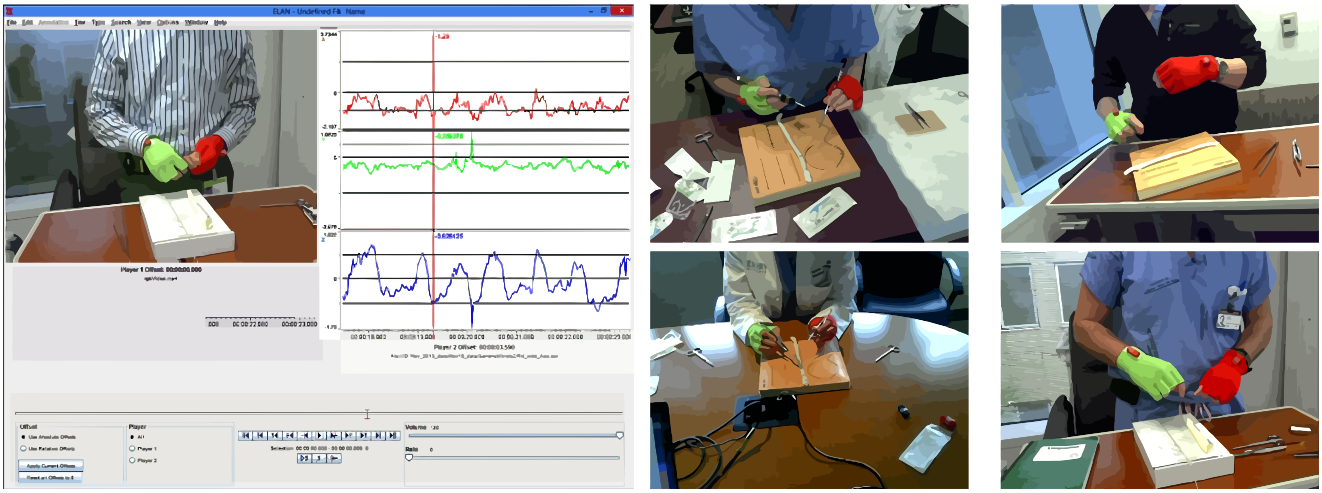

Fig. 3 Image on left shows a screenshot from ELAN software for synchronization of video and accelerometer data. Middle column and right most columns show sample frames for suturing and knot tying, respectively. The accelerometers can also be seen placed on the wrists and the needle-holder

4 Experimental Evaluation

4.1 Data Set

Our data set consists of video and accelerometer data for evaluating the performance of proposed and previous state-of-the-art features for skill assessment. We use the surgical skills dataset from [6] for direct comparisons. This dataset had 18 participants. We augmented this dataset with additional 23 participants to a total of 41 participants consisting of surgical residents and nurse practitioners, essentially doubling the data set from previous studies. In suturing, the participants were asked to perform a "*running suture*" using an instrument (needle holder) for a specified amount of time, resulting in varied number of sutures completed. For knot tying, the participants were asked to tie knots for a given time using their hands only (without any instruments). In this data set, each participant undertook two instances each of suturing and knot tying tasks. For each instance, video data was captured at 30 frames per second at a resolution of 640×480 using a standard RGB camera. We captured a fixed number of frames for each surgical task: 4000 for suturing and 1000 for knot tying. Each video was captured in different lighting conditions and from varying camera angles to make the data set invariant to lighting and viewing angle. Figure 3 shows some sample frames from the videos. Due to acquisition errors, some videos had to be excluded from the data set resulting in 74 videos (from 38 participants) for each surgical task.

The acceleration data was captured using Axivity WAX9[1] sensors. Two accelerometers were used for each surgical task. For knot tying, one accelerometer was attached to each hand wrist whereas for suturing, one accelerometer was attached to the dominant hand wrist and one to the needle-holder. This was done because for suturing, there was very little movement of the non-dominant hand and would not contribute much. On the other hand, needle holder is the main instrument used for suturing. Hence we capture the motion of the dominant hand and the needle holder for suturing. The data captured consisted of x, y and z acceleration values resulting in a 3-dimensional time series for each accelerometer. At the start of each instance, all participants were asked to rapidly shake the hands/instruments with the accelerometers to get the synchronization waveform that is used to align the starting point of acceleration data with the video using the ELAN software

[1] https://axivity.com/downloads/wax9

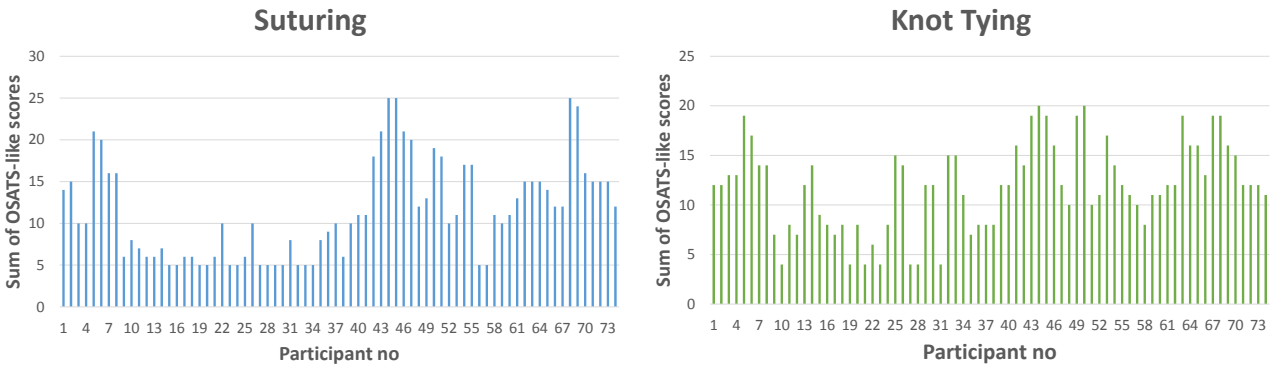

Fig. 4 OSATS-like score distribution for both tasks in the dataset. For this plot, the individual scores for each criteria were summed for each participant.

[26] (a snapshot shown in figure 3). The accelerometer data had some additional noise as the accelerometers were not being attached properly, resulting in unwanted jerks. For some cases, the accelerometer even fell off during a session and had to be reattached. All such samples were removed from the data set resulting in a final 54 acceleration data samples for knot tying (from 30 participants) and 62 for suturing (from 33 participants). The average length with standard deviations of the acceleration data was 8434 ± 2030 for suturing and 1919 ± 507 for knot tying.

In order to generate the ground truth skill levels, we asked an expert to watch the videos and give OSATS-like scores (on a scale of 1 to 5) for each participant. The scores were then divided into three categories: beginner ($score = [1,2]$), intermediate ($score = 3$) and expert ($score = [4,5]$). A complete class distribution for video and accelerometer data is given in Table 1. We also show the distribution of the sum of OSATS-like scores in Figure 4 for both tasks. Please note that we only use the OSATS-like criteria being used in our partner hospital for actual assessment. For example, RT and IH were not used for knot tying since there is no direct tissue contact with no instrument being used. Scores for OP in suturing and KP in both tasks, were not available. This dataset is not currently publicly available for download. However, we do plan to release it in the future.

4.2 Parameter Selection

All the free parameters for the entropy based features (ApEn, XApEn) and previous state-of-the-art features (SMT, DCT, DFT) were tuned on the data set that we extended in this work to find the optimal values. For SMT, DCT and DFT, parameter tuning led to the same parameter values as presented in [2,6,3]. Traditional methods such as HMM, BoW and A-BoW were reported to perform poorly as compared to SMT and DCT/DFT features in [6] and hence were excluded from the experiments.

We used $K \in [2,3,\ldots,10,12,\ldots,20]$ for k-means clustering to learn motion classes (the number of time series dimensions used) for analysis of video data. The accelerometer data, however, did not have this dependency with a 6-dimensional time series (concatenation of 3-dimensional time series from two accelerometers used) for all evaluations. As described in the previous section, entropy based features are dependent on some parameters which need to be specified. These are the embedding dimension (m), time delay (τ) and the radius (r). In order to differentiate time series data on the basis of regularity, radius (r) needs to be equal to $r_{coeff} \times std$, where r_{coeff} can range from 0.1 to 0.25 and std denotes the standard deviation of the time series. For the embedding dimension, $m = 1$ and $m = 2$ both work equally well according to [24]. The time delay τ essentially represents the factor by which the input data is downsampled for further calculations.

For $ApEn$, the approximate entropy for each dimension of the time series is calculated for values of $r_{coeff} = [0.1, 0.13, 0.16, 0.19, 0.22, 0.25]$ resulting in a feature vector $\theta_{ApEn} \in \Re^{R_{ApEn}K}$, where R_{ApEn} represents the number of radius values used for $ApEn$ and K is the dimension of time series used (6 for acceleration data but variable for video data). However, for $XApEn$, we use the same values of r_{coeff} for accelerometer data but only use $r_{coeff} = 0.2$ for videos. This was done since it was observed that the value of $XApEn$ did not vary much for different values of r_{coeff} for videos. Moreover, the computation time for $XApEn$ also increases significantly with increasing dimensionality of time series as is the case for videos. We obtain a final feature vector for cross entropy $\theta_{XApEn} \in \Re^{\frac{R_{XApEn}K(K-1)}{2}}$, where R_{XApEn} denotes the number of radius values

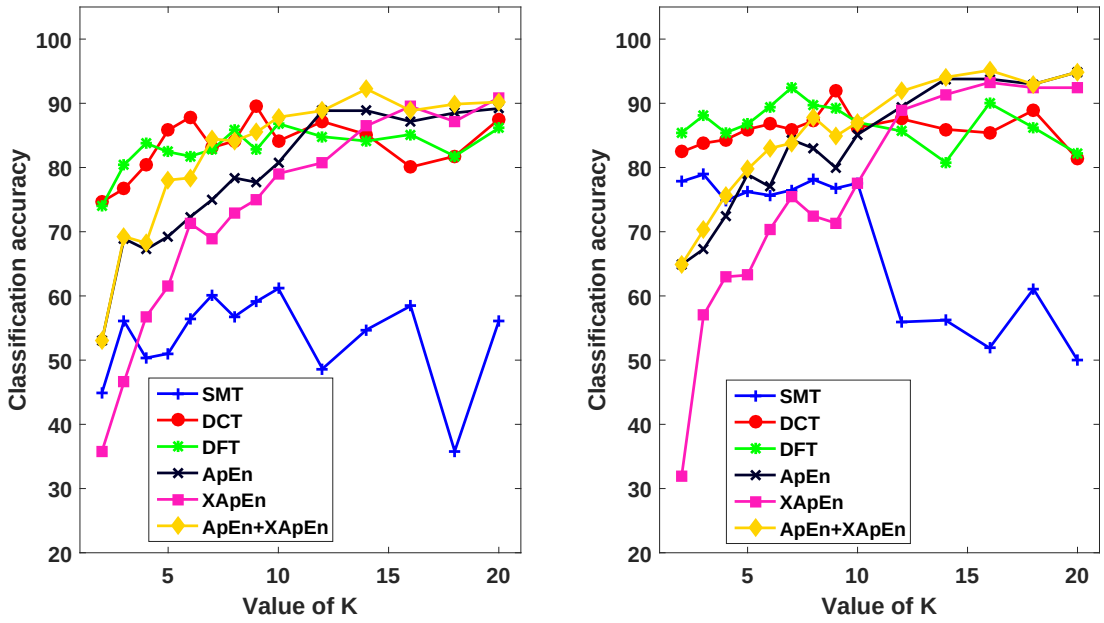

Fig. 5 Average classification accuracy ($\hat{A}_k$) versus K (number of dimensions of time series) for video data. (Best viewed in color)

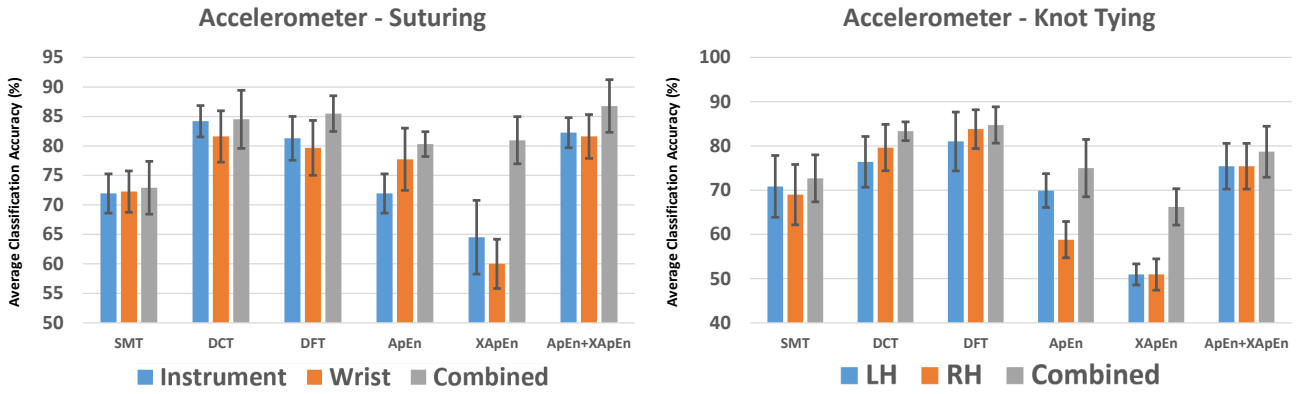

Fig. 6 Average classification accuracies with standard deviations for accelerometer data using individual and combination of the two accelerometers. (Best viewed in color)

Table 2 Highest average classification accuracies with standard deviations for different techniques using multi-modality data. For video data, K corresponding to highest accuracy is also shown.

	Video		Accelerometer	
	Suturing	Knot Tying	Suturing	Knot Tying
SMT	78.9 ± 5.7 (K=3)	61.1 ± 2.3 (K=10)	72.9 ± 4.5	72.7 ± 5.3
DCT	91.9 ± 3.4 (K=9)	89.5 ± 2.8 (K=9)	84.5 ± 4.9	83.3 ± 2.1
DFT	92.4 ± 3.7 (K=7)	86.8 ± 2.8 (K=10)	85.5 ± 3.0	**84.7 ± 4.1**
ApEn	93.7 ± 2.2 (K=20)	89.2 ± 5.3 (K=20)	80.3 ± 2.1	75.0 ± 6.5
XApEn	91.4 ± 3.0 (K=16)	90.9 ± 4.3 (K=20)	81.0 ± 4.0	66.2 ± 4.1
ApEn+XApEn	**95.1 ± 3.1 (K=16)**	**92.2 ± 3.0 (K=14)**	**86.8 ± 4.5**	78.7 ± 5.8

used for *XApEn*. We also check the performance of fusing *ApEn* and *XApEn* before classification by concatenation, resulting in a feature vector $\theta_{ApEn+XApEn} \in \Re^{\frac{K(2R_{ApEn}+R_{XApEn}(K-1))}{2}}$.

Table 3 Per class average precision and recall values over all OSATS-like criteria with standard deviations using video data corresponding to Table 2. The values in each cell are in the format *Precision* | *Recall*.

	Suturing			Knot Tying		
	Beginner	Intermediate	Expert	Beginner	Intermediate	Expert
SMT	89.0±5.7 | 82.3±4.5	66.2±11.8 | 73.2±15.5	60.9±14.7 | 72.6±18.1	68.4±5.6 | 65.5±2.0	51.8±9.4 | 58.4±4.7	63.7±6.9 | 59.4±11.1
DCT	97.3±2.8 | 94.0±1.6	79.2±14.6 | 90.7±7.6	86.0±10.1 | 85.4±10.6	86.5±7.4 | 92.2±7.0	88.3±5.5 | **92.2±5.7**	**93.1±5.9** | 85.1±2.1
DFT	96.5±2.8 | 94.0±2.5	82.1±11.8 | 90.7±8.9	91.0±9.3 | 89.3±5.2	88.1±7.9 | 85.6±4.9	**91.5±7.3** | 85.3±5.4	79.7±9.9 | 90.8±7.7
ApEn	97.6±1.9 | **96.3±2.9**	86.7±8.4 | 90.1±3.0	**94.6±5.0** | 95.3±4.4	91.1±4.4 | 90.0±5.3	86.8±4.5 | 84.8±7.8	89.5±8.1 | **93.8±3.7**
XApEn	97.6±2.4 | 92.8±3.6	80.6±9.0 | **92.6±6.9**	93.8±6.2 | 96.6±4.6	91.6±4.1 | **94.2±3.9**	88.6±3.3 | 87.9±7.6	91.9±9.4 | 91.8±7.5
ApEn+XApEn	**98.1±2.2** | 95.2±3.2	**92.4±7.0** | 92.2±5.2	89.3±8.6 | **100.0±0.0**	**95.0±3.9** | 93.0±6.8	89.7±5.1 | 91.4±3.1	91.6±8.4 | 93.7±6.3

Table 4 Per class average precision and recall values over all OSATS-like criteria with standard deviations using accelerometer data corresponding to Table 2. The values in each cell are in the format *Precision* | *Recall*.

	Suturing			Knot Tying		
	Beginner	Intermediate	Expert	Beginner	Intermediate	Expert
SMT	82.3±4.1 | 79.0±5.0	60.7±7.9 | 69.0±7.8	63.9±11.0 | 61.6±13.4	54.8±8.4 | 75.4±11.5	81.0±7.2 | 66.9±3.2	80.4±3.5 | 80.6±9.6
DCT	**95.8±4.4** | 83.1±5.5	80.2±7.5 | **88.0±5.2**	60.5±7.9 | 84.7±15.0	83.6±9.1 | 79.7±9.7	85.3±7.6 | **84.7±3.3**	80.4±3.5 | 87.3±9.3
DFT	94.2±5.5 | **88.7±3.5**	**82.1±7.5** | 82.8±7.3	67.2±12.7 | 81.9±11.0	**84.9±3.8** | **87.8±6.8**	**88.1±5.7** | 78.3±2.0	80.7±6.9 | **91.4±3.8**
ApEn	91.9±4.2 | 82.5±3.0	64.1±10.0 | 76.1±10.0	69.1±6.0 | 76.4±6.0	74.0±15.1 | 69.0±10.3	67.7±6.6 | 76.5±6.2	82.7±10.7 | 80.9±10.5
XApEn	90.7±5.5 | 82.9±6.7	73.0±17.2 | 78.2±9.5	61.9±15.5 | 82.9±7.4	54.3±7.4 | 70.6±15.4	70.4±5.3 | 63.6±6.8	72.7±10.7 | 67.7±4.2
ApEn+XApEn	93.9±2.1 | 86.2±6.8	75.5±13.6 | 86.4±5.9	**81.7±14.0** | **93.2±7.5**	77.7±8.9 | 75.8±10.7	72.3±10.0 | 81.3±1.8	**86.0±9.8** | 79.3±8.3

The value of m is set as 1 for all evaluations since a higher value did not improve the performance in our case. It was observed that down-sampling the input data (keeping $\tau > 1$) deteriorated the skill assessment performance. Therefore, τ was set as 1. For fair comparisons with previously proposed techniques, we use similar classification methodology and adopt leave-one-out cross validation (LOOCV) using a Nearest Neighbor (NN) classifier with a cosine distance metric. We also use sequential forward selection (SFS) for feature selection. For a feature set $\Psi = \{\psi_j | j = [1, \ldots, P]\}$, SFS finds a subset of features $\hat{\Psi} = \{\hat{\psi}_i | i = [1, \ldots, Q]\}$, where $Q < P$. In our evaluation, we used a Nearest-Neighbor (NN) classifier (i.e. k-NN classifier for $k = 1$) with a cosine distance metric as a wrapper function for SFS. For fair comparison with previous works [2,6,3], the value of Q was set to be 20 (reducing the dimensionality of the final feature vector to be less than or equal to 20). However, we did note that in our analysis, almost always the number of features selected were less than 20. Not using SFS significantly deteriorated performance of all features.

4.3 Evaluation Metrics

Different metrics were used to compare performances of various features on our data set. For video, we calculate the average classification accuracy over all OSATS-like criteria for different features for all values of K in order to find the optimum number of clusters for each feature type. The average accuracy $\hat{A}_k$ is calculated using $\hat{A}_k = \frac{1}{O} \sum_{OSATS} A_K$, where A_K is the accuracy using K clusters for a specific OSATS-like criteria, and O represents the total number of applicable OSATS-like criteria for that task. For accelerometer data, we evaluate the different features for both the accelerometers attached for each task; wrist and needle-holder for suturing and hand wrists for knot tying. Accuracies are averaged over all OSATS-like criteria for accelerometer data as well.

We also calculate the class wise precision and recall values as $precision = \frac{tp}{tp+fp}$ and $recall = \frac{tp}{tp+fn}$, where tp is true positive, fp is false positive and fn denotes the false negatives for the corresponding class. Again, the per-class precision and recall values are averaged over all OSATS-like criteria for a more compact representation.

5 Results

The features and evaluation metrics described in the previous section were evaluated on video and accelerometer data for suturing and knot tying tasks for all applicable OSATS-like criteria. Figure 5 shows the comparison of different features for suturing and knot tying tasks using video data while using different values of K. Figure 6 shows the average classification results achieved using accelerometer data. The highest average accuracy and the corresponding standard deviations achieved for different techniques are given in Table 2. Along with highest average accuracies, we also show the results for individual OSATS-like criteria using optimal K for each feature type (as indicated in Table 2) in Figure 7. The per-class precision and recall values corresponding to accuracies given in Table 2 are given in Tables 3 and 4.

In order to check the statistical significance of the presented results in Table 2, we conducted McNemar's test [27]. The best performing feature for each modality and surgical task was compared with the rest of the

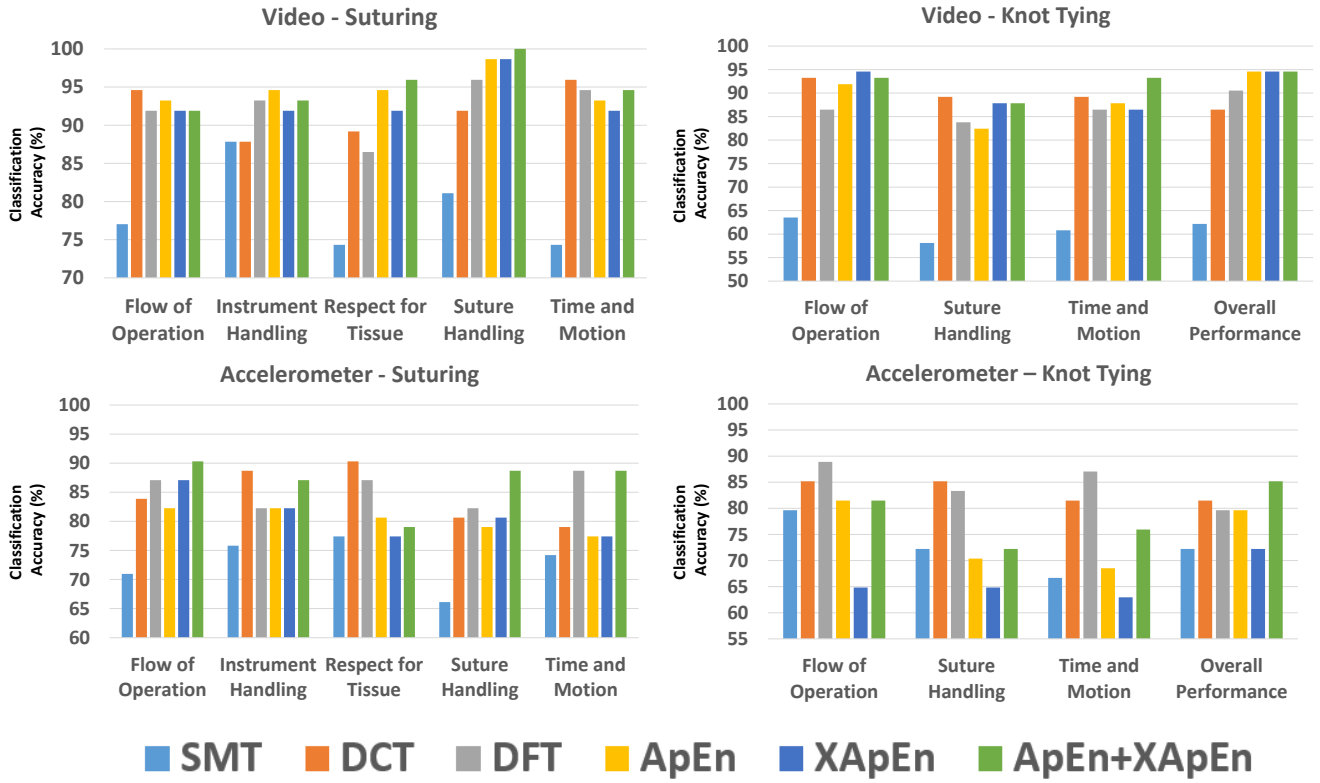

Fig. 7 Individual OSATS-like criteria results for video and accelerometer data. For each feature, the optimal value of K (as indicated in Table 2) was used. (Best viewed in color)

Table 5 McNemar's test of statistical significance for results presented in Table 2. For each column, the highest performing feature (denoted by *"HPF"*) was compared with all other features to check if the higher accuracy achieved is statistically significant by evaluating the p-value. For example, in the first column, ApEn+XApEn performance was compared to rest. The improvement in accuracy is statistically significant if p-value<0.05.

	Video		Accelerometer	
	Suturing	Knot Tying	Suturing	Knot Tying
SMT	<0.01	<0.01	<0.01	<0.01
DCT	<0.01	<0.01	<0.05	<0.05
DFT	<0.01	<0.01	<0.05	HPF
ApEn	>0.05	<0.01	<0.01	<0.01
XApEn	<0.01	<0.05	<0.05	<0.01
ApEn+XApEn	HPF	HPF	HPF	<0.01

features. For comparing performance of different classifiers, a p-value < 0.05 indicates that the difference in classification accuracies is statistically significant. Table 5 shows the p-values achieved conducting the McNemar's test. It can be observed that the improvement in average classification accuracy by the highest performing feature for each column is statistically significant for almost all cases. This shows that the improvements achieved by the proposed entropy based features, when using video data for both tasks and using accelerometer data for suturing, is statistically significant.

We also perform experiments to compare how an early fusion of video and accelerometer data performs for frequency (DCT and DFT) and top performing entropy features (*ApEn+XApEn*). The features are fused via concatenation. Since some of the accelerometer data had to be excluded (as described in Section 4), we only use videos for which the corresponding accelerometer data is available i.e 54 for knot tying and 62 for suturing. Tables 6 and 7 show the average accuracies (over all OSATS-like criteria) with standard deviations using different modalities for suturing and knot tying, respectively.

Lastly, for a more thorough comparison, we perform another experiment using harder cross validation schemes. We again compare *ApEn+XApEn* with DCT and DFT. For this analysis, we use the Video+ Acceleration data for each feature type. Figure 8 shows the average accuracies with standard deviation over all OSATS-like criteria for 2, 5, and 10 fold cross validation schemes. Tables 8 and 9 show results for *'hold-out'* cross validation schemes for suturing and knot tying, respectively. For hold-out validation scheme, $h\%$ of the data was kept as testing data (corresponding to each column in the tables) while the remaining $(100 - h)\%$ was used for training. Within the training data, 10% was used as validation set. Both validation and testing

accuracies are given in Tables 8 and 9. We do not show training accuracy since that will always be 100% using a nearest-neighbor classifier (each point in the training data will be closest to itself, always).

6 Discussion

From the results presented in the previous section, we can see that entropy based features perform better for video data as compared to state-of-the-art techniques in terms of accuracy. For accelerometer data, entropy based features attain a higher accuracy for suturing but not for knot tying (Table 2). The reasons for this is mainly because entropy based features are dependent on the dimension of the time series used (can also be seen in Figure 5 for increasing values of K); the higher the dimension of time series being evaluated, the more information is captured especially for cross entropy ($XApEn$). In case of accelerometer data, we only have 3-axis acceleration values so entropy based features cannot capture enough information. However, entropy based features still have a higher accuracy for suturing task. From Tables 3 and 4, we can see that entropy based features perform well overall, however, there isn't a conclusive trend in terms of precision/recall values.

Comparing the performances of using individual or a combination of accelerometers from Figure 6, we can observe that the combination of data from both accelerometers performs better than individual accelerometers. However, these differences in the performance can potentially give us some valuable insights for skill assessment. For example, in suturing, instrument data works slightly better than wrist for most of the feature types. The reason for this could be that there is relatively more movement of the instrument in suturing as compared to the wrist. Therefore, more motion information would be available to differentiate between different skills. This information can help surgeons improve on their skills by focusing on their instrument motion a bit more.

Comparing results for individual modalities shows us that using video data performs much better than accelerometer for all feature types. This can be explained by the fact that accelerometers only capture the hands/needle-holder 3-D acceleration data whereas videos can be used to extract all motions (both hands, instruments etc.). From the results of our video and accelerometer features fusion experiment (Table 6 and Table 7), we can see that combining video and accelerometer data deteriorates performance for DCT and DFT features as compared to video data. For $ApEn$+$XApEn$, the performance improves for knot tying but slightly decreases for suturing. Overall, the highest performance is achieved using $ApEn$+$XApEn$ features for each task (shown in bold). Even while using harder cross validation schemes, the proposed $ApEn$+$XApEn$ features outperform frequency based features for both tasks for most setups (Figure 8, Table 8, Table 9).

While out-performing the previously proposed features for skill assessment, $ApEn$ and $XApEn$ also have some limitations. Firstly, these features are somewhat dependent on the dimensionality of the time series data; they work better for high dimensional data, especially for $XApEn$ (since it can capture more information). However, increasing dimensionality also leads to potential over-fitting. Moreover, $XApEn$ is computationally expensive and can take a long time if extracted using CPU. However, this can be overcome if a GPU implementation is used. In [28], the authors showed that using GPU for extracting $XApEn$ from a multi-dimensional time series can be more than 250x faster than using CPU. This would be particularly important for real time feedback.

Although, previously proposed frequency features perform reasonably well (especially for accelerometer data), we think that they perform well on repetitive surgical tasks like suturing and knot tying. We believe that the proposed entropy based features would perform better in other surgical procedures as well since they try to capture the irregularity in motion instead of just the repetitiveness. Specifically, it would be interesting to see how these features perform in the recently published JIGSAWS dataset [29] since it contains similar surgical tasks being performed on a da Vinci robot.

7 Conclusion

We presented a comparison of the proposed entropy based features for assessment of surgical skills using video and accelerometer data with previous state-of-the-art techniques. Overall, our analysis showed that videos are better for extracting skill relevant information as compared to accelerometer. However, a fusion of video and accelerometer features can improve the performance. Also, the proposed combination of $ApEn$ and $XApEn$ outperforms state-of-the-art features.

Having an automated system for surgical skills assessment can significantly improve the quality of surgical training. It would allow the surgical trainees to practice their basic skills a lot more with valuable feedback. Moreover, such a system could also help save expert surgeon's time that is spent on trainee assessment.

Table 6 Average accuracies with standard deviations for corresponding feature types using different data modalities for suturing task. Highest performance across all modalities and feature types is shown in bold

	Video	Accelerometer	Video+Accelerometer
DCT	90.6 ± 3.1	84.5 ± 4.9	86.8 ± 7.7
DFT	87.1 ± 1.1	85.5 ± 3.0	86.1 ± 2.1
ApEn+XApEn	**93.9 ± 3.7**	86.8 ± 4.5	93.2 ± 6.6

Table 7 Average accuracies with standard deviations for corresponding feature types using different data modalities for knot tying task. Highest performance across all modalities and feature types is shown in bold

	Video	Accelerometer	Video+Accelerometer
DCT	91.7 ± 6.1	83.3 ± 2.1	83.8 ± 4.9
DFT	86.1 ± 1.9	84.7 ± 4.1	81.0 ± 5.5
ApEn+XApEn	90.3 ± 3.1	78.7 ± 5.8	**94.0 ± 2.8**

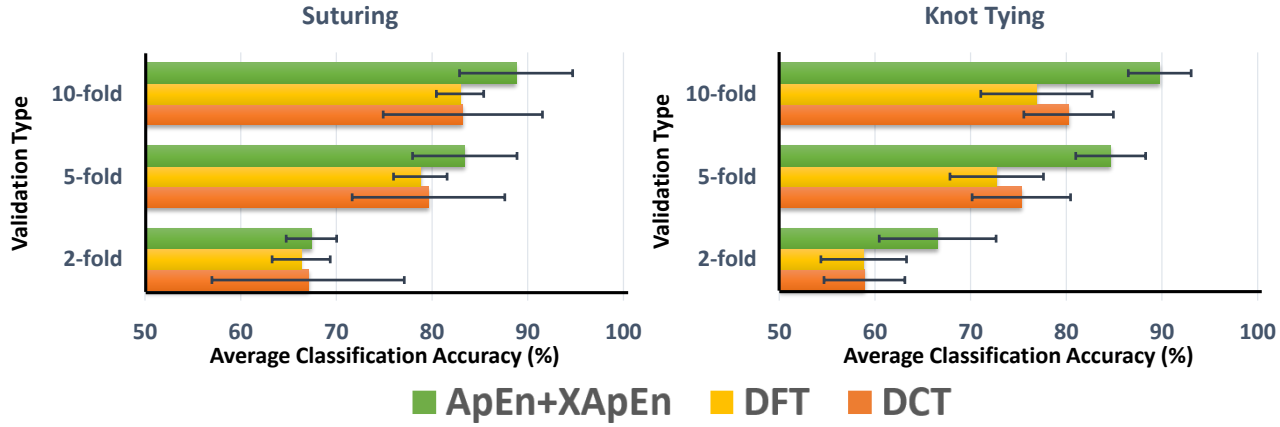

Fig. 8 Average classification accuracy bars with standard deviations for different cross validation schemes by using Video+Accelerometer data. (Best viewed in color)

Table 8 Average validation and testing accuracies over all OSATS-like criteria with standard deviations using hold-out cross-validation for suturing with Video+Accelerometer data. The values in each cell are in the format *Validation Accuracy* | *Testing Accuracy*. Each column corresponds to the amount of data that was *left-out* for testing.

	Testing Set Percentage											
	80%		70%		60%		50%		40%		30%		20%	
DCT	50.3±8.7	51.8±7.6	55.8±8.4	57.7±9.0	**60.5±8.9**	61.9±9.2	64.3±9.6	66.3±9.3	69.1±9.5	71.8±8.7	72.5±8.7	75.4±8.9	76.3±8.8	79.3±8.3
DFT	**53.6±1.8**	**54.0±1.3**	**57.8±2.3**	**58.7±1.7**	**60.5±1.5**	**63.0±1.9**	65.0±1.9	67.3±2.2	69.3±2.2	71.6±2.4	73.1±1.9	75.3±2.4	76.2±2.4	79.0±2.4
ApEn+XApEn	51.6±2.8	51.5±2.3	56.0±3.0	56.9±3.2	59.9±3.5	62.7±4.0	**65.5±3.8**	**67.8±4.3**	**71.3±5.0**	**73.7±4.7**	**75.3±5.0**	**78.4±5.3**	**79.8±5.3**	**83.7±5.7**

Table 9 Average validation and testing accuracies over all OSATS-like criteria with standard deviations using hold-out cross-validation for knot tying with Video+Accelerometer data. The values in each cell are in the format *Validation Accuracy* | *Testing Accuracy*. Each column corresponds to the amount of data that was *left-out* for testing.

	Testing Set Percentage											
	80%		70%		60%		50%		40%		30%		20%	
DCT	42.4±3.7	45.0±3.6	48.5±5.0	50.6±3.9	53.9±5.1	54.9±4.4	57.9±4.0	60.4±4.5	63.2±4.5	65.0±4.1	67.6±4.4	70.2±4.6	71.8±4.8	75.9±5.0
DFT	45.7±3.2	45.4±4.5	50.9±5.4	50.4±5.0	52.7±4.7	54.9±4.9	57.8±4.9	58.7±5.3	6.3±5.0	63.9±5.3	65.6±5.7	68.1±5.1	70.2±5.5	73.3±6.1
ApEn+XApEn	**46.9±5.8**	**47.0±6.3**	**54.6±5.8**	**54.5±6.7**	**58.5±5.7**	**60.9±6.2**	**64.2±5.5**	**66.6±5.6**	**70.2±4.8**	**73.7±5.2**	**75.4±4.9**	**79.0±4.7**	**80.8±4.2**	**85.4±4.2**

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. 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

2. 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) -International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI. (2014)
3. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Ploetz, T., Clements, M.A., Essa, I.: Automated video-based assessment of surgical skills for training and evaluation in medical schools. International Journal of Computer Assisted Radiology and Surgery **11**(9) (2016) 1623–1636
4. 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)
5. 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)
6. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer (2015) 430–438
7. Trejos, A., Patel, R., Naish, M., Schlachta, C.: Design of a sensorized instrument for skills assessment and training in minimally invasive surgery. In: Biomedical Robotics and Biomechatronics, 2008. BioRob 2008. 2nd IEEE RAS & EMBS International Conference on, IEEE (2008) 965–970
8. Nisky, I., Che, Y., Quek, Z.F., Weber, M., Hsieh, M.H., Okamura, A.M.: Teleoperated versus open needle driving: Kinematic analysis of experienced surgeons and novice users. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE (2015) 5371–5377
9. Ershad, M., Koesters, Z., Rege, R., Majewicz, A.: Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer International Publishing (2016) 508–515
10. Brown, J., O'Brien, C., Leung, S., Dumon, K., Lee, D., Kuchenbecker, K.: Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Transactions on Biomedical Engineering (2016)
11. Rosen, J., Hannaford, B., Richards, C.G., Sinanan, M.N.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE transactions on Biomedical Engineering **48**(5) (2001) 579–591
12. Reiley, C., Hager, G.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI. (2009)
13. Haro, B.B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer (2012) 34–41
14. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical Image Analysis **17**(7) (2013) 732–745
15. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1) (2017) 86–97
16. DiPietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S.S., Lee, G.I., Lee, M.R., Hager, G.D.: Recognizing surgical activities with recurrent neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer International Publishing (2016) 551–558
17. Krishnan, S., Garg, A., Patil, S., Lea, C., Hager, G., Abbeel, P., Goldberg, K.: Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. In: Robotics Research. Springer (2018) 91–110
18. Zia, A., Zhang, C., Xiong, X., Jarc, A.M.: Temporal clustering of surgical activities in robot-assisted surgery. International Journal of Computer Assisted Radiology and Surgery **12**(7) (Jul 2017) 1171–1178
19. Goh, A.C., Goldfarb, D.W., Sander, J.C., Miles, B.J., Dunkin, B.J.: Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of Urology **187**(1) (2012) 247–252
20. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: European Conference on Computer Vision, Springer International Publishing (2014) 556–571
21. Venkataraman, V., Vlachos, I., Turaga, P.: Dynamical regularity for action analysis. In: Proceedings of the British Machine Vision Conference (BMVC). (2015) 67–1
22. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2-3) (2005) 107–123
23. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters **15**(11) (1994) 1119–1125
24. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences **88**(6) (1991) 2297–2301
25. Pincus, S., Singer, B.H.: Randomness and degrees of irregularity. Proceedings of the National Academy of Sciences **93**(5) (1996) 2083–2088
26. Sloetjes, H., Wittenburg, P.: Annotation by category: ELAN and ISO DCR. In: Language Resources and Evaluation Conference - LREC. (2008)
27. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2) (1947) 153–157
28. Martínez-Zarzuela, M., Gómez, C., Pernas, F.J.D., Fernández, A., Hornero, R.: Cross-approximate entropy parallel computation on GPUs for biomedical signal analysis. Application to MEG recordings. Computer Methods and Programs in Biomedicine **112** (2013) 189–199
29. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., Chen, C.C.G., Vidal, R., Khundanpur, S., Hager, G.D.: JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In: International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI) -International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI. Volume 3. (2014)