2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897932

REPRESENTATION LEARNING USING RANK LOSS FOR ROBUST NEUROSURGICAL SKILLS EVALUATION

Britty Baby[⋆†], *Mustafa Chasmai*[‡], *Tamajit Banerjee*[‡], *Ashish Suri*[†], *Subhashis Banerjee*[‡], *Chetan Arora*[⋆‡]

[⋆]Amar Nath and Shashi Khosla School of Information Technology,
Indian Institute of Technology Delhi, India
[‡] Department of Computer Science Engineering,
Indian Institute of Technology Delhi, India
[†]Department of Neurosurgery, All India Institute of Medical Sciences, New Delhi, India

ABSTRACT

Surgical simulators provide hands-on training and learning of the necessary psychomotor skills. Automated skill evaluation of the trainee doctors based on the video of a task being performed by them is an important key step for the optimal utilization of such simulators. However, current skill evaluation techniques require accurate tracking information of the instruments which restricts their applicability to robot assisted surgeries only. In this paper, we propose a novel neural network architecture that can perform skill evaluation using video data alone (and no tracking information). Given the small dataset available for training such a system, the network trained using ℓ_2 regression loss easily overfits the training data. We propose a novel rank loss to help learn robust representation, leading to 5% improvement for skill score prediction on the benchmark JIGSAWS dataset. To demonstrate the applicability of our method on non-robotic surgeries, we contribute a new neuro-endoscopic technical skills (NETS) training dataset comprising of 100 short videos of 12 subjects. Our method achieved 27% improvement over the state of the art on the NETS dataset. Project page with source code, and data is available at nets-iitd.github.io/nets-v1.

Index Terms— Representation learning, rank loss, surgical skill evaluation, neurosurgery, action quality assessment

1. INTRODUCTION

Different surgical techniques demand deliberate and specific training for hand-eye coordination and hands-on skills [1]. Simulators designed to provide feedback on the level of skills and evaluate the skills acquisition are considered appropriate due to ethical concerns. However, existing training methods rely on subjective evaluation and demands the presence of an expert evaluator to provide the feedback to the trainee. The objective evaluation using scoring scales are burdensome for the evaluator. Further, there are very less professional evaluators to meet the demand of expertise, time and energy required for domain-specific skills assessment. There exists inter-rater variability between different experts on the same input data and evaluation criteria [2, 3]. All these factors demand for an automated skills assessment system.

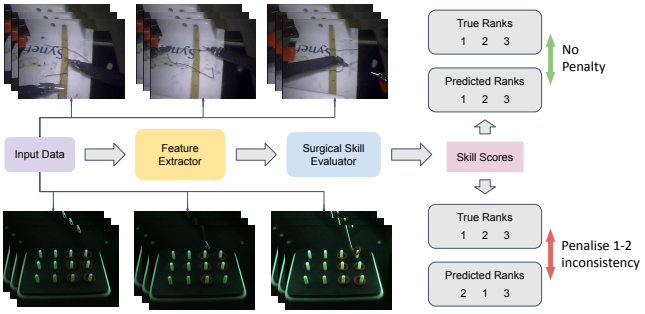

Fig. 1: Block diagram of a general automated surgical skills assessment system. The top, and bottom rows show samples from JIGSAWS, and NETS dataset respectively.

The surgical skills are specific for surgical procedures, and specialised training modules and assessment methods have been developed to address various domains. There are automated skills assessment methods available for robotic assisted minimally-invasive surgery (R-MIS) [4, 5, 6, 7, 8, 9, 10, 11], fundamental laparoscopic surgery (FLS) [3, 12], general surgery (GS) [2, 13, 14] but are very limited for neurosurgery [15]. In neurosurgery, the evaluation is furthermore important due to constraint in anatomy, low margin of error, and high risk of neurological issues in case of any error during surgery. The methods for various surgical branches are customized for the input data, feature extraction module and surgical skills evaluation module.

Fig. 1 shows the fundamental components of a general automated surgical skills assessment system. The majority of the studies address the skills evaluation of R-MIS with the

This work was supported by Department of Biotechnology, Ministry of Science and Technology, India (Project No. BT/PR13455/CoE/34/24/2015)

help of publicly available JIGSAWS dataset [16]. The studies use either the video data [9, 8], or a combination of video and kinematic data [17, 11]. The other surgical skills assessment methods use private datasets with video [3, 2], kinematic data [18] or inertial measurement unit or accelerometer data [14]. On the other hand, most real surgical scenarios [19] and majority of the simulation methods produce only video data. To perform skill evaluation in such scenarios, the feature extractors used in literature have evolved from hand-picked features like path length, time, video spatiotemporal descriptors, discrete cosine transform (DCT), discrete fourier transform (DFT) [3, 2, 13, 20] to data-driven features to deep learning based models [9, 21, 8, 10, 11]. The skills evaluation component has also evolved from posing the skills evaluation as a classification problem [17, 22] to a more detailed regression problem [9, 10, 11].

Recently Liu *et al.* [11] proposed a multi-path framework with surgical tool usage, event patterns, and other skill proxies as the input. This enables a single network to be used for cross-domain surgical skills assessment. We tested this network with a new dataset generated by neurosurgery physical simulator in our lab for neuro-endoscopic skills evaluation. We call this dataset neuroendoscopic technical skills (NETS) training dataset. Our experiments showed that the method does not generalize well on our dataset. Hence, in this paper we suggest focusing on representation learning of the simulator videos. Instead of fixed-value learning (fvl) approach of the regression, we propose a dynamic variant of rank-loss for contrastive learning of video representation along with the MSE loss. This enables the network to learn the discriminative features that predict the relative rank of the skills for pairs and thereby better correlates with the expert surgeon's ranking.

Contributions: The specific contributions of our works are:

- We propose representation learning using rank-loss for cross-domain surgical skills assessment and robust correlation with the expert's ranking.
- We propose a one of its kind dataset for evaluation of neuro-endoscopic surgical skills. The dataset contains 100 short videos obtained from tasks performed by 6 experts and 6 trainees having experiences of > 10 and < 1 man years respectively. For each video, we also release evaluation score given by an expert surgeon with over 20 years of the experience performing neurosurgery.
- We show that our proposed method generalizes well on the JIGSAWS dataset for R-MIS, as well our neuro-endoscopic NETS dataset, showing an improvement of 5% and 27 % respectively over the SOTA method [11].

2. PROPOSED METHODOLOGY

We extend the work on unified framework for skill evaluation [11] with the concepts from representation learning using rank-loss. This allows our model to generalise well even with relatively small, real-world datasets, as well as without using difficult procedures to obtain tracking data. In this section, we first start with a description of the model architecture for the overall evaluation framework, and then describe our rank loss that allows us to learn a better representation space.

2.1. Unified Skill Evaluation

The framework uses separate computation paths to explicitly model different skills aspects relevant for evaluating a surgical activity. Each of these paths expect different input features specific to the respective skill aspect, and the individual scores predicted by each of these paths are combined together to obtain a unified skill score. The rich inter-dependency of the different skill aspects at different stages of the surgical activity is explicitly modelled by a path dependency module while combining individual scores.

In [11], 4 paths are used to model 4 important aspects of a surgical activity, namely visual, proxy, tracking, and events. The visual path expects features extracted from the video frames, the proxy path expects a proxy performance measure like time taken for the activity, while tracking and events paths expect kinematics and event-level features respectively. While the visual features are readily available in most datasets, the last two are often hard to obtain. The majority of works use R-MIS datasets like JIGSAWS, which contains the video, robotic kinematic data and manual annotations for skills score and events[16]. For our analysis on the JIGSAWS dataset, we use all the available features corresponding to four paths, while we use only the visual path for the proposed NETS dataset.

Each computation path in the framework consists of TCN [23] as a feature encoder, followed by scoring functions. The path dependency module takes aggregated features from all the paths to provide temporal importance weights, and the scores from each path are combined together to obtained a unified score. Denoting the encoder for the i^{th} path as $\mathcal{F}_i$, the scoring function as S_i, the path dependency module as W_i and the input features as X_i, the final score predicted by the model is given by:

$$\hat{y} = \sum_{i=0}^{4} \sum_{t} \sigma_t(W_i^t(\mathcal{E}_{\text{all}}))S_i^t(\mathcal{F}_i(X_i)) \tag{1}$$

where σ_t is a *softmax* over time, $\mathcal{E}_{all}$ is a feature vector formed by concatenating the features from all paths, and the suffix t is used to indicate the features at a particular time instant in the surgical video. An overview of this architecture can be seen in Fig. 2.

2.2. K-Rank loss

The score predicted by the model is a scalar real number, and so, it is only natural to use the Mean Squared Error (MSE)

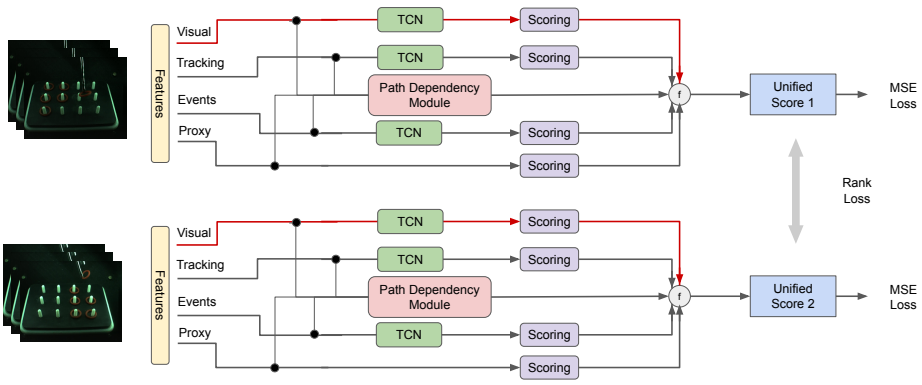

Fig. 2: An overview of the model architecture for skill evaluation. Different features are passed through different paths in the network followed by a path dependency module. The scores are combined using the dependency module as given by Eq. (1)

loss for training the model. However, there are a few disadvantages with this. Firstly, MSE loss tends to allow models to overfit when the training data is scarce. Thus, the representation space learned by the model does not generalise well to unseen test samples. In the absence of sufficient data, contrastive losses have proven to be quite effective [8] in fully, semi as well as self-supervised settings, because of their tendency to drive models towards more structured and robust representation spaces.

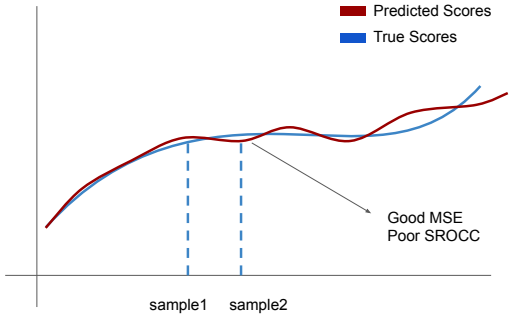

Fig. 3: An example highlighting differences between MSE and SROCC. In the given figure, the MSE will be very good because of small absolute differences. On the other hand, SROCC will be poor because relative ranking between samples 1 and 2 is inconsistent.

Secondly, the skill evaluation algorithm is evaluated using the Spearman's rank correlation (SROCC) as the metric, and not MSE loss. There are subtle differences between the natures of these two scores, and minimising MSE may not lead to a higher SROCC value. An example of such a case can be seen in Fig. 3, where the predicted scores achieve a very low (good) MSE value, but the SROCC value is very poor. While MSE provides a global optimization objective, it alone is not sufficient to fully capture the expectations of the skill evaluator due to the absence of comparative evaluation with respect to other trainees.

To this end, we propose the use of ranking loss for generalized representation learning. While comparing any two pairs of surgical videos, we penalise the model if the relative ordering in the two videos' true and predicted scores are not consistent. For each video, we maintain the predicted and true scores of the previous $k-1$ videos, and compare its scores with each of the $k-1$ previous scores. Looping over the entire dataset once, this would mean comparing each video with all videos within a distance of k videos from it. With randomly shuffled iterations, we expect this to allow us to cover most such pairs. For each of such pair, we compute the k-rank loss as follows:

$$\mathcal{L}_{\text{k-rank}} = \frac{1}{m}\sum_{i=k}^{m}\sum_{j=i-k}^{i-1} \Big[\mathbb{1}\{y^{(i)} > y^{(j)}\}\big\|\text{ReLU}(\hat{y}^{(j)} - \hat{y}^{(i)})\big\|^2 + \mathbb{1}\{y^{(j)} \geq y^{(i)}\}\big\|\text{ReLU}(\hat{y}^{(i)} - \hat{y}^{(j)})\big\|^2\Big], \quad (2)$$

where ReLU is the rectified linear unit, $y^{(i)}$ is the actual score of the i^{th} sample, and $\hat{y}^{(i)}$ is the predicted score, as described in Eq. (1). With this loss, the model is pushed to learn a more robust representation space where the relative ordering of different videos is maintained. Note however, that using the rank loss alone may allow the model to learn a trivial mapping where it predicts the same score for all videos, leading to a 0 loss. To avoid this, we include both the MSE loss and k-rank loss. Thus, the overall loss is:

$$\mathcal{L} = \frac{1}{m}\sum_{i=0}^{m}\|y^{(i)} - \hat{y}^{(i)}\|^2 + \lambda\mathcal{L}_{\text{k-rank}} \quad (3)$$

where λ is for the trade-off between the two losses. Thus, the MSE loss provides a global optimization while the k-rank loss provides local relative optimization leading to an overall better representation space more suited for skill evaluation.

3. EXPERIMENTS AND RESULTS

JIGSAWS Dataset: We used JHU-ISI gesture and skill assessment working set (JIGSAWS) [16], which is one of the largest publicly available dataset for R−MIS. It includes three tasks performed on a bench-top model using da-Vinci system: knot tying, needle passing and suturing. There are synchronized data available for robot kinematics and video at 30 Hz. We use 72 videos for knot tying, 78 videos for suturing task, 56 for needle passing for training our network separately for four-fold cross validation as mentioned in [11].

NETS Dataset: We introduce a new dataset for neuro-endoscopic technical skills (NETS) training obtained from Neuro-endoscope box trainer, which is a box trainer used for imparting neuro-endoscopic skills [24, 15, 25]. The activity performed on the trainer includes transferring 6 rings in a pre-defined manner using a biopsy forceps and endoscope. The activity of ring transfer is recorded by an auxiliary camera placed on the top of the trainer at 25 fps. The whole activity is split into small videos or *surgemes* containing one ring transfer per video (approach, grasp, transfer and retract) and is used for evaluation. The videos were obtained from 6 neuro-endoscopic experts and 6 trainees with experience of > 10 and < 1 years respectively and were divided into 100 surgemes. All 100 short videos were given for blinded subjective evaluation by an expert neurosurgeon on a scoring scale of 1-10 (1-least, 10-highest). 25 videos from this set are randomly selected as a test set and evaluated again at a later point in time to find intra-rater correlation. The correlation between the scores were obtained as 0.92.

3.1. Experimental Setup

For NETS dataset, we use only the vision path and use a ResNet101 model pretrained on ImageNet with 10 crop augmentation to extract 10×2048 dimensional image level features for the visual path similar to the feature extraction strategy used in [11]. The features are first encoded into a lower dimensional embedding sequence, followed by a TCN [23] module to obtain rich spatio-temporal feature representations of the video. The dimensions are reduced from $10 \times 2048 \times T$ to $10 \times 8 \times T$ in this step. The features are passed through MLP layers to obtain the score from them. For JIGSAWS, we directly use all paths features provided by [11].

3.2. Training Hyperparameters

We train on NETS dataset for 950 epochs using Adam optimizer with a learning rate of 0.001 and weight decay of $1e^{-5}$. For JIGSAWS, we use the same training hyperparameters as in [11]. We use a linearly increasing scheduling strategy for the trade-off parameter λ in Eq. (3) because rank-loss becomes more reliable after some iterations. A high value of 'k' in rank-loss would require higher memory to store the k-1 previous features. Hence, a trade-off is required and we experiment with only up to k=5.

Method	KT	NP	SU	Average
JRG [7]	0.36	0.54	0.75	0.57
AIM [6]	0.63	0.65	0.82	0.71
VTPE [11]	0.82	0.76	0.83	0.80
(Ours) 2-rank	**0.85**	**0.80**	**0.86**	**0.84**
(Ours) 3-rank	0.84	0.80	0.85	0.83
(Ours) 4-rank	0.85	0.80	0.80	0.82

Table 1: SROCC values on JIGSAWS for the 4-Fold setting

3.3. Results on JIGSAWS Dataset

We compare the performance of our rank aware model with VTPE [11] on NETS and some other existing methods on JIGSAWS. We consistently obtain better SROCC performance on all tasks for JIGSAWS in the 4-fold cross validation setting and improve upon SOTA by 5%. The results are tabulated in Table 1. In each of the three tasks Knot Tying, Needle Passing and Suturing, we obtained improvements of 3.6%, 5.2% and 3.6% respectively over VTPE [11].

3.4. Results on NETS Dataset

We further evaluate on our NETS dataset and show the generalisability of our method on a very different dataset. We compared our performance with VTPE [11] using only the visual path. We consistently obtain better SROCC performance for different variants of rank loss as shown in Table 2. For our train-test split we obtain an improvement of 27% over the baseline VTPE method.

Method	Test SROCC
Only Vision	0.507
(Ours) 2-rank	0.595
(Ours) 3-rank	0.636
(Ours) 4-rank	**0.648**
(Ours) 5-rank	0.581

Table 2: SROCC values on NETS dataset for test set

4. CONCLUSION AND FUTURE WORK

This paper proposes a general representation learning based deep learning model to automatically assess surgical skills. The effectiveness of the proposed method is validated on benchmark simulated R−MIS dataset. We also contribute a new neurosurgery dataset for skill evaluation, and use it to validate generalization of our method. In future, we plan to explore other feature extractors like I3D and C3D which can potentially capture rich spatio-temporal features for more accurate score prediction.

5. REFERENCES

[1] Patricia L Figert, Adrian E Park, Donald B Witzke, and Richard W Schwartz, "Transfer of training in acquiring laparoscopic skills," *Journal of the American College of Surgeons*, vol. 193, no. 5, pp. 533–537, 2001.

[2] Yachna Sharma et al., "Automated surgical osats prediction from videos," in *ISBI*. IEEE, 2014, pp. 461–464.

[3] Gazi Islam, Kanav Kahol, Baoxin Li, Marshall Smith, and Vimla L Patel, "Affordable, web-based surgical skill training and evaluation tool," *Journal of biomedical informatics*, vol. 59, pp. 102–114, 2016.

[4] Henry C Lin et al., "Automatic detection and segmentation of robot-assisted surgical motions," in *MICCAI*. Springer, 2005, pp. 802–810.

[5] Benjamín Béjar Haro, Luca Zappella, and René Vidal, "Surgical gesture classification from video data," in *MICCAI*. Springer, 2012, pp. 34–41.

[6] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai, "An asymmetric modeling for action assessment," in *ECCV*. Springer, 2020, pp. 222–238.

[7] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng, "Action assessment by joint relation graphs," in *ICCV*, 2019, pp. 6331–6340.

[8] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *CVPR*, 2018, pp. 6057–6066.

[9] Tianyu Wang, Yijie Wang, and Mian Li, "Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels," in *MICCAI*. Springer, 2020, pp. 668–678.

[10] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *CVPR*, 2020, pp. 9839–9848.

[11] Daochang Liu et al., "Towards unified surgical skill assessment," in *CVPR*, 2021, pp. 9522–9531.

[12] Fernando Pérez-Escamirosa et. al, "Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches," *IJCARS*, vol. 15, no. 1, pp. 27–40, 2020.

[13] Aneeq Zia et al., "Automated assessment of surgical skills using frequency analysis," in *MICCAI*. Springer, 2015, pp. 430–438.

[14] Xuan Anh Nguyen et al., "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1–8, 2019.

[15] Britty Baby, Vinkle Kumar Srivastav, Ramandeep Singh, Ashish Suri, and Subhashis Banerjee, "Neuro-endo-activity-tracker: An automatic activity detection application for neuro-endo-trainer: Neuro-endo-activity-tracker," in *ICACCI*. IEEE, 2016, pp. 987–993.

[16] Yixin Gao et al., "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, 2014, vol. 3, p. 3.

[17] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in *MICCAI*. Springer, 2018, pp. 214–221.

[18] Balakrishnan Varadarajan, Carol Reiley, Henry Lin, Sanjeev Khudanpur, and Gregory Hager, "Data-derived models for segmentation with application to surgical assessment and training," in *MICCAI*. Springer, 2009, pp. 426–434.

[19] Amy Jin et al., "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *WACV*. IEEE, 2018, pp. 691–699.

[20] Aneeq Zia and Irfan Essa, "Automated surgical skill assessment in rmis training," *IJCARS*, vol. 13, no. 5, pp. 731–739, 2018.

[21] Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran, "S3d: Stacking segmental p3d for action quality assessment," in *ICIP*. IEEE, 2018, pp. 928–932.

[22] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel, "Video-based surgical skill assessment using 3d convolutional neural networks," *IJCARS*, vol. 14, no. 7, pp. 1217–1225, 2019.

[23] Colin Lea et al., "Temporal convolutional networks for action segmentation and detection," in *CVPR*, 2017, pp. 156–165.

[24] Britty Baby et al., "Design and validation of an open-source, partial task trainer for endonasal neuroendoscopic skills development: Indian experience," *World neurosurgery*, vol. 86, pp. 259–269, 2016.

[25] Ramandeep Singh et al., "Neuro-endoscope box trainer," Jan. 26 2021, US Patent 10,902,745.