ACS 2026 Surgeons and Engineers: A Dialogue on Surgical Simulation

O-10

Research Abstracts

Multimodal Graph Convolutional Networks for Action Recognition in the Operating Room

Benjamin Liu; Ryan Bui; Kiran Nijjer; Lauren Pak; Derek Jiu; Pranav Siddineni; Nicholas Rennie; Jeffrey K. Jopling, MD; Isabele Van Herzeele; and Serena Yeung-Levy

Stanford University, Palo Alto, CA; University of California, Irvine, Winnetka, CA, University of Southern California, Los Angeles, CA; United World College of South East Asia (Dover), Singapore, Singapore; Stanford University, Stanford, CA; University of California, Irvine, Irvine, CA; Ghent University Hospital, Ghent, Belgium; The Johns Hopkins University School of Medicine, Baltimore, MD

Introduction: Graph convolutional networks (GCNs) trained on 3D joint trajectories demonstrate potential for motion analysis in classifying operating room activity. Here, we examine whether adding empirically-derived features improves model performance and which signals are most informative for action recognition.

Methods: We trained a spatiotemporal GCN on 864 five-second clips from simulated endovascular surgical videos, labeled as hand-tool interaction, walking, or peer observation. A multilayer perceptron (MLP) head integrated distance walked, average speed, speed variability, engagement events, and attention switches. Engagement events were defined as sustained gaze within an angular threshold, and attention switches as gaze reorientations above a cosine threshold, both computed over 1-4s gaze windows. We individually evaluated each, then categorically evaluated motion and attention features. Finally, ablations were staged by sequentially adding speed, attention switches, and engagement atop distance.

Results: Our baseline model achieved an F1 and AUPRC of 0.586 and 0.742, respectively. While appending motion metrics improved baseline performance, engagement events led to the strongest individual gains; a 2-second engagement window achieved an F1 and AUPRC of 0.728 and 0.782, respectively. Modeling attention switches over longer windows led to improved performance, with peaks at 4s (F1=0.721; AUPRC=0.771). When examining a combination of empirical features, motion features yielded modest gains, while combining all metrics at 2s produced the strongest performance (F1=0.741; AUPRC=0.771). In scaffolded feature experiments, adding velocity features to distance features reduced performance, but attention and engagement features progressively recovered these drops, with engagement at 3s yielding the best gains (F1=0.722; AUPRC=0.780).

Conclusions: Engagement events were the most informative individually and attention switches also improved performance over longer gaze windows. The full feature set was strongest, with attention features more useful than motion features. These findings highlight potential for real-time deployment of attention-based features in workflow analysis and decision support in the OR.

GCN performance with appended empirical features. Baseline is featureless.

Appended Action Feature	F1	AUPRC	Precision	Recall
None (Baseline)	0.585	0.741	0.630	0.585
Total Distance	0.717	0.752	0.712	0.722
Average Speed	0.610	0.702	0.602	0.652
Speed Variance	0.705	0.777	0.711	0.703
Engagement Events	0.693	0.758	0.694	0.699
Attention Switch Events	0.671	0.748	0.682	0.673
Motion Features	0.631	0.723	0.628	0.634
Interaction Events	0.673	0.733	0.673	0.678
All Features	0.698	0.751	0.692	0.710