AUTOMATED SURGICAL OSATS PREDICTION FROM VIDEOS

Yachna Sharma[1], *Thomas Plötz*[2], *Nils Hammerla*[2], *Sebastian Mellor*[2], *Roisin McNaney*[2]
Patrick Olivier[2], *Sandeep Deshmukh*[3], *Andrew McCaskie*[3], *and Irfan Essa*[1]
[1] Georgia Institute of Technology
[2] Culture lab, School of Computing Science Newcastle University, United Kingdom
[3] Newcastle University, United Kingdom

ABSTRACT

The assessment of surgical skills is an essential part of medical training. The prevalent manual evaluations by expert surgeons are time consuming and often their outcomes vary substantially from one observer to another. We present a video-based framework for automated evaluation of surgical skills based on the Objective Structured Assessment of Technical Skills (OSATS) criteria. We encode the motion dynamics via frame kernel matrices, and represent the motion granularity by texture features. Linear discriminant analysis is used to derive a reduced dimensionality feature space followed by linear regression to predict OSATS skill scores. We achieve statistically significant correlation (p-value <0.01) between the ground-truth (given by domain experts) and the OSATS scores predicted by our framework.

Index Terms— OSATS, motion texture, surgical skill, video analysis

1. INTRODUCTION

Developing high quality surgical skills is a time-consuming process, requiring expert supervision and evaluation throughout all stages of the training procedure. This manual assessment of surgical skills poses a substantial resource problem to medical schools and teaching hospitals. In addition, the assessment criteria used are typically domain specific and often subjective where even domain experts do not always agree on the assessment scores (inter-observer variability).

Structured manual grading systems, such as the Objective Structured Assessment of Technical Skills (OSATS) [1], represent the gold standard for (manual) assessments of surgical skills aiming for alleviating the problem of subjective assessments. In this work, we propose a framework for automated assessment of OSATS criteria using video data to alleviate the manual observation requirements and provide objective skill assessments for (prospective) surgeons. Figure 1 shows sample frames from our video dataset. By using video data, the system has minimal requirements of the infrastructure, which is of benefit for large scale deployments.

Automated assessment of surgical OSATS is challenging due to several reasons. First, the OSATS criteria are diverse in nature (Table 1). For example, the "respect for tissue" criterion is based on the trainee's capability in handling the tissue without injuring it. On the other hand, criteria such as "knowledge of procedure" and "time and motion" depend on the trainee's knowledge and orderly task execution. Thus, it is very challenging to encode motion dynamics corresponding to diverse OSATS criteria within a common framework and the task is further complicated by the style variations among surgeons in performing different procedures.

To extract skill relevant information, first we encode the motion dynamics in the videos into frame kernel matrices [2]. We observed that the patterns in the frame kernel matrices vary according to the skill level of the subject. To extract skill relevant information from these patterns, we compute texture features from the frame kernel matrices. Our approach thus enables encoding of motion dynamics into texture features.

We obtained statistically significant correlation between the expert and predicted OSATS scores. With encouraging results, we envision our system to be potentially useful for evaluating medical students in their early training phases.

2. RELATED WORK

There are two domains where assessment of surgical skills has been studied. The first one pertains to skill assessment of surgeons performing robotic minimally-invasive surgery (RMIS). The second domain is assessment of skills in medical schools and teaching hospitals. The state-of-the-art in computerized surgical skill evaluation is dominated by RMIS using robots such as *da-Vinci* [3–5]. In most of the RMIS works, the analysis goal is the automatic recognition of surgical gestures using robotic kinematic data. Very few works have addressed the automated OSATS score prediction. Datta et al. [6] defined surgical efficiency score as the ratio of OSATS "end product quality score" and the number of detected hand movements. Their results indicate significant correla-

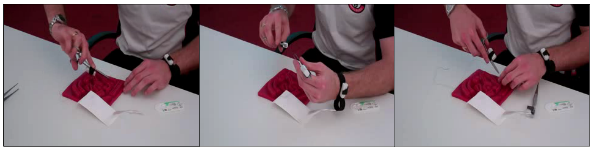

Fig. 1. Sample video frames showing surgical suturing task

Table 1. Summary of OSATS scoring system [1].

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

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

tions between the overall OSATS rating and the surgical efficiency. However, they did not correlate the hand movements to individual OSATS criteria.

With advances in video data acquisition, the attention has shifted towards video based analysis in both RMIS and teaching domains [4,5,7]. Haro et al. [4] and Zapella et al. [5] used linear dynamical systems (LDS) and bag-of-features (BoF) for surgical gesture (surgeme) classification in RMIS surgery using both video and kinematic data. Most of the video based works have reported surgical gesture recognition with few works on surgical skill classification [7]. However, the automated video-based prediction of OSATS scores has not been reported in the published literature.

In this work, we demonstrate that motion texture analysis can be effectively used for prediction of OSATS skill scores. Our results on a diverse data collected in a general surgical lab setting indicate the potential of our framework for OSATS score prediction in medical schools and teaching hospitals.

3. APPROACH

Figure 2 gives an overview of the proposed procedure. The input to the system is a video recording of a trainee surgeon performing suturing procedure and the output is the predicted skill scores corresponding to the seven OSATS criteria. In the following, we will discuss the technical details of the developed framework.

Encoding motion dynamics into frame kernel matrices: A frame kernel matrix defines the similarity between two frames using a kernel function. Frame kernel matrices provide a suitable representation to encode the skill relevant motion dynamics because mapping of data points to the kernel feature space ensures that the motion dynamics depend only on the relative locations of the data points with respect to each other and not on the global origin. In addition, expert motions are more organized, distinct and uncluttered [3], and they are expected to yield well-organized patterns in the frame kernel matrix as compared to the non experts.

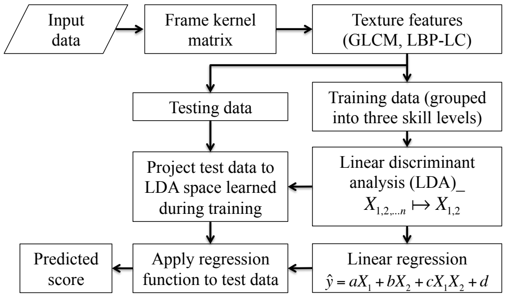
Input data
Frame kernel matrix
Texture features (GLCM, LBP-LC)
Testing data
Training data (grouped into three skill levels)
Project test data to LDA space learned during training
Linear discriminant analysis (LDA)_ $X_{1,2,...n} \mapsto X_{1,2}$
Predicted score
Apply regression function to test data
Linear regression $\hat{y} = aX_1 + bX_2 + cX_1X_2 + d$

Fig. 2. Framework for skill assessment using motion texture analysis.

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be a d-dimensional time series of length n, then the frame kernel matrix $\mathbf{K}$ is given by $\mathbf{K} = \phi(\mathbf{X})^T \phi(\mathbf{X})$, where each entry in $\mathbf{K}$, κ_{ij} defines similarity between two frames x_i and x_j using a kernel function $\phi(x_i)^T \phi(x_j)$. We use the Gaussian kernel function, $\kappa_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$, where σ is the standard deviation. The parameter σ controls the flexibility of the kernel. Small values of σ tend to make the kernel matrix close to an identity matrix. Large values of σ result in a constant kernel matrix. In general, σ is selected empirically to avoid these extremes. We obtain the frame kernel matrices from videos using Algorithm 1.

Algorithm 1 - Frame kernel matrices from videos

```
Require: Surgical videos in set $V$
    Step 1: $\forall v \in V$, compute STIPs (spatio-temporal interest points) and 162-element HoG-HoF (histogram of oriented gradients-histogram of optical flow) descriptors [8].
    Step 2: Cluster STIPs from two experts by applying k-means (k=5) to HoG-HoF features. We select k=5 since we expect approximately five moving entities in the videos: surgeon's two hands and the three instruments (forceps, needle-holder, and scissors).
    Step 3: Assign STIPs for remaining videos to the $k$ clusters learnt in step 2 using minimum Mahalanobis distance.
    Step 4: Compute motion class counts, $\mathbf{X}$, for each of the $k$ clusters. Each entry in the $N \times k$ motion class count matrix $\mathbf{X}$, $x(n, q)$ represents the number of STIPs belonging to the $n^{th}$ frame and the $q^{th}$ cluster, where $N$ is the number of frames in the video.
    Step 5: Compute the frame kernel matrix $\mathbf{K}$.
        $\mathbf{K} = \phi(\mathbf{X})^T \phi(\mathbf{X})$
        Each entry in $\mathbf{K}$, $\kappa_{ij}$, defines similarity between two frames $x_i$ and $x_j$ using a kernel function $\phi(x_i)^T \phi(x_j)$.
        $\kappa_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ where $\sigma$ is the standard deviation. We set $\sigma$ to be the average distance from 20 percent of the closest neighbors resulting in textured frame kernel matrices.
```

Computing motion texture features: We compute texture features from frame kernel matrices to encode motion dynamics in our surgical videos. To test our concept of revealing skill relevant features via motion texture analysis, we use two techniques for texture analysis – Gray Level Co-occurrence Matrix (GLCM) [9], and Local Binary Pattern (LBP) [10].

We employ $N_g \times N_g$ GLCM, calculated for N_g gray levels and eight directions ($0° - 360°$ in steps of $45°$) at a spatial offset of 1 pixel. We compute twenty texture features after averaging and normalizing over the GLCM. These features are [9,11,12]: Autocorrelation, Contrast, Correlation, Cluster prominence, Cluster shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability, Sum of squares variance, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, Information measure of correlation 1, Information measure of correlation 2, Inverse difference normalized, and Inverse difference moment normalized. For LBP, we use the Local Configuration Pattern (LCP) with rotation invariant Local Binary Patterns (LBP) as patterns of interest for the LCP [10]. We use the multi-scale LBP-LC approach with different neighborhood and radius values to capture the motion dynamics at different granularity.

OSATS prediction: We test our framework for predicting OSATS score in a leave-one-out cross validation (LOOCV) scheme. After obtaining textural features using the GLCM and LBP-LC methods, we create a linear regression model using the training data for each OSATS criteria. First, we use Linear Discriminant Analysis (LDA) to find a lower-dimensional feature space in which three (grouped based on OSATS range) skill levels in training data can be discriminated. This provides a discriminating two-dimensional feature space which can be used to predict the skill score of test data. For LDA, we group the training data into three skill levels: low (OSATS score ≤ 2), intermediate ($2 <$ OSATS score ≤ 3.5) and high ($3.5 <$ OSATS score ≤ 5). A linear regression model is obtained using the two dimensions in the reduced LDA feature space.

To predict OSATS scores of a test sample, the test features are first projected to the LDA space learnt during training. The reduced test features are then used to predict the score using the regression function obtained during the training. To evaluate the efficacy of our framework, we calculate the normalized root mean square error (NRMSE) given by, $NRMSE = \sqrt{\frac{\sum(y_n - \hat{y_n})^2}{\sum(y_n)^2}}$ where y_n is the ground truth skill score and $\hat{y_n}$ is the predicted skill score of the nth sample. We also compute the Pearson correlation coefficient R and the corresponding p value between the true and predicted scores.

4. RESULTS AND DISCUSSION

Surgical video data: We recruited 16 participants (medical students) for our case study. Previous suturing expertise and background of the participants varied. Every participant performed suturing activities involving tasks such as stitching, knot tying, *etc.* using a needle-holder, forceps and the silicone suture pads. These training sessions were recorded using a standard video camera (50fps, 1280×720 pixels), which was mounted on a tripod. Fifteen participants performed two sessions of a suturing task. An expert surgeon also performed three sessions giving a total of thirty-three videos. The average duration of the videos is 18 minutes. Ground truth assessment was provided by the expert surgeon based on the OSATS scoring scheme on a scale of [1-5].

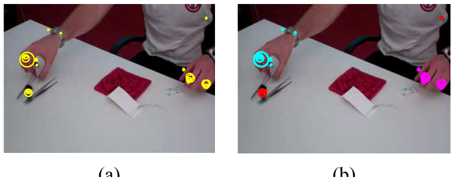

Fig. 3. (a) Detected STIPS in a sample frame represent the moving objects in the scene, (b) STIPs classified into distinct motion classes.

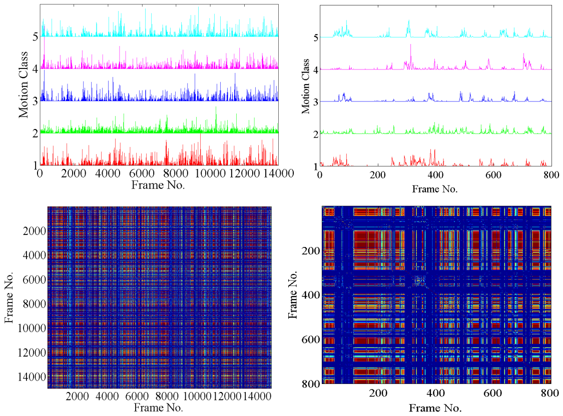

Fig. 4. Top row: Motion class counts for a novice (left), and an expert (right) surgeon. The five classes are plotted at an offset on y axis for clarity. Note that the novice motions exist in almost all frames for all motion classes as compared to fewer motion for expert surgeon. The plots correspond to a single suturing and knot tying task and demonstrate that experts use fewer motions than novices as reported in [6]. Bottom row: Frame kernel matrices corresponding to motion class counts in top row.

Skill relevant motion dynamics: Figure 3 shows the detected STIPs and motion classes in a sample frame. Figure 4 (top row) shows the time frequency counts for five motion classes. It is clear that the pixel intensity transitions in the frame kernel matrices (Figure 4, bottom row) correspond to motion dynamics and vary according to the skill level of the surgeon. Thus, frame kernel matrices provide a suitable representation to encode skill relevant motion.

OSATS prediction: Figure 5 shows the prediction results for three OSATS criteria using LBP-LC features. Similar results were obtained for other 4 criteria using LBP-LC features. Table 2 shows the NRMSE and correlation coefficient R between the ground truth and the predicted OSATS criteria using LBP-LC and GLCM features at different texture granularity. Multi-scale LBP-LC features resulted in high correlation between the true and predicted scores (Table 2, column 4). We also achieved significant correlation with GLCM features for several OSATS criteria, however, overall better performance was achieved with LBP-LC features.

5. SUMMARY AND CONCLUSION

We proposed a video based automated framework for surgical OSATS score prediction in training scenarios using silicone suture pads. Our approach does not involve manual gesture

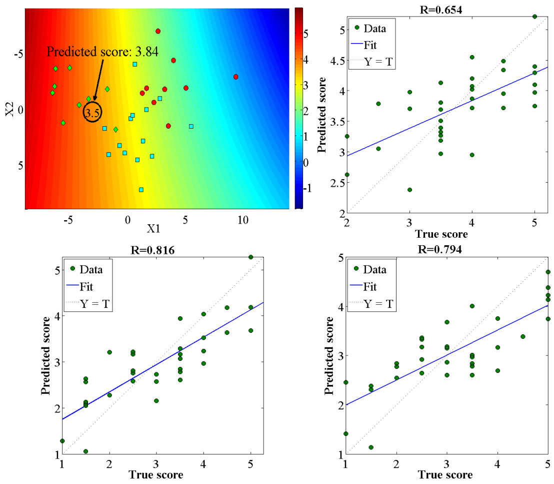

Fig. 5. Top left: Single instance prediction for OSATS criterion TM in LOOCV scheme. Note the separation of experts (green diamonds), intermediates (blue squares) and novices (red circles) in the LDA feature space. The true score is marked with a circle. The colormap corresponds to predicted scores obtained by linear regression. True vs. predicted scores with best fit for OSATS criteria – RT (top right), TM (bottom left), and IH (bottom right).

segmentation. We achieve high correlation between ground truth and predicted OSATS for diverse criteria on videos covering varying skill levels. We plan to extract motion features from surgeon's hands to discount other background motions (*e.g.* moving tissues) and to provide dexterity feedback.

An additional future research direction comprises the incorporation of alternative sensing modalities. For example, accelerometry, with its substantially higher temporal resolution, has the potential of focusing even more on finer-grained details of surgical procedures, which is important for the assessment of a number of OSATS criteria such as "time and motion" or "flow of operation". A few approaches exist that focus on accelerometry based skill assessment (*e.g.*, [13]. These alternative modalities would need, however, modified feature extraction approaches to capture the essentials of the underlying data (*e.g.*, [14]).

Table 2. OSATS prediction using motion texture features

Criteria	Texture feature	NRMSE	R
RT	$\{N_{12}(r_2)\,N_{10}(r_4)\,N_8(r_8)\}$	0.16	0.65^{**}
TM	$\{N_8(r_2)\,N_8(r_4)\,N_8(r_8)\}$	0.20	0.81^{**}
IH	$\{N_8(r_2)\,N_8(r_4)\,N_8(r_8)\}$	0.22	0.79^{**}
SH	$\{N_{12}(r_2)\,N_{10}(r_4)\,N_8(r_8)\}$	0.26	0.67^{**}
FO	$\{N_{12}(r_2)\,N_{10}(r_4)\,N_8(r_8)\}$	0.19	0.71^{**}
KP	$\{N_8(r_2)\,N_8(r_4)\,N_8(r_8)\}$	0.24	0.68^{**}
OP	$\{N_8(r_2)\,N_8(r_4)\,N_8(r_8)\,N_{10}(r_2)\}$	0.17	0.82^{**}
RT	$N_g = 64$	0.26	0.45^{**}
TM	$N_g = 128$	0.34	0.56^{**}
IH	$N_g = 8$	0.30	0.56^{**}
SH	$N_g = 128$	0.39	0.43^{*}
FO	$N_g = 128$	0.36	0.33
KP	$N_g = 128$	0.49	0.45^{**}
OP	$N_g = 64$	0.31	0.52^{**}

"$**$" refers to p value < 0.01, "$*$" refers to p value < 0.05, $N_i(r_j)$ represents the LPB-LC feature evaluated for neighborhood size i around the radius j, N_g refers to the number of gray levels used to compute the GLCM matrix.

6. 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] F. Zhou, F. De la Torre, and J.K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion.," *PAMI*, 2012.

[3] H. Lin, I. Shafran, T. Murphy, A. Okamura, D. Yuh, and G. Hager, "Automatic detection and segmentation of robot-assisted surgical motions," *MICCAI 2005*.

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

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

[6] 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.

[7] V. Bettadapura, G. Schindler, T. Plötz, and I. Essa, "Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition," in *CVPR*, 2013.

[8] H. Wang et al., "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[9] R.M. Haralick, K. Shanmugam, and I.H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics*, , no. 6, pp. 610–621, 1973.

[10] Y. Guo, G. Zhao, and M. Pietikäinen, "Texture classification using a linear configuration model based descriptor.," 2011, pp. 1–10.

[11] L.K. Soh and C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 780–795, 1999.

[12] D.A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Canadian J. Remote Sens.*, vol. 28, no. 1, pp. 45–62, 2002.

[13] A. Möller, L. Roalter, S. Diewald, J. Scherr, M. Kranz, N. Hammerla, P. Olivier, and T. Plötz, "Gymskill:A personal trainer for physical exercises," in *Percom*, 2012.

[14] T. Plötz, N. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *IJCAI*, 2011.