Accelerated Video-Based Surgical Skills Assessment through Saliency-Guided Surgical Highlights Reel

Marzie Lafouti, Liane S. Feldman, and Amir Hooshiar
Surgical Performance Enhancement and Robotics Centre (SuPER)
Department of Surgery, McGill University, Montreal, Canada.
Email: amir.hooshiar@mcgill.ca

INTRODUCTION

Objective evaluation of surgical skill is critical for training, credentialing, and quality assurance. However, current approaches rely heavily on expert raters manually reviewing surgical videos, which is time-consuming, labour-intensive, and provides delayed feedback. These limitations restrict the scalability of performance assessment and slow the feedback cycle for trainees [1,2]. Video-based artificial intelligence (AI) offers a promising pathway to automate parts of this process while maintaining clinician oversight. To address these gaps, we developed an interpretable spatiotemporal AI pipeline that classifies surgical expertise directly from operative video while generating saliency-based “highlight reels”. These highlights reveal the critical frames and time points most influential to the model’s inference.

MATERIALS AND METHODS

We proposed an RGB spatiotemporal pipeline that integrates two complementary deep models (Fig 1). InceptionV3 [3], pre-trained on large-scale imagery, extracts compact spatial representations from each frame. A gated recurrent unit (GRU) [4] then modeled temporal dependencies across the sequence to capture motion patterns and consistency of technique. The fused representation is passed to a fully connected classifier that outputs one of three expertise levels: novice (N), intermediate (I), or expert (E). We trained and evaluated the pipeline on the publicly available JIGSAWS dataset [5], which contains videos of suturing, needle-passing, and knot-tying labelled by expertise. Inception and GRU features were precomputed, stored locally, and loaded in batches in training time to accelerate the training epochs. For evaluation, full-length videos at 30 fps were used. A rolling window with stride 10 frames, preserving temporal context while enabling low-latency predictions was adopted for continual evaluation (first-in first-out frames sequence). Two pipeline architectures 1) IncpetionV3+GRU and 2) InceptionV3-OF+GRU. We tested the motion-augmented variant that fused an optical-flow stream with the RGB stream via late feature fusion. The motivation was to capture motion quality, speed, smoothness, and bimanual coordination, thought to be informative for expertise. To enhance AI transparency and identify *‘highlights’ times*, we generated spatial and temporal saliencies.

RESULTS AND DISCUSSION

The RGB spatiotemporal model achieved 95% overall accuracy. Per-class F1 scores were 92% (E), 86% (I), and 99% (N) (Table 1). The confusion matrix demonstrated strong discrimination between novice and expert, with remaining errors primarily between adjacent categories (intermediate vs expert). Saliency overlays consistently focused on tool–tissue interactions, and temporal curves highlighted intuitive peaks during technically demanding phases, providing explanations that align with surgical judgment (Fig 2).

On JIGSAWS this variant underperformed (accuracy ~38%, macro F1 ~0.40, macro recall ~0.49; Table 1). Likely contributors include: noisy flow from camera drift, specular highlights, and bench-top phantom movement, where motion magnitude is a poor proxy for skill allowing noisy motion features to dilute discriminative RGB cues.

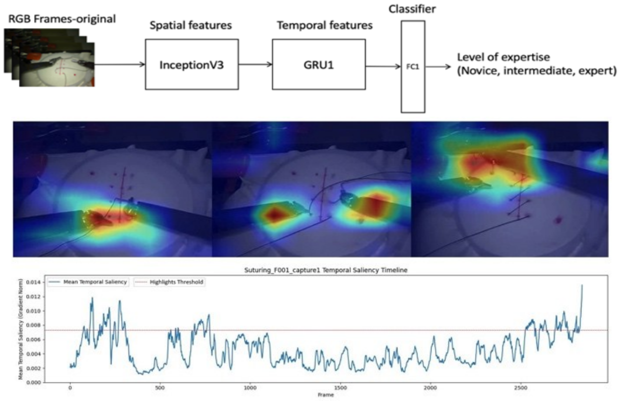

Fig 2 (top) the implemented learning model architecture, (middle) spatial and (bottom) temporal saliency maps.

Table 1: Quantitative comparison of RGB vs. RGB + Optical-Flow.

	Accuracy	Precision	Recall	F1	Class E	Class I	Class N
InceptionV3 + GRU	**0.9524**	**0.9420**	**0.9175**	**0.9234**	P: 0.861 R: 0.989 F1: 0.921	P: 0.990 R: 0.764 F1: 0.863	P: 0.975 R: 0.999 F1: 0.987
InceptionV3-OF + GRU	0.3803	0.6403	0.4911	0.3958	P: 0.244 R: 0.945 F1: 0.388	P: 0.851 R: 0.312 F1: 0.456	P: 0.826 R: 0.217 F1: 0.343

CONCLUSIONS

An interpretable RGB spatiotemporal pipeline classified surgical expertise from video with high accuracy while generating spatial–temporal “highlight reels” that align with surgical judgment. In future work, we will apply a Minimal Sufficient Evidence method to determine the optimal highlights threshold that preserves model accuracy. Validation will be performed on laparoscopic inguinal hernia repairs, with extensions from global labels to task-level metrics and longitudinal trainee tracking, positioning this approach as an objective complement to faculty assessment.

REFERENCES

[1] Madani A et al. *J Surg. Endo*, 2024.
[2] Lafouti M et al. *Hamlyn Symp. Med. Rob*. 2024.
[3] Szegedy C et al. *Proc IEEE CVPR*: 2818-26, 2016.
[4] Cho K et al. *arXiv*:1406.1078, 2014.
[5] Gao Y et al. *MICCAI Workshop*: 1-10, 2014.