Measuring surgical knot tying with 3D vision and VR gloves

J Bhattacharya, P Ridgway, G Lacey

Trinity College Dublin
{jbhattac, ridgwayp, Gerard.Lacey}@tcd.ie

Abstract

One handed surgical knot tying is one of the core skills of surgeons. The development of this skill takes time and one-on-one expert supervision. Our motivation is to provide good quality automatic evaluation in order to reduce the need for expert supervision and provide continuous evaluation that can support a deliberate practice training paradigm. Our work demonstrates the recognition of knot tying gestures a combination of 3D vision tracking and finger motion analysis from a pair of Manus VR gloves. We demonstrate segmentation and classification of the knot tying into its component knot types and estimate student skill level based on motion analysis.

Keywords: Vision, motion capture, gesture recognition, VR glove.

1 Introduction

Surgeons are expected to obtain high quality knot tying expertise through training and practice. Knots are tied one-handed with both the dominant and non-dominant hand. The development of these skills is a time-consuming process and require significant expert supervision. However, expert supervision is not perfect as there is significant inter-assessor variability. The use of Objective Structured Assessment of Technical Skills (OSATS)[1] attempts to minimize this problem by setting standard evaluation critera. The different factors included in the OSATS grading scheme are Respect for Tissue (RT), Time and Motion (TM), Instrument Handling (IH), Suture Handling (SH), Flow of Operation (FO), Knowledge of Procedure (KP), and Overall Performance (OP).

This work aims to develop an automated evaluation for the OSATS skill of surgical knot-tying. We segment the overall knot tying task into its component knot types and evaluate the expertise of the user based on the component knot type, the sequence of ties and flow of the procedure. Other than using common computer vision tools for region-of-interest detection and segmentation, this work exploits the use of pose information and motion analysis from the pose time series data for action recognition. The rest of the paper is organized as follows: Section 2 discusses related work, problem statement and methodology is discussed in sections 3 and 4. Section 5 and 6 provide the experimental results and conclusion.

2 Background

The objective evaluation of surgical proficiency using simulators is a common theme of the medical education literature. [2, 3] use both video and kinematic data for linear dynamical systems (LDS) and bag-of-features (BoF) in order to classify Robotic Minimally Invasive Surgery (RMIS) gestures while [4] use a semi-automated method for skill evaluation. In [5] stereo vision tracked 3D path of Minimally Invasive Surgical (MIS) instruments is used to classify the skill levels of surgeons preforming suturing exercises in a mixed reality MIS simulator. In [6] the motion dynamics data from videos were first encoded in a frame kernel using clustered

 ISBN 978-0-9934207-4-0

HOG-HOF descriptors from spatio-temporal interest points. This data was further used for texture feature extraction using standard GLCM matrices and LBP. OSATS scores were computed from LDA reduced texture features via a linear regression technique. Frame kernel encoded data was also used in [7] but the videos were classified into novice, beginner or expert category using NN classifier on texture descriptors of the frame kernel. In [8] the authors focus on assessment of suturing and knot tying tasks using video and accelerometer data.

Skill classification of endoscopic sinus surgery using hand and eye data from electromagnetic tracker and infrared eye tracker respectively[9] has used Hidden Markov Models. It was also used at a subtask level for identifying skill level of individual surgeons [10] in a robotic surgery task. Further [11] demonstrated the effect of surgeons level and dexterity level segmentation of surgical videos using Hidden Markov Models (HMMS). An extensive comparison of HMM, Bag-of-Words(BoW), Augmented-BoW, Spatial-Motion-Texture(SMT), Discrete-Fourier-Transform(DFT), Discrete-Cosine-Transform(DCT) was provided in [12]. Surgical skills for RMIS was also classified using machine learning tools such as logistic regression and linear SVM [13]by using features like task duration,curvature,depth perception smoothness,path length; computed from the da-Vinci instrument data.

3 Problem statement

Similar to [6], the current work aims at an automatic evaluation a surgical knot tying task. The objective is that students, with the help of the automated evaluation tool, can learn and practice independently with objective feedback. To validate the performance of the system we will perform a construct validity test, in which use the system to separate subjects in differnet skill levels. These should correlate with their prior experience of one handed knot tying. We use both 3D video of the knot-tying task captured using a Intel RealSense RGBD camera as well as a ManusVR glove. The glove provides joint positions for 16 points of the hand; a hand pivot point on the palm and 3 joints per finger.

We decompose the task into sub-parts and analyze each separately. The knot-tying task is divided into 2 major knot types; i.e. a Fore Finger Throw (FFT) and a Back Hand Throw (BHT). The correct surgical knot is made using a FFT followed by a BHT. Another common error is related to rope positioning, a subject tying an FFT with the left hand should move the right part of the rope upwards to snug the knot. If the the left part of the rope is moved upwards, this is treated as an error. The third requirement is to identify the overall flow of the procedure.

4 Methodology

4.1 Segment the video into a set of knot sequences

On the basis of the current task, a simple yet elegant approach is used to automatically segment the videos into individual knots. Each knot ends with an action to snug down the knot, this requires the two hands to move away from each other. This action is used to identify the boundaries between the different component knots. The hands are tracked in each frame and their distances from each other are recorded. The peaks are considered as the frame boundaries. The depth image is used to segment the hand regions based on the fact that they are nearest to the camera. The RGB frame is synchronized based on this ROI mapping. Due to the presence of the rope and the knot-tying kit the depth synchronized RGB image I_D is further used to obtain hand clusters via Gaussian Mixture Model. k-Means clustering algorithm is used to track the cluster across the frames and obtain their inter-cluster distance map D_M. D_M is smoothed using conventional box filters and peaks are marked as the knot boundaries. Two consecutive peaks are checked for a local minima (below a threshold) between them, in the absence of which, the latter is not considered as a qualifying frame boundary. This is depicted in figures 1 and 4.2.

 ISBN 978-0-9934207-4-0

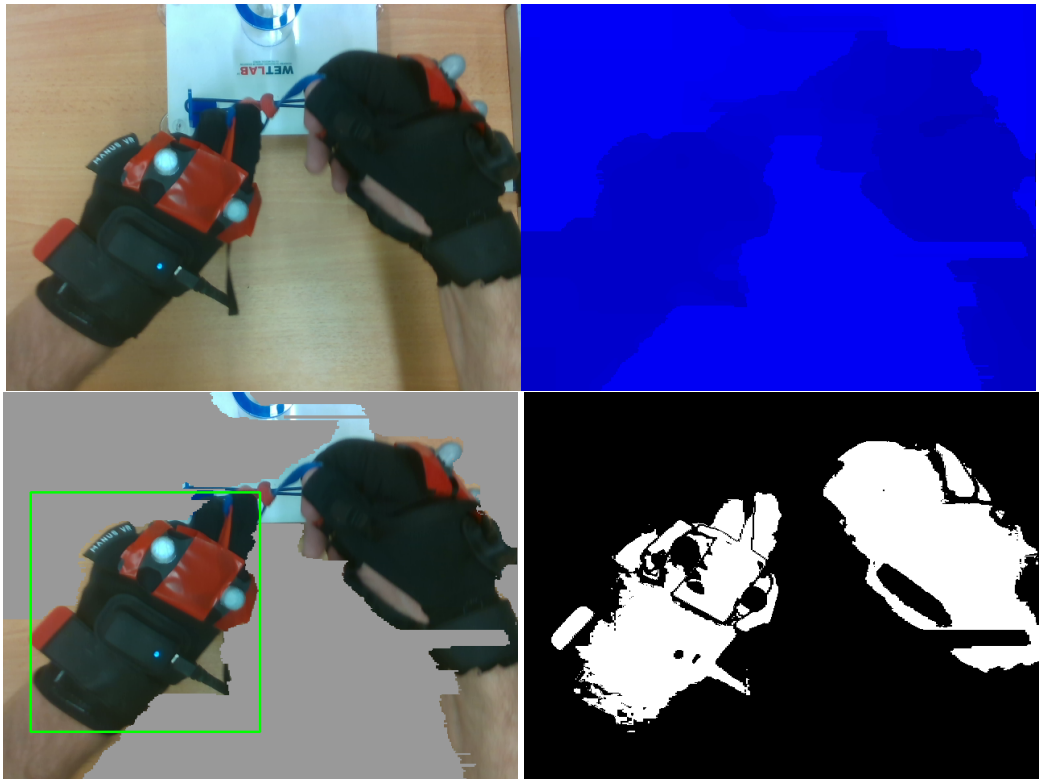

Figure 1: Top image depicts the RGB/Depth Capture from Intel Camera , bottom left depicts I_D with non-dominant detected cluster, bottom right depicts the hand clusters from GMM

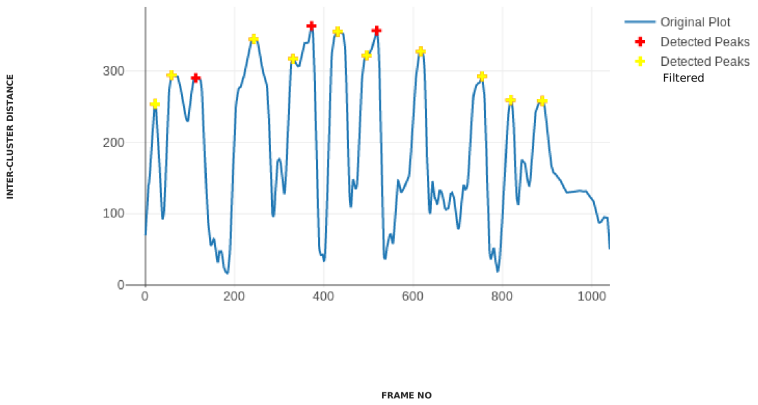

Figure 2: D_M.Red peaks are all peaks detected. Yellow marked peaks are filtered red peaks and denote the Frame boundaries separating the knots

 ISBN 978-0-9934207-4-0

4.2 Represent a knot-tying action as a set of pose information

The ManusVR gloves provide joint position information and the time stamps of the video data and ManusVR data are synchronized so that the knot boundaries can be marked identified within the pose data. The frame rate in both the cases are different, i.e the RealSense uses 15fps while the gloves use 90fps. The knot set $\mathbf{X}$ is a set of n knots $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ...\mathbf{X_n}\}$ where $\mathbf{X_i} \in \Re^{l.90\times16}$ where l is the knot boundary in seconds. In order to get a smooth mapping between the camera and glove data, the peak frame number was converted to the nearest multiple of 15 so that its equivalent second length could be easily obtained. It is verified from the captured videos that doing so does not hamper the knot boundary. This is mainly because the knot boundaries are located at the snugs, the duration of which takes up multiple frames.
It should be noted that the non-dominant hand has the main role in knot tying. The dominant hand is just used for holding the rope and snugging, hence pose information for dominant hand is not important and avoided to reduce complexity. Further, the critical information on pose was described by the fore finger during the FFT and the middle finger during the BHT. The remaining finger movements were not important and can be ignored. It was observed that expert surgeons had almost no movement in these fingers whereas novice learners had more movements in these three fingers.

4.3 Classify knot types

In order to use $\mathbf{X_i}$ for classification, there are a few factors which needs to be considered. As $\mathbf{X_i}$ denotes a set of frames, it can be used with a standard bayesian or LDA classifier where each frame will contribute towards the total score in the distribution. The varying length of the sequence in each case may be dealt with a suitable sampling function for applying any distance classifier. Alternatively, a bag-of-words[14] approach can be used to process the entire frame set without any sampling operations. The well known KLDA set based classifier[15] deemed useful in this scenario. Though applied for image set recognition tasks, this is its first application for a surgical video sequence classification task. In general sets are represented by their covariance matrix $\mathbf{C}$. The Riemmanian distance metrics d_{RD} as shown in equation 1 between two sets is a reduced version of the Euclidean space due to the use of a logarithmic map. The Riemannian kernel function $\kappa(a, b)$ can be derived by the inner product of the logarithm maps.

$$\begin{aligned} d_{RD}(\mathbf{C}_i, \mathbf{C}_j) &= \| \log(\mathbf{C}_i) - \log(\mathbf{C}_j) \|_F \\ \sigma(\mathbf{X_i}), \sigma(\mathbf{X_j}) &= \kappa(\mathbf{C}_i, \mathbf{C}_j) \\ \kappa(\mathbf{C}_i, \mathbf{C}_j) &= tr[\log(\mathbf{C}_i).\log(\mathbf{C}_j)] \end{aligned} \tag{1}$$

The kernel is learned by optimizing equation 2, where $\mathbf{C}_j$ is the covariance of $\mathbf{X_j}$, K is the kernel matrix and α is a diagonal matrix representing the inverse of the number of samples in each class.

$$w = argmax \frac{w^T K \alpha K w}{w^T K K w} \tag{2}$$

Another approach taken apart from processing the pose data of all the frames individually, was using the finger motion along the sequence, obtained using the rate of movement of all the joints. This was measured as

$$ke_{t-1} = \frac{1}{2} \sum_j \frac{\| x_t^j - x_{t-1}^j \|_F}{f} \tag{3}$$

where x_t^j is the position vector of the j^{th} joint at frame t and f is the frame rate. The motion vector per sequence ke can be used with standard gaussian, LDA or distance classifiers.

 ISBN 978-0-9934207-4-0

4.4 Identify Skill level

In order to identify the skill level of the candidate, features for Intermediates and novices are learned separately. The available data from the FFT knot and BHT knot sequences are split into 2 classes, each representing the novice and intermediate subjects. A test video sample is first segmented as mentioned in section 1. The frame boundaries obtained are then used to extract and parse motion data from the ManusVR and these are further transformed using w as $w^T K_t$. The transformed feature is then compared to each of the 4 sets such that it is classified(with a k-nearest neighbor) as belonging to FFT or BHT of a novice or intermediate learner.

5 Experimental results and discussion

120 knot sequences from novice (N) and intermediate learners (I) with an equal number of FFT and BHT are used and the sequences are classified into knot-types and expertise level. K-fold validation results obtained using (a)Kernal-Set based classifier on ManusVR pose data (b) KE-LDA: LDA classifier on motion vector (obtained using equation 3) of the ManusVR pose data (c) KE-Bayes: Bayes classifier on motion vector of pose data obtained from ManusVR (d) KE-Set: Set based classification on motion vector data (e) LDA-vote on a knot sequence (f) Bayesian voting on individual frames on a knot sequence is shown in figure 3.
It is observed that the kernel set based classification and LDA voting is able to classify knot types satisfactorily. It is observed that all the motion vector based techniques achieve an average of 65% accuracy. However on critical analysis, it is seen that they fail in case of novice learners. For example, an accuracy of 75% on k=2 fold validation is obtained for the intermediate learners while 40% knot identifying accuracy is obtained for the novice learners. This motivates the use of the motion vector feature along with the set-based classifier for depicting the skill level of a learner. Figure 5 on the left, depicts the ratio of FFTs and BHTs identified correctly by individual techniques. The top-3 performing classifiers from figure 3 are used for this purpose. Confidence values of these 3 techniques for correct classification (in figure 4) shows that the Bayesian classifier has the highest confidence for the correct guesses, however the number of correct guesses is lower than that of the other two. Set based classification provides the maximum correct predictions with acceptable average confidence scores. Figure 5 on the right shows the knot detection accuracies using set based classification, obtained by individual learners when the training set has not seen them. It is seen that the results are far worse when the learners are novices instead of intermediates. This is mainly because different learners make different types of mistakes and hence the system needs to see many such cases before generalizing the errors. On the contrary intermediate learners in most cases follow a procedure which can be identified by the system. This is further confirmed from the motion vectors and hand pose maps in figures 6 and 7. Table 1 depicts some skill level identification results. LOU depicts Leave out User, i.e. the user was not seen in the training set, whereas LOT stands for Leave out Test i.e some knots were not performed by the user. The score used in the table is out of 1 and it maps the skill level of the learners without considering whether the knot type was identified correctly or not. In case of the learners (I) , in maximum cases both the skill and knot type was marked correctly, however for novice users this is not the case.
Our work differs from similar work in this field [6, 8] in the fact that pose information from Manus VR gloves was exploited for surgical knot evaluation tasks for the first time. Also, along with classifying the learner as an expert or novice, the type of knot performed is identified. This will give a detailed understanding of where the learner is making more errors. We do not provide a direct comparison of the recognition rates between [6, 8] and our technique as both the dataset as well as the sensor is different.

6 Conclusion

The paper presents an automated evaluation of a knot tying procedure by segmenting it into individual knot sequences. It combines 3D vision and ManusVR hand gloves to accurately segment the component knot ties and uses a kernel set based classification technique classify the sequences as by knot type as well as skill levels subjects. It is also observed that only a few data points from an intermediate learner is required to identify their

 ISBN 978-0-9934207-4-0

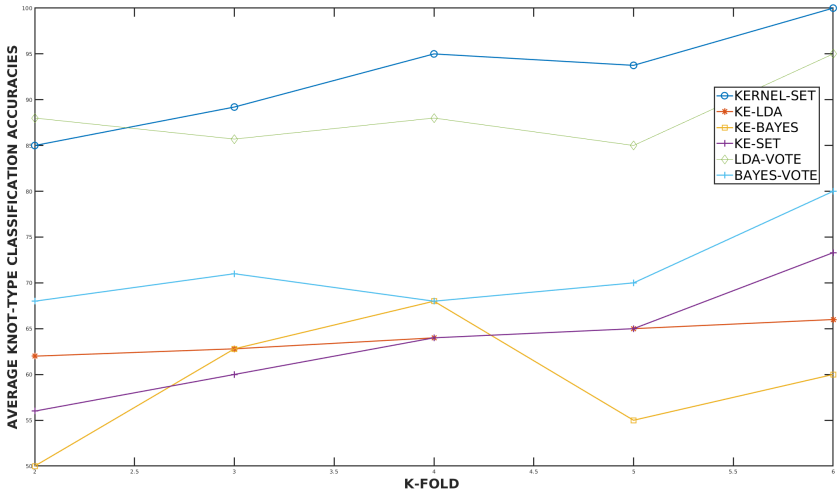

Figure 3: K-fold validation result

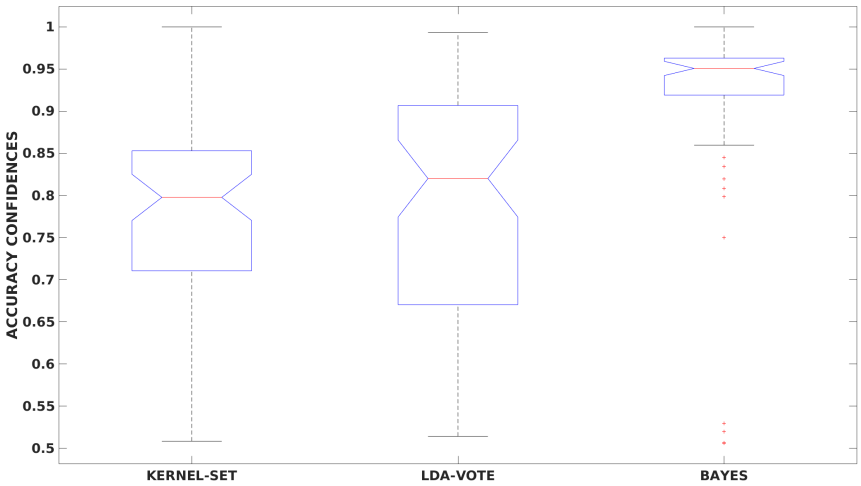

Figure 4: Confidence of correct predictions for 3 techniques. Although Bayes classifier has the maximum confidence for predicted classes, the total number of correct predictions is lower when compared to Kernel-Set.

 ISBN 978-0-9934207-4-0

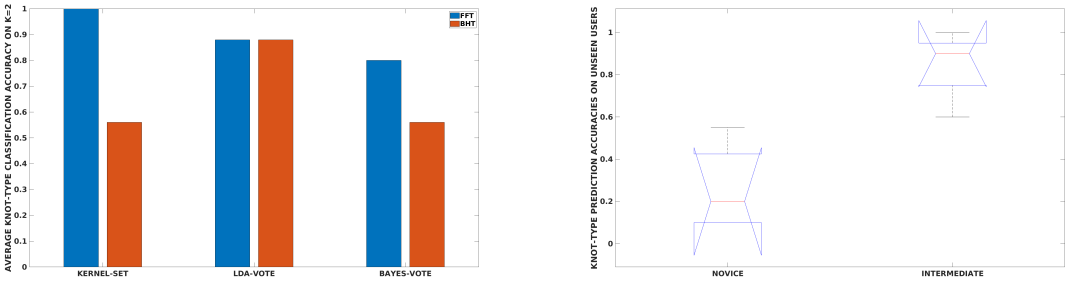

Figure 5: [Left] Classification accuracy of Individual knot-types with each technique, [right] Classification accuracy of knot-types for a set of candidates not seen in training data

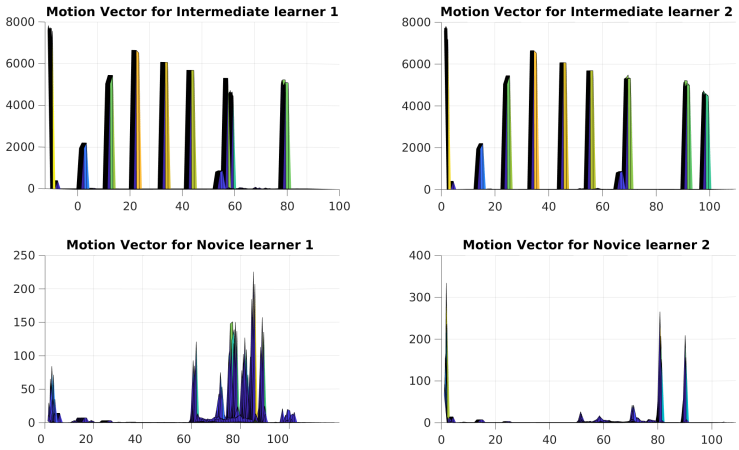

Figure 6: Typical motion vector pattern (y-axis) across the frames (x-axis) for knot sequence of intermediate(I) and novice(N) learners. It is seen that the motion vector are much more structured in case of (I) learners.

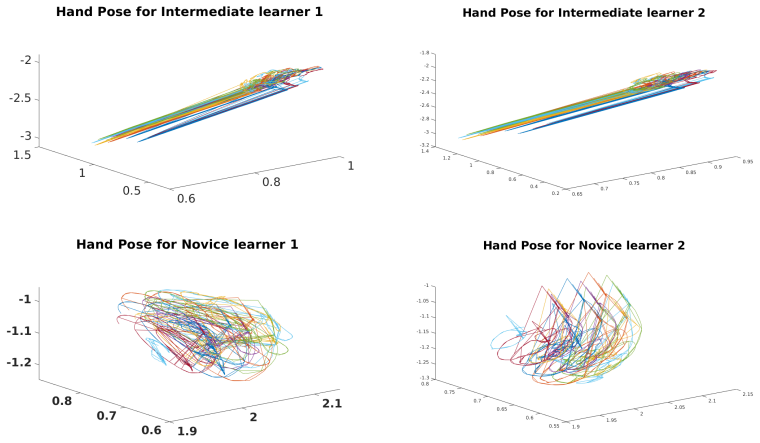

Figure 7: Hand pose (3D joint locations) for knot sequence of intermediate(I) and novice(N) learners. The different colors denote the different fingers. It is seen that the poses are much more structured in case of (I) learners. Also samples of 2 different (I) learners resemble each other when compared to (N) sample handposes.

 ISBN 978-0-9934207-4-0

Features	Level	LOU	LOT
Kernel	I	0.9	0.9
Set	N	0.5	0.6
KE	I	0.88	0.9
SET	N	0.45	0.8
LDA	I	0.9	0.9
Vote	N	0.5	0.7

Table 1: Skill level Identification with different techniques for LOU (User not seen in training set) and LOT (the knot to be classified is not performed by that user in the training set) cases.

motion pattern whereas, the motion patterns of novice users is highly variable. The primary purpose of using a ManusVR glove is to explore knot tying performance with joint pose vectors. The idea can be extended to the use of 3D hand poses from RGB-D image.

References

[1] J.A. Martin et al., "Objective structured assessment of technical skill (osats) for surgical residents," British Journal of Surgery, vol. 84, no. 2, pp. 273–278, 1997

[2] B. B. Haro, L. Zappella, and R. Vidal, "Surgical gesture classification from video data," in MICCAI. 2012

[3] L. Zappella, B. Béjar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," Medical image analysis, 2013.

[4] V. Datta, S. Bann, M. Mandalia, and A. Darzi, "The surgical efficiency score: a feasible, reliable, and valid method of skills assessment," The American journal of surgery, vol. 192, no. 3, pp. 372–378, 2006

[5] G. Lacey, D. Ryan, D. Cassidy, D. Young, "Mixed-Reality Simulation of Minimally Invasive Surgeries" IEEE MultiMedia, vol 14, no 4, pp 76-87, Oct 2007

[6] Yachna Sharma ; Thomas Plötz ; Nils Hammerld ; Sebastian Mellor ; Roisin McNaney ; Patrick Olivier ; Sandeep Deshmukh ; Andrew McCaskie ; Irfan Essa, Automated surgical OSATS prediction from videos, IEEE International Symposium on Biomedical Imaging,2014

[7] Sharma, Y., Bettadapura, V., Plötz, T., Hammerla, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Video based assessment of OSATS using se- quential motion textures. In: International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-Workshop. (2014)

[8] A. Zia, Y. Sharma, V. Bettadapura, E.Sarin, and I. Essa (2017), "Video and Accelerometer-Based Motion Analysis for Automated Surgical Skills Assessment," in Proceedings of Information Processing in Computer-Assisted Interventions (IPCAI), 2017

[9] Ahmidi, N., Ishii, M., Fichtinger, G., Gallia, G., Hager, G.: An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool- motion data. In: International Forum of Allergy Rhinology, Wiley Online Library (2012)

[10] Reiley, C., Hager, G.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: MICCAI. (2009)

[11] Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. Springer (2009) 426–434

[12] 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

[13] Mahtab J. Fard, Sattar Ameri, Ratna B. Chinnam, Abhilash K. Pandya, Michael D. Klein, and R. Darin Ellis, Machine Learning Approach for Skill Evaluation in Robotic-Assisted Surgery,Proceedings of the World Congress on Engineering and Computer Science 2016 Vol I WCECS 2016

[14] Bettadapura, V., Schindler, G., Plotz, T., Essa, I.: Aug- menting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In: CVPR, IEEE (2013)

[15] Ruiping Wang ; Huimin Guo ; Larry S. Davis ; Qionghai Dai ,Covariance discriminative learning: A natural and efficient approach to image set classification, IEEE Conference on Computer Vision and Pattern Recognition,2012.

 ISBN 978-0-9934207-4-0