Automated Surgical Skill Assessment in RMIS Training

Aneeq Zia · Irfan Essa

Abstract *Purpose*: Manual feedback in basic RMIS training can consume a significant amount of time from expert surgeons' schedule and is prone to subjectivity. While VR-based training tasks can generate automated score reports, there is no mechanism of generating automated feedback for surgeons performing basic surgical tasks in RMIS training. In this paper, we explore the usage of different holistic features for automated skill assessment using only robot kinematic data and propose a weighted feature fusion technique for improving score prediction performance. Moreover, we also propose a method for generating *'task highlights'* which can give surgeons a more directed feedback regarding which segments had the most effect on the final skill score.

Methods: We perform our experiments on the publicly available JIGSAWS dataset and evaluate four different types of holistic features from robot kinematic data - Sequential Motion Texture (SMT), Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Approximate Entropy (ApEn). The features are then used for skill classification and exact skill score prediction. Along with using these features individually, we also evaluate the performance using our proposed weighted combination technique. The task highlights are produced using DCT features.

Results: Our results demonstrate that these holistic features outperform all previous HMM based state-of-the-art methods for skill classification on the JIGSAWS dataset. Also, our proposed feature fusion strategy significantly improves performance for skill score predictions achieving up to 0.61 average spearman correlation coefficient. Moreover, we provide an analysis on how the proposed task highlights can relate to different surgical gestures within a task.

A. Zia
E-mail: aneeqzia@gmail.com

College of Computing, Georgia Institute of Technology
Atlanta, GA, USA 30332

Conclusions: Holistic features capturing global information from robot kinematic data can successfully be used for evaluating surgeon skill in basic surgical tasks on the da Vinci robot. Using the framework presented can potentially allow for real time score feedback in RMIS training and help surgical trainees have more focused training.

Keywords Robot-assisted surgery · Surgical skill assessment · Feature fusion

1 Introduction

With the rapidly increasing amount of Robot-Assisted Minimally Invasive Surgery (RMIS) around the world, the focus on robotic surgical training has increased tremendously. Typical robotic surgery training includes simulator based and dry lab exercises like suturing, knot tying and needle passing. Training on these tasks is crucial since it forms the base for advanced training procedures on pigs, cadavers and eventually, humans. However, the current assessment on such dry lab exercises is done manually by supervising surgeons which makes it prone to subjectivity and reduces the overall efficiency of training. In order to reduce subjectivity, many medical schools are starting to adopt Objective Structured Assessment of Technical Skills (OSATS) as a grading system [1]. OSATS consists of different grading criteria like Respect for Tissue (RT), Time and Motion (TM), Flow of Operation (FO), Overall Performance (OP) and Quality of Final Product (QP). However, this grading is still done manually making it extremely time consuming.

Much of the literature in basic RMIS training has focused on developing methods for recognizing surgical gestures [2,3,4,5]. Although recognizing surgical gestures within a task can be helpful for skill assessment, treating the data from tasks as a whole reduces the complexity of the problem and has been shown to work well enough [6,7,8,9]. Some of the recent approaches for automated surgical skills assessment in RMIS training have tried to use variants of HMM [10] given data from a task. While HMM's can be effective in modeling temporal data, we hypothesize that extracting features capturing global information from time series data can be more indicative of surgeon skill. Moreover, to the best of our knowledge, all the works in the surgical skills assessment domain have only proposed methods for predicting overall scores of a task. However, although that is the important first step towards better feedback, we feel that it is extremely important to give surgeons more directed feedback as to which part of a particular task contributed to their high or low score. In this paper, we present a detailed analysis on skill assessment for basic RMIS training and list our main contributions below.

Contributions: (1) We propose a framework for automated surgical skills assessment in RMIS training, and show that texture, frequency and entropy based features outperform all previous HMM based state-of-the-art techniques on JIGSAWS dataset using kinematic data. (2) We propose a weighted feature fusion technique for skill score prediction. (3) We provide a detailed analysis

on skill assessment on JIGSAWS dataset and show the role played by different features in score predictions. (4) We propose a technique for generating task highlights that can provide surgeons with more directed feedback as to which parts of a task had the most positive/negative impact on the final score prediction.

2 Background

Automated surgical skills assessment has been of interest to researchers for a long time. Recent works have shown promising results in both RMIS and video based basic surgical skills assessment.

In video based approaches, most works employ Spatio-Temporal Interest Points (STIP) [11] to capture motion information and use them to develop models for skill prediction [12,6,7,8,13,9]. [9] proposed Sequential Motion Texture (SMT) that used texture features of frame kernel matrices for skill prediction. In [6,7], the authors used the repeatability in motions via frequency features (DCT and DFT) to classify surgeon skill level. More recently, [8] proposed to encode the predictability in surgical motions using approximate entropy (ApEn) and cross approximate entropy (XApEn) for skill assessment. In the computer vision literature, frequency and entropy based features have been shown to perform good for sports quality assessment as well [14,15]

For assessment of surgical skills in RMIS, one of the earlier works proposed a variant of HMM - sparse HMM [10]. Other works like [16] studied the differences in needle-driving movements and reported significant differences between beginner and expert surgeons. In [17], the authors proposed descriptive curve coding-common string model (DCC-CSM) for simultaneous surgical gesture recognition and skill assessment. [18] used SVM on basic metrics like time for completion, path length, speed etc, for skill evaluation. More recently, some works have explored the use of crowd sourcing techniques to evaluate surgeon skill [19].

Although the previous works have shown promising results on RMIS based skill prediction, none of them explored the usage of features capturing the repeatability and predictability in surgical motions (like frequency [6,7] and entropy based [8]) from robot kinematic data. We hypothesize that such features would be able to capture more skill relevant information since expert robotic surgeons tend to have smoother and predictable motions as compared to beginners. Moreover, inspired by the work in [14], we hypothesize that frequency based features can be used to evaluate the impact any short segment has on the final score prediction for surgical tasks using inverse transforms.

Fig. 1 Flow diagram of the proposed framework for robotic surgical skills assessment.

3 Methodology

3.1 Skill Classification/Score Prediction

As opposed to previous proposed works on using different variants of HMMs for skill assessment, we evaluate holistic features for predicting skill level using robot kinematics data. Figure 1 shows the proposed pipeline. For a given D-dimensional time series $S \in \Re^{D \times L}$, where L is the number of frames, we extract 4 different types of features: Sequential Motion Texture (SMT), Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Approximate Entropy (ApEn). The dimensionality of the features is reduced using Principal Component Analysis (PCA) before classification/prediction. We give details of the feature types, fusion method and the prediction model below.

SMT: Sequential motion texture was implemented as presented in the original paper [9]. The time series is divided into N_w number of windows. A frame kernel matrix is calculated after which Gray Level Co-Occurence Matrices (GLCM) texture features (20 in total) are evaluated resulting in a feature vector $\phi_{SMT} \in \Re^{20N_w}$.

DCT/DFT: Frequency based features have proven to work well for video based assessment of actions like olympic sports [14] and basic surgical tasks [6,7]. We evaluate DCT and DFT coefficients for each dimension of the robot kinematics time series. This results in a matrix of frequency components $F \in \Re^{D \times L}$. The lowest Q components from each dimension are then concatenated together to make the final feature vector $\phi_{DCT/DFT} \in \Re^{DQ}$. Using low frequency features would eliminate any high frequency noise that could have resulted during data capture.

ApEn: Expert surgeons tend to have a more fluent and predictable motion as compared to beginners. Therefore, a measure of predictability in temporal kinematic data can potentially help differentiate between varying skill levels. Approximate entropy is a measure of predictability in a time series data [20] and has been used in recent literature for activity assessment [15,8]. We extract ApEn features from our robot kinematic time series data as presented in [8]. Evaluating ApEn for all dimensions of the time series data results in a feature vector $\phi_{ApEn} \in \Re^{DR}$, where R is the number of radius values used in evaluation per dimension.

Feature Fusion: We propose a weighted feature fusion technique for skill prediction (as shown in Figure 2). The outputs of different prediction models are combined to produce a skill score. We take our training time series data
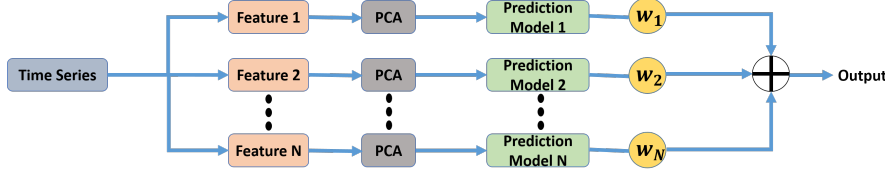
Fig. 2 Weighted feature fusion for modified-OSATS score and GRS prediction.

and evaluate each feature type to produce a training feature matrix $\phi_f \in \Re^{n \times D}$, where f corresponds to a the feature type used, n is the number of training samples and D is the dimensionality of the feature type. The output $y_f \in \Re^n$ corresponding to each ϕ_f is then evaluated using the prediction model. A matrix of outputs from different features $Y \in \Re^{n \times F}$ is generated by concatenating all the y_f, where F corresponds to total number of features used. Given the ground truth predictions $G \in \Re^n$, the optimal weights vector $w^* \in \Re^F$ is then evaluated by solving a simple least squares as $w^* = \underset{w}{\mathrm{argmin}} ||Yw - G||_2^2$. For a given test set, the output $\hat{y_{test}}$ is then calculated using $\hat{y_{test}} = Y_{test} w^*$.

Classification/Prediction: We use a simple nearest neighbor classifier for classification of skill levels. For exact score prediction, we use a linear support vector regression (SVR) model [21].

3.2 Task Highlights

Apart from giving feedback to surgeons in terms of skill score predictions, it could be of great help to surgeons if they knew which parts of the task impacted their final score the most. This can potentially allow surgeons to focus more on specific gestures that contribute to low scores. We define the impact of a segment as the amount by which the predicted score would change if that segment was not observed. In order to do this, we need to evaluate the inferred feature vector had we not observed a particular segment of the data.

Few works have presented approaches for measuring impact of different segments on overall skill score predictions. For example, in [14], the authors presented a frequency features based approach using human pose for evaluating the impact of a particular segment on final score prediction in Olympic sports. Similar to their work, we present a DCT feature based approach for generating *'task highlights'* using robot kinematics data. For a given d-th dimension of the kinematic time series $S(d) \in \Re^L$, the corresponding DCT features $F(d) \in \Re^L$ are evaluated using $F(d) = AS(d)$, where $A \in \Re^{L \times L}$ is the DCT transformation matrix. Taking $B = A^+$ as the inverse cosine transformation matrix (where A^+ denotes the pseudo-inverse of A), the DCT equation can be written as $F(d) = B^+ S(d)$. Now, if the data from frames n_1 till n_2 were to be removed, we can evaluate the inferred DCT feature vector by $\hat{F}(d) = (B_{n_1:n_2})^+ S(d)$, where $B_{n_1:n_2}$) is the matrix B with rows n_1 till n_2 removed. $\hat{F}$ will essentially

Suturing   **Knot-Tying**   **Needle-Passing**

Fig. 3 Sample frames from the 3 tasks in the JIGSAWS dataset [22].

have inferred the missing segment by the most likely kinematics signal given the frequency spectrum of the rest of the signal. Since $\hat{F}$ will have the same dimensionality as F, we can use the same SVR model for score prediction. The final impact of the segment on skill score is then evaluated by $impact = \psi - \hat{\psi}$, where ψ is the predicted score using whole sequence and $\hat{\psi}$ is the inferred score with a missing segment. The surgical task highlights can be generated by evaluating the *impact* on a running window.

4 Experimental Evaluation

Dataset: Our proposed framework is evaluated on the publicly available JIGSAWS dataset [22]. This dataset consists of kinematics and video data from 8 participants for three robotic surgical tasks: Suturing, Knot Tying and Needle Passing. Figure 3 shows sample frames for each task. We only use kinematic data for our analysis and employ the standard LOSO (*leave-one-supertrial-out*) and LOUO (*leave-one-user-out*) cross validation setups. For LOSO, we leave one randomly selected trial from each surgeon out for testing and repeat this 20 times. For LOUO, we leave all trials from one surgeon out for testing. The dataset has ground truth skill labels of three categories: self-proclaimed, modified-OSATS and global gating score (GRS). Self-proclaimed category has three skill levels (dependent on the amount of hours spent on the system) − novice (< 10 hrs), intermediate ($10 - 100$ hrs) and expert (> 100 hrs). The modified-OSATS scores are based on six criteria on a scale of 1-5 and are generated by an expert watching the videos while grading them. This is different from the original OSATS [1] (as described in introduction section) since it contains an extra criteria of suture handling (SH) and that none of the criteria are graded as Pass/Fail. The GRS is a sum of all individual modified-OSATS scores.

Parameter estimation: We use the original feature implementations as presented in [9,6,8]. In SMT, we use number of windows $N_w = 10$ and evaluate Gray Level Co-Occurence Matrices (GLCM) texture features with 8 gray levels resulting in a 200-dimensional feature vector. For frequency features, we take the lowest 50 components ($Q = 50$) for each dimension of the time series

Table 1 Table showing optimal number of PCA components estimated. For prediction, the optimal value of the regularization parameter C is given within parantheses.

	SMT	DCT	DFT	ApEn
Classification	50	150	150	40
Prediction	10 (10^2)	1000 (10^{-6})	250 (10^{-6})	40 (10^4)

and concatenate them resulting in a $50D$-dimensional feature vector, where D is the dimension of time series (76 in our case). In calculating approximate entropy ($ApEn$), we use radius $r = [0.1, 0.13, 0.16, 0.19, 0.22, 0.25]$ resulting in a $6D$-dimensional feature vector. A value of 1 was used for both m and τ.

We use Principal Component Analysis (PCA) for dimensionality reduction before passing features onto the classifier or the regression model. This was done since a lower performance was observed using original feature dimensionality. In order to estimate the optimal number of PCA componenets D_{PCA}, we evaluate performance for D_{PCA} ranging from 10 to 3000 for all tasks for each feature type. The value of D_{PCA} corresponding to highest average performance accross all tasks was selected. For score predictions, we need to estimate an optimal value for the regularization parameter C in SVR. For each feature type, we evaluated the average correlation coefficient (over all modified-OSATS) for $C \in [10^{-7}, 10^{-6}, \ldots, 10^6, 10^7]$ and selected the best performing value of C for evaluations. The optimal values of D_{PCA} and C are given in Table 1. Please note that all parameters were strictly tuned on the training data only for both validation setups. This includes the weights being estimated for the fusion of different prediction models.

For task highlights generation, we use 50 lowest DCT features (same as for classification/prediction) with a running window of length 100.

5 Results and Discussion

We evaluate the proposed features for skill classification and modified-OSATS based score prediction using the JIGSAWS dataset. For classification, we compare the performance of these features with previous HMM based state-of-the-art methods [10]. Table 2 shows results for self proclaimed skill level classification in the JIGSAWS dataset. As evident, using holistic features significantly out-perform previous approaches of using different variants of HMMs. Specifically, ApEn performs significantly better than all other methods. This is interesting to note since experts (with > 100 hrs of practice) would have smoother motions as compared to beginners (with < 10 hrs of practice) making their movements more 'predictable', and hence easily differentiated using ApEn features.

Table 3 shows the results for modified-OSATS and global rating score predictions. We use spearman's correlation coefficient 'ρ' as an evaluation metric and check for statistical significance using the p-value. For modified-OSATS score prediction, we show the value of ρ averaged over all six criteria, whereas,

Table 2 Self proclaimed skill classification results

	Suturing		Knot Tying		Needle Passing	
	LOSO	LOUO	LOSO	LOUO	LOSO	LOUO
MFA-HMM	92.3	38.5	86.1	44.4	76.9	46.2
KSVD-HMM	97.4	59	94.4	58.3	96.2	26.9
SMT	99.70	35.3	99.6	32.3	99.9	57.1
DCT	**100**	64.7	99.7	54.8	99.9	35.7
DFT	**100**	64.7	**99.9**	51.6	99.9	46.4
ApEn	**100**	**88.2**	**99.9**	**77.4**	**100**	**85.7**

Table 3 OSATS scores and GRS prediction results. Each cell contains two numbers in the form $\rho_{OSATS} \mid \rho_{GRS}$, where the first number is the value of ρ averaged over all OSATS and the latter is the value of ρ for GRS prediction. "*" means a $p-$value < 0.05 for the corresponding ρ.

	Suturing				Knot Tying				Needle Passing			
	LOSO		LOUO		LOSO		LOUO		LOSO		LOUO	
SMT	0.25	0.46*	-0.08	-0.28	0.41*	0.39*	0.18	0.21	-0.12	0.09	0.07	-0.60*
DCT	0.57*	0.68*	0.10	0.08	0.59*	**0.76***	0.49	0.73*	0.22	0.26*	-0.16	0.09
DFT	0.45*	0.49*	-0.28	-0.29	0.31	0.32*	0.46*	0.47*	0.44*	**0.53***	0.37	0.19
ApEn	0.31*	0.49*	0.43	0.40*	0.26	0.14*	0.02	0.12	0.16	0.06	0.21	-0.21
SMT+DCT	0.48*	0.61*	0.01	0.01	**0.66***	0.71*	0.46	**0.78***	0.14	-0.16	-0.23	-0.14
SMT+DFT	0.40*	0.60*	-0.21	-0.49*	0.36	0.39*	0.52*	0.48*	0.39*	0.54*	0.33	0.13
SMT+ApEn	0.28*	0.35*	0.41	**0.42***	0.18	0.36*	0.06	0.12	0.12	-0.06	0.15	-0.29
SMT+DCT+DFT	0.57*	0.64*	0.16	0.10	0.58*	0.70*	**0.56***	0.73*	0.36*	0.38*	0.50*	0.23
DCT+DFT	0.56*	0.66*	0.13	0.14	0.53*	0.68*	0.55*	0.73*	0.41*	0.47*	**0.53***	**0.28**
DCT+DFT+ApEn	**0.59***	**0.75***	0.43*	0.37*	0.57*	0.63*	0.48	0.60*	0.37	0.46*	0.23	0.25
SMT+DCT+DFT+ApEn	0.47*	0.66*	**0.45***	0.37*	0.55*	0.61*	0.49	0.62*	**0.45***	0.45*	-0.21	-0.19

Table 4 Values of ρ averaged over all three tasks for the corresponding feature types in the form $\rho_{OSATS} \mid \rho_{GRS}$.

	LOSO	LOUO
SMT	0.18 | 0.31	0.05 | -0.22
DCT	0.46 | 0.57	0.14 | 0.24
DFT	0.40 | 0.45	0.19 | 0.12
ApEn	0.24 | 0.23	0.22 | 0.10
SMT+DCT	0.43 | 0.39	0.08 | 0.22
SMT+DFT	0.38 | 0.51	0.22 | 0.04
SMT+ApEn	0.20 | 0.22	0.21 | 0.08
SMT+DCT+DFT	0.50 | 0.57	**0.41**| 0.36
DCT+DFT	0.50 | 0.60	0.40 | 0.38
DCT+DFT+ApEn	**0.51** | **0.61**	0.38 | **0.41**
SMT+DCT+DFT+ApEn	0.49 | 0.58	0.24 | 0.27

the GRS ρ values are given as is. Feature combination results presented in Table 3 are evaluated using weighted feature fusion as described in methodology section. Overall, we can see that individual features and their combinations achieve good results for the LOSO setup. On the other hand, we see a comparatively low performance on LOUO setup. This is because LOUO is a harder validation scheme due to less data for training phase. However, using the proposed feature combination significantly improves performance over individual features. In general, frequency features seem to perform well when used individually or in combination with other features. We can also see an overall lower performance across all features for the needle-passing task. The reason for this could be that needle-passing is a relatively less repetitive task as compared to
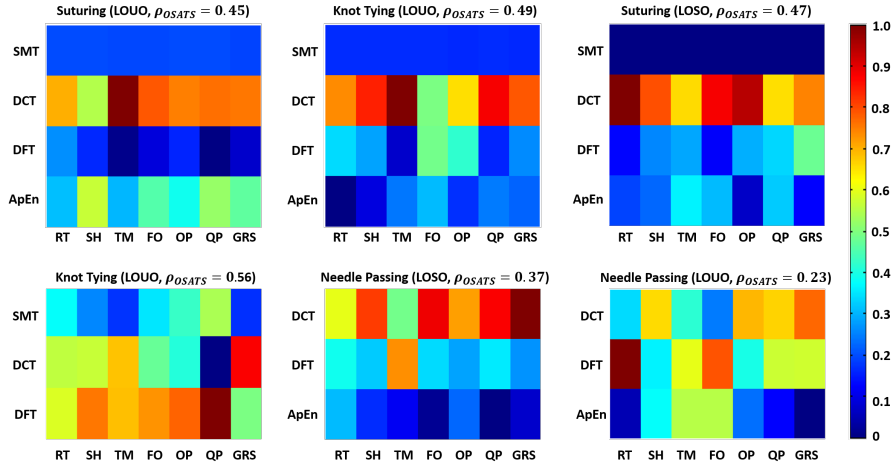
Fig. 4 Heatmaps of weight assignments of different features. Each column shows the weight vector $w*$ (scaled from 0 to 1) for the corresponding OSATS criteria or GRS. For each heatmap, the features used in combination are shown next to each row and the corresponding task, validation scheme and average ρ (over OSATS) are also shown. (Please view this figure in color)

the other two. Since the features we use try to differentiate between different skill levels using data repeatability, they perform less well for needle-passing. Table 4 shows the average of ρ values over all three tasks (as given in Table 3) for each feature type. We observe that DCT+DFT+ApEn performs best on average for OSATS and GRS score prediction.

In order to analyze the role of different features in the proposed weighted late fusion for skill prediction, we generate heatmaps of the weight vectors learned and show a few of them in Figure 4. It can be seen that DCT features get assigned the highest weight in most of the cases. DFT and ApEn features generally have similar weight assignments whereas SMT always gets assigned a low weight. This shows that DCT features capture the most skill relevant information which is also evident from its high performance compared to other individual features in Table 3.

Figure 5 shows some sample task highlights constructed by following the procedure described in the methodology section. We overlay the impact scores plot on color coded gestures for getting better insights. The gestures used are the same as presented in the original dataset paper [22]. For completeness, the gesture vocabulary of JIGSAWS dataset is given in Table 5. The segments where the impact scores are negative indicate that these parts had a adverse effect on the final score, and vice versa. There are some interesting points that we can note from these plots that make intuitive sense. For example, in the suturing plot, we can observe that the impact score has maximum variations for G3 (i.e. Pushing needle through tissue). Since we predict for RT criteria in this case, one would expect that a 'good' or 'bad' push of a needle through the tissue should have the maximum impact on final skill score prediction.
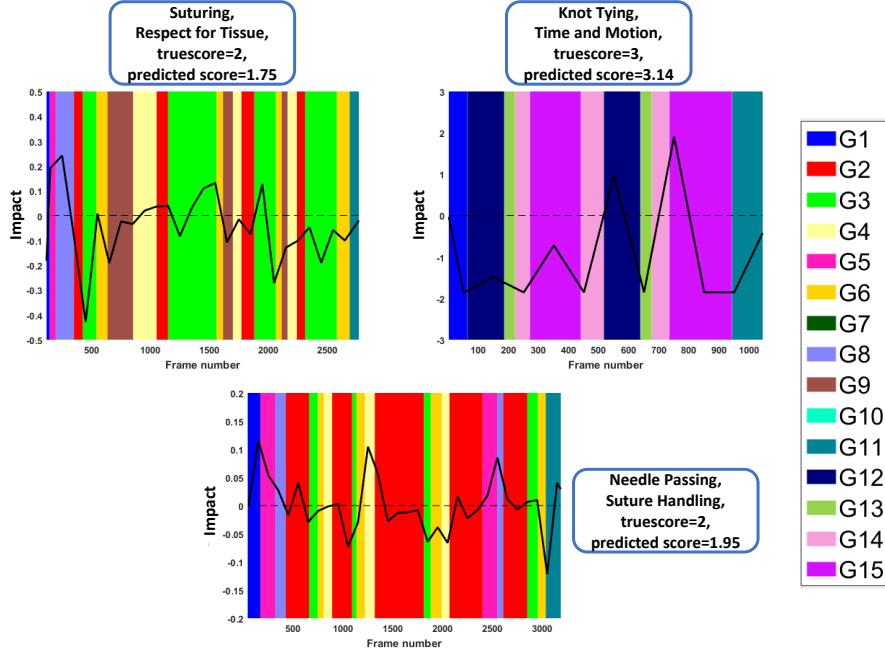
Fig. 5 Sample task highlights. The y-axis on each plot corresponds to the impact (as defined in methodology section) with number of frames on the x-axis. The task type, modified-OSATS criteria, ground truth score, and the predicted score from our model using DCT features on the whole sequence, are given in boxes next to each plot. The color coding for the different gestures is also provided.

Similarly, for knot tying, we can see high positive and negative impact scores for G15 (i.e. Pulling suture with both hands). Again, this makes intuitive sense since G15 is important for knot tying task. We can draw similar insights for needle-passing (considering G2, G4 and G5) as well. Although there are no ground-truth highlights to compare our results to (and it probably would be an extremely tedious task to generate such ground-truths), we believe that such impact score plots can tremendously help surgeons in understanding the parts within a task that they need to improve on. As a result, surgical trainees can direct their time and training on specific gestures within a task which can potentially allow them to move through their learning curves much faster.

6 Conclusion

In this paper we propose to use holistic features like SMT, DCT, DFT and ApEn for skill assessment in RMIS training. Our proposed framework outperforms all existing HMM based approaches. We also present a detailed analysis of skill assessment on the JIGSAWS dataset and propose a weighted feature combination technique that further improves performance on score predictions. We do not use any video data making our method computationally

Table 5 Gesture vocabulary [22].

Gesture ID	Description
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G7	Pulling suture with right hand
G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

feasible for real time feedback. Our framework can easily be integrated in a robotic surgery platform (like the daVinci system) to generate automated modified-OSATS based score reports in training. Moreover, our proposed task highlights generation method could be extremely valuable for giving surgeons more focused feedback.

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. Reiley, C.E., Hager, G.D.: Decomposition of robotic surgical tasks: an analysis of sub-tasks and their correlation to skill. In: M2CAI workshop. MICCAI, London. (2009)
3. Haro, B.B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: MICCAI 2012. Springer (2012) 34–41
4. 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 (2016) 551–558
5. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Bejar, B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE transactions on bio-medical engineering (2017)
6. 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
7. 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
8. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. arXiv preprint arXiv:1702.07772 (2017)
9. 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)

10. Tao, L., Elhamifar, E., Khudanpur, S., Hager, G.D., Vidal, R.: Sparse hidden markov models for surgical gesture classification and skill evaluation. In: International Conference on Information Processing in Computer-Assisted Interventions, Springer Berlin Heidelberg (2012) 167–177

11. Laptev, I.: On space-time interest points. IJCV (2005)

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, Georgia Institute of Technology (2014)

13. 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)

14. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: ECCV. Springer (2014) 556–571

15. Venkataraman, V., Vlachos, I., Turaga, P.K.: Dynamical regularity for action analysis. In: BMVC. (2015) 67–1

16. 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

17. Ahmidi, N., Gao, Y., Béjar, B., Vedula, S.S., Khudanpur, S., Vidal, R., Hager, G.D.: String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer (2013) 26–33

18. Fard, M.J., Ameri, S., Chinnam, R.B., Pandya, A.K., Klein, M.D., Ellis, R.D.: Machine learning approach for skill evaluation in robotic-assisted surgery. arXiv preprint arXiv:1611.05136 (2016)

19. 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

20. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences **88**(6) (1991) 2297–2301

21. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In Jordan, M.I., Petsche, T., eds.: Advances in Neural Information Processing Systems 9. MIT Press (1997) 155–161

22. 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., Khudanpur, S., Hager, G.D.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2CAI. Volume 3. (2014)