Surgical Skill Assessment on In-Vivo Clinical Data via the Clearness of Operating Field

Daochang Liu[1], Tingting Jiang[1], Yizhou Wang[1,3,4], Rulin Miao[2], Fei Shan[2], and Ziyu Li[2]

[1] NELVT, Department of Computer Science, Peking University, China
[2] Peking University Cancer Hospital
[3] Peng Cheng Lab
[4] Deepwise AI Lab
{daochang, ttjiang}@pku.edu.cn

Abstract. Surgical skill assessment is important for surgery training and quality control. Prior works on this task largely focus on basic surgical tasks such as suturing and knot tying performed in simulation settings. In contrast, surgical skill assessment is studied in this paper on a real clinical dataset, which consists of fifty-seven in-vivo laparoscopic surgeries and corresponding skill scores annotated by six surgeons. From analyses on this dataset, the clearness of operating field (COF) is identified as a good proxy for overall surgical skills, given its strong correlation with overall skills and high inter-annotator consistency. Then an objective and automated framework based on neural network is proposed to predict surgical skills through the proxy of COF. The neural network is jointly trained with a supervised regression loss and an unsupervised rank loss. In experiments, the proposed method achieves 0.55 Spearman's correlation with the ground truth of overall technical skill, which is even comparable with the human performance of junior surgeons.

Keywords: surgical skill assessment · clinical data · neural networks.

1 Introduction

Surgical skill assessment is crucial for the improvement of surgeons' competency during both training and practice [16]. Traditionally, surgical skills are evaluated by experts with onsite observation, which is prone to subjective biases. Although OSATS [13] can reduce such subjectivity to some extent, it requires intensive efforts from experts for manual grading. Thus computer-aided approaches, which provide objective and scalable skill predictions, draw increasing attention from the community. This paper works on computer-aided surgical skill assessment.

Prior works on this task [11, 20, 17, 19, 6, 8, 7, 21, 18, 15, 1] have been mainly set up on simulated datasets such as JIGSAWS [3], in which basic suturing/tying tasks are performed on benchtop models. Although some benchtop models can have high fidelity of human anatomy and accurate imitation of procedure steps, there is still a large gap between simulated scenarios and real-world clinical ones

in terms of varying patient condition, working dynamics, stress level and so on. Only few studies [4, 12, 2, 14] have been conducted on clinical data. As for surgery type, clinical open surgeries have been studied in [4, 2] and laparoscopies in [12, 14]. The modality of these clinical datasets can be surgery videos [4, 12] or the data from external sources [2, 14]. The limitations of the two existing clinical video datasets are that [4] only targets at short suturing/tying clips manually segmented from long surgeries and [12] annotates only four videos with skill labels. Different from previous works, we construct a new clinical laparoscopic video dataset, which consists of *fifty-seven long* procedures.

On the other hand, from the perspective of assessment method, most of the existing works are motion-based. In these works, surgical skills are determined with hand/tool/eye motions, obtained from robotic kinematics [11, 17, 19, 8, 1] or external sensors [20, 7, 2, 14] or visual tracking/interest points [4, 12, 20, 21, 18, 15]. However, acquiring motion trajectories is difficult for clinical data because that 1) robotic kinematics are restricted to only robotic surgeries 2) external sensors interrupt normal workflows and are hard to implement in operating rooms 3) visual tracking is not robust enough and often involves manual corrections. Unlike these prior works, we utilize the clearness of operating field (COF) rather than motion for skill assessment. The COF reflects the amount of bleeding and the visibility of anatomy landmarks. Prediction of the COF relies on appearance information carried in surgery videos, which is more obtainable and robust than motions in clinical settings. Statistical analyses on our clinical dataset identify that the COF is a good skill proxy, for its strong correlation with overall skills and high consistency across annotators. For a detailed review of previous studies, please refer to this survey [16].

In this paper, a new clinical dataset is collected, which includes fifty-seven videos of laparoscopic gastrectomy conducted in operating rooms (OR). Six surgeons annotate the videos with not only technical OSATS scores but also newly designed procedural scores and the COF score. Then we propose an objective video-based method to predict surgical skills via the proxy of COF. A neural network taking in color and semantic features is trained with a supervised regression loss and an unsupervised rank loss collectively. The proposed method outputs skill scores at frame-level to provide feedback. Experiments show that predicting overall skills via the proxy of COF performs better than predicting them directly without proxy. The Spearman's correlation between predicted overall technical skill and ground truth ratings is 0.55, which is promising and even comparable with the human performance of junior surgeons. In summary, our contributions are three-fold: 1) An in-vivo clinical dataset collected from real operating rooms. 2) The identification of COF as a good skill proxy. 3) A video-based method without extra equipment and intensive human efforts.

2   Dataset

Data. The dataset includes videos of 57 in-vivo laparoscopic gastrectomy surgeries for gastric cancer. Videos are captured by the built-in camera (Karl Storz,

Table 1. Skill Metrics

ID	General Metrics	ID	Procedure-specific Metrics
1	Gentleness	7	Dissection in Correct Planes
2	Time and Motion	8	Vessel Exposure and Transection
3	Instrument Handling	9	Venous Breakpoint Selection
4	Flow of Operation	10	Arterial Breakpoint Selection
5	Tissue Exposure	11	Infrapyloric Artery Exposure
6	Overall Technical Skill (OTS)	12	Care for Adjacent Organs
14	Clearness of Operating Field	13	Overall Procedural Skill (OPS)
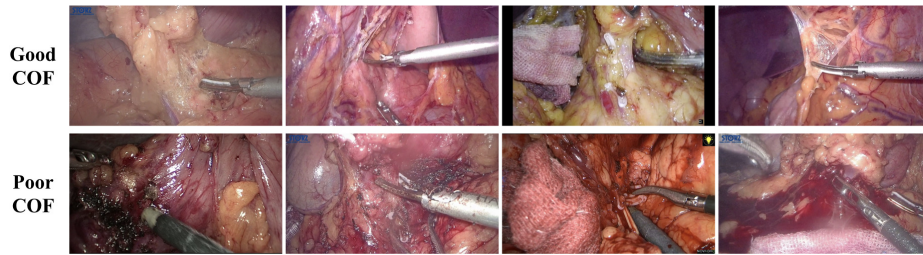
Fig. 1. Example frames from cases with good or poor COF scores.

Olympus or Sony) and are formatted to 960×540 and 25 fps. The procedures are performed by one surgeon. It is assumed that the performance of a same primary surgeon can vary across cases, due to different case complexity, fatigue condition, team members, operating time and so on. To lighten the annotation burden, the infrapyloric area [10], which is one of the four major parts of gastrectomy, is used for skill assessment in this study. The duration of infrapyloric procedure ranges from 8 to 57 minutes and the average is 26 minutes. Surgical skills are annotated by 6 surgeons on 14 metrics with the Likert scale of 1-5. The 6 surgeons include 3 senior surgeons with more than 8 years experience and 3 junior surgeons with less than 4 years experience. For each metric, the ground truth is defined as the mean score of the three seniors.

Skill Metrics. As listed in Table 1, the surgeries are evaluated on 14 metrics, including technical OSATS metrics, procedural metrics, and the COF metric. For OSATS metrics (ID 1-6), a modified version from [5] is employed. Since OSATS metrics only measure the general surgical technique and are procedure-independent, we also design 7 new procedure-specific metrics (ID 7-13) according to [10] to provide fine-grained ratings of surgeons' compliance with gastrectomy instructions. In addition, based on the suggestion of surgeons, the COF metric (ID 14) is proposed to represent how much the bleeding and burned tissues impact on the identification of anatomical structures. Examples are shown in Fig. 1. In cases with good COF, there is no obvious bleeding and the anatomical
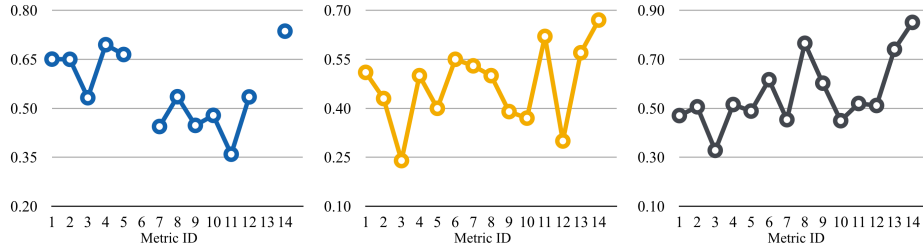
Fig. 2. Metric analyses. **Left:** Correlation with overall skills. **Middle:** Inter-senior consistency. **Right:** Senior-junior consistency. COF (ID 14) is a good skill proxy.

planes and landmarks can be recognized clearly, while in poor cases the bleeding is excessive and severely affects the recognition of planes and landmarks.

Analysis. As shown in Fig. 2, skill metrics are analyzed statistically on all 57 cases in terms of correlation with overall surgical skills and inter-annotator consistency. In the following analyses, $\mathbf{r}_{im}$ denotes the 57 scores from surgeon i on metric m. $\mathcal{S}$ and $\mathcal{J}$ denote the sets of seniors and juniors respectively. And $\mathrm{srocc}(\cdot,\cdot)$ is the function of Spearman's rank correlation coefficient (SROCC).

Correlation with overall skills. We consider the correlations with both overall technical skill (OTS) and overall procedural skill (OPS). Concretely, the correlation with overall skills for a non-overall metric m is defined as:

$$CORR_m = \frac{1}{2 \times |\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{k \in \{6,13\}} \mathrm{srocc}(\mathbf{r}_{im}, \mathbf{r}_{ik}) \quad m \notin \{6, 13\} \tag{1}$$

where the SROCC between this metric and either OTS or OPS is computed, and the results from three seniors are averaged. This value is the largest when $m = 14$, indicating that COF has the best correlation with overall skills.

Inter-senior consistency. This value reflects how seniors agree on a metric. Specifically, the inter-senior consistency for metric m is defined as the averaged SROCCs between the scores of this metric from each two seniors:

$$ISC_m = \frac{1}{|\mathcal{S}| \times (|\mathcal{S}|-1)} \sum_{i,j \in \mathcal{S}, i \neq j} \mathrm{srocc}(\mathbf{r}_{im}, \mathbf{r}_{jm}). \tag{2}$$

The higher this value is, the metric is less affected by subjective biases. Among the 14 metrics, the COF achieves the highest inter-senior consistency.

Senior-junior consistency. To examine whether juniors and seniors have a similar understanding of each metric m, we define senior-junior consistency as the SROCC between the ground truth and the mean scores of the three juniors:

$$SJC_m = \mathrm{srocc}(\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{r}_{im}, \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mathbf{r}_{jm}). \tag{3}$$

A high value means that this metric can be correctly understood by juniors and can thus provide effective feedback to juniors. The COF gives the best value.

In addition to the above reasons, the assessment of COF only relies on appearance information carried in video data. Hand/tool motions and extra devices are not involved. Therefore the COF is identified as a good proxy for overall skills.

3   Method

In this section, an objective video-based method is devised to regress surgical skills via the proxy of COF. The proposed approach takes a video as input and predicts its COF score, which is directly regarded as overall surgical skills.

Preprocessing. The input video is first downsampled to 1 fps to reduce data redundancy. Then extra-abdominal views are removed manually. Note that this can be efficiently done by non-professionals.

Feature Extraction. Since the COF is a measure for bleeding amount and anatomical recognition, we extract color features to describe the severity of bleeding and semantic features to provide anatomy information. For color features, color histograms in RGB space, HSV space, and Red-ratio space (R/G and R/B) are computed in every video frame. In our dataset, videos can have inconsistent color distributions due to different recording devices and patient conditions. Thus the color features are normalized using the first 30% of each video, given the observation that the first 30% is unlikely to contain heavy bleeding and should be of similar color and COF across cases. In detail, for each video, the mean color feature of the first 30% is subtracted from the color feature of each frame. As for the semantic features, the ResNet-101 [9] pretrained on ImageNet is employed in each frame. Then the two types of features are concatenated. After feature extraction, the video is transformed into a feature sequence, denoted by $X \in \mathbb{R}^{T \times D}$. T is the number of frames and D is the feature dimension.

Model. A neural network model is designed for automated skill assessment, which consists of a score branch to evaluate frame quality and a weight branch to provide frame importance. Both of the branches are frame-wise multilayer perceptrons (MLP). Given the feature of a video frame, the score branch produces a score of this frame, while the weight branch outputs a frame weight. In this way, the feature sequence X of the input video is transformed into a score sequence denoted by $A \in \mathbb{R}^{T \times 1}$ and a weight sequence represented by $U \in \mathbb{R}^{T \times 1}$. The weight branch is additionally followed by a softmax function so that the weights of all frames sum to one. Then the video-level score q is obtained by the weighted sum of frame-level scores: $q = \sum_{t=1}^{T} U_t A_t$.

Loss. The loss function comprises a supervised regression loss and an unsupervised rank loss. The regression loss is a standard L1 loss: $L_{reg} = |y - q|$, where y denotes the ground truth COF score and q is the predicted score. In addition, we devise a rank loss based on the observation that the quality of COF decreases over time as the bleeding accumulates. It is assumed that the COF is better at the start of the surgery than in the end. As recommended by surgeons, we define the start section as the first 30% of the surgery and the end section as 60% to 90%. The last 10% is not used for the end section because surgeons commonly clean the operating field thoroughly when finishing. Then the rank loss is proposed to enforce the score of the start section to be higher than the end section by a margin:

$$L_{rank} = \max(0, 1 - (q_s - q_e)) \tag{4}$$

Table 2. Performance of COF Prediction

Method	Feature	PLCC	SROCC
Baseline	Red	0.164	0.130
Baseline	Saturation	0.233	0.178
Baseline	Duration	0.229	0.191
$\mathcal{L}_{reg}$	Color + ResNet	0.580	0.595
$\mathcal{L}_{rank}$	Color + ResNet	0.433	0.457
$\mathcal{L}_{reg} + \mathcal{L}_{rank}$	Color	0.447	0.446
$\mathcal{L}_{reg} + \mathcal{L}_{rank}$	ResNet	0.622	0.601
$\mathcal{L}_{reg} + \mathcal{L}_{rank}$	Color + ResNet	**0.641**	**0.647**
Junior Surgeon	-	0.670	0.657
Senior Surgeon	-	**0.880**	**0.869**

Table 3. Performance of Overall Skills Prediction

Method	PLCC (OTS/OPS)	SROCC (OTS/OPS)
No Proxy ($\mathcal{L}_{reg}$)	0.46 / 0.18	0.47 / 0.18
No Proxy ($\mathcal{L}_{reg} + \mathcal{L}_{rank}$)	0.45 / 0.21	0.45 / 0.24
With Proxy ($\mathcal{L}_{reg} + \mathcal{L}_{rank}$)	**0.56 / 0.40**	**0.55 / 0.41**
Junior Surgeon (COF)	0.56 / 0.61	0.56 / 0.60
Junior Surgeon (OTS/OPS)	0.42 / 0.64	0.41 / 0.62
Senior Surgeon (COF)	0.74 / 0.74	0.73 / 0.73
Senior Surgeon (OTS/OPS)	**0.82 / 0.84**	**0.82 / 0.83**

where $q_s = \sum_{t=1}^{0.3T} U_t A_t$ is the predicted score for the start section and $q_e = \sum_{t=0.6T}^{0.9T} U_t A_t$ is for the end section. This rank loss is only applied to the training cases with COF score no more than 3, since in good cases the COF quality might stay high through the whole surgery. Note that frame weights U are normalized with softmax within the defined ranges when computing q_s and q_e. Then the neural network is trained with the regression loss and the rank loss jointly.

4   Experiments

The proposed method is evaluated on the newly introduced clinical dataset. We repeat three-fold cross-validation 15 times, with 45 runs in total. In each run, 38 videos are chosen randomly for training and the rest 19 videos are for testing. We report the Spearman's rank correlation coefficient (SROCC) and the Pearson linear correlation coefficient (PLCC) between the ground truth and the prediction, which are averaged over all 45 runs. Results of COF prediction and overall skills prediction are both presented.

COF Prediction. Results of COF prediction are provided in Table 2. First, three baselines are tested, which simply depend on the mean red value, the mean saturation value, or duration of the video. Then ablation studies are performed to investigate the impact of feature and loss design. At last, human performances of
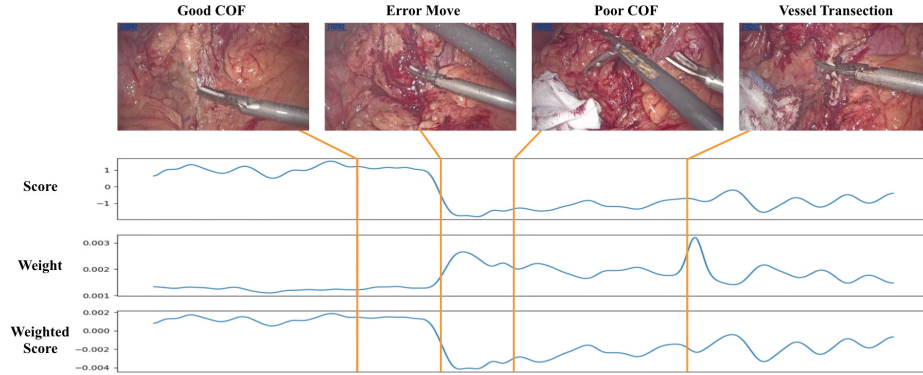
Fig. 3. Frame-level COF scores for a test surgery. The scores have been normalized.

senior and junior surgeons are computed by using annotations of each individual surgeon as the prediction and then averaging the results over juniors or seniors. Our full model surpasses the baselines and the ablated models, which is even comparable with junior surgeons.

Overall Skills Prediction. For overall skills, both OTS and OPS are predicted with or without the proxy of COF. When with the proxy, the predicted overall skills are set the same as the predicted COF. When without proxy, same models are trained to regress overall skills directly by setting the OTS or OPS as the learning target y in the loss. To facilitate the comparison between the proposed method and human performances, performances of surgeons are also computed with proxy (COF as overall skills) or without the proxy. As the results in Table 3, predicting overall skills via the proxy is better than directly regressing them without proxy, which further validates that COF is a good skill proxy. For OTS, the SROCC is 0.55, which is promising and comparable with the human performance of junior surgeons. For OPS, the SROCC is 0.41, which can be improved in the future by explicitly modeling the procedure steps. Note that previous works are not applicable due to different data modality and surgery setting.

Feedback. As shown in Fig. 3, the proposed method generates frame-level scores to indicate which parts are of low quality and frame-level weights to identify which parts should receive more attention. In this test example, the score drop (the 2nd image) corresponds to an error move that the surgeon needs to improve, and the increase of weight (the 4th image) corresponds to the vessel transection that the surgeon should perform with special care. Video demos are provided in the supplementary. The quantification of such feedback is left for future research.

5 Conclusion and Future Work

In this work, a clinical dataset for surgical skill assessment is compiled. Statistical analyses and empirical results both show that the clearness of operating

field is a good skill proxy. A video-based objective model is proposed to predict overall skills via this proxy, which achieves promising results in experiments. The limitation of this work is that the dataset only includes gastrectomy procedures performed by a single surgeon. The generalizability to other procedures and across surgeons needs to be further validated. Future works should also focus on 1) temporal structure modeling 2) incorporating more domain knowledge 3) explicitly modeling the surgery steps and human anatomy.

References

1. 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: MICCAI (2013)
2. Ahmidi, N., Poddar, P., Jones, J.D., Vedula, S.S., Ishii, L., Hager, G.D., Ishii, M.: Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. IJCARS (2015)
3. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE TBE (2017)
4. Azari, D.P., Frasier, L.L., Quamme, S.R.P., Greenberg, C.C., Pugh, C.M., Greenberg, J.A., Radwin, R.G.: Modeling surgical technical skill using expert assessment for automated computer rating. Ann Surg (2019)
5. Birkmeyer, J.D., Finks, J.F., O'reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J., Banerjee, M., Birkmeyer, N.J.: Surgical skill and complication rates after bariatric surgery. N. Engl. J. Med (2013)
6. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? Who's best? Pairwise deep ranking for skill determination. In: CVPR (2018)
7. Ershad, M., Koesters, Z., Rege, R., Majewicz, A.: Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In: MICCAI (2016)
8. Fard, M.J., Ameri, S., Darin Ellis, R., Chinnam, R.B., Pandya, A.K., Klein, M.D.: Automated robot-assisted surgical skill evaluation: Predictive analytics approach. Int J Med Robotics Comput Assist Surg. (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Huang, C.M., Zheng, C.H.: Laparoscopic gastrectomy for gastric cancer: surgical technique and lymphadenectomy. Springer (2015)
11. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Evaluating surgical skills from kinematic data using convolutional neural networks. In: MICCAI (2018)
12. Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L.: Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: WACV (2018)

13. 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 (1997)
14. Richstone, L., Schwartz, M.J., Seideman, C., Cadeddu, J., Marshall, S., Kavoussi, L.R.: Eye metrics as an objective assessment of surgical skill. Ann Surg (2010)
15. Sharma, Y., Plötz, T., Hammerld, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Automated surgical OSATS prediction from videos. In: ISBI (2014)
16. Vedula, S.S., Ishii, M., Hager, G.D.: Objective assessment of surgical technical skill and competency in the operating room. Annu. Rev. Biomed. Eng (2017)
17. Wang, Z., Fey, A.M.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. IJCARS (2018)
18. Zhang, Q., Li, B.: Relative hidden markov models for video-based evaluation of motion skills in surgical training. TPAMI (2015)
19. Zia, A., Essa, I.: Automated surgical skill assessment in RMIS training. IJCARS (2018)
20. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. IJCARS (2018)
21. 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. IJCARS (2016)