ACS 2024 Surgeons and Engineers: A Dialogue on Surgical Simulation Meeting

P-C-06

Research Abstracts

Localizing Steps in Cataract Surgical Videos

Nisarg A. Shah; Shameema Sikder, MD; S. Swaroop Vedula; and Vishal M. Patel

Johns Hopkins University, Baltimore, MD

Introduction: Automated surgical step recognition is pivotal for elevating patient safety and surgical quality. Introducing GLSFormer, a pioneering vision transformer-based method that adeptly merges spatial and temporal information via a gated-temporal attention mechanism. GLSFormer showcases exceptional performance on cataract surgery datasets, offering a pathway to precise and efficient surgical step recognition. This innovation holds immense promise for optimizing surgical outcomes.

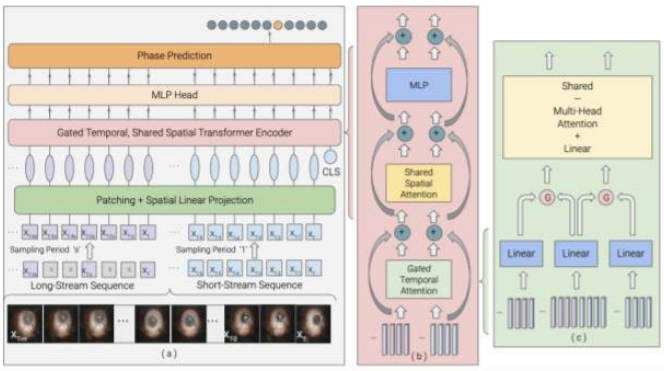
Methods: Our GLSFormer method targets surgical step recognition in videos, utilizing short-term and long-term sequences for predictions. Short-term frames offer recent context, while long-term frames provide broader insights. This fusion enables GLSFormer to capture intricate surgical dynamics.
Patch Encoding: Frames are segmented and transformed into embeddings, followed by merging short-term and long-term patches with positional data.

Temporal and Spatial Attention: Gated Temporal Attention and Shared Spatial Attention modules manage temporal and spatial relationships. This aids in detecting step transitions and local context.
Classification: A cross-entropy-trained MLP processes encoded data for step prediction.
In summary, GLSFormer leverages mixed sequences, temporal-spatial dynamics, and MLPs, advancing surgical step recognition in videos.

Results: GLSFormer was evaluated on Cataract-101 and D99 datasets containing annotated surgical steps. Cataract-101 featured 101 videos with 10 steps, while D99 had 99 videos with 12 steps. Videos were resized and subsampled. Evaluation metrics included Accuracy, Precision, Recall, and Jaccard index.

A comprehensive comparison established GLSFormer's superiority. As a vision transformer-based model, it seamlessly integrated short and long-term spatiotemporal cues, achieving a 6%-11% Jaccard index enhancement. Results underscored GLSFormer's effectiveness in surgical step recognition and its potential for advancing complex video analysis.

Conclusions: In conclusion, our GLSFormer approach, based on vision transformers, excels in surgical step recognition. By fusing short and long-term cues via gated temporal attention, it outperforms existing models. This promises more effective step recognition during surgeries, aiding doctors in improved procedural understanding.



Method	Cataract-101				D99			
	Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard
ResNet [15]	82.64 ± 1.54	76.68 ± 1.86	74.73 ± 1.27	62.58 ± 1.92	72.06 ± 2.12	54.76 ± 2.77	52.28 ± 2.89	37.98 ± 2.97
SV-RCNet [15]	86.13 ± 0.91	84.96 ± 0.94	76.61 ± 1.18	66.51 ± 1.30	73.39 ± 1.64	58.18 ± 1.67	54.25 ± 1.86	39.15 ± 2.03
OHFM [25]	87.82 ± 0.71	85.37 ± 0.78	78.29 ± 0.81	69.01 ± 0.93	73.82 ± 1.13	59.12 ± 1.33	55.49 ± 1.63	40.01 ± 1.68
TeCNO [5]	88.26 ± 0.92	86.03 ± 0.83	79.52 ± 0.90	70.18 ± 1.15	74.07 ± 1.78	61.56 ± 1.41	55.81 ± 1.58	41.31 ± 1.72
TMRNet [16]	89.68 ± 0.76	85.09 ± 0.72	82.44 ± 0.75	71.83 ± 0.91	75.11 ± 0.91	61.37 ± 1.46	56.02 ± 1.65	41.42 ± 1.76
Trans-SVNet [12]	89.45 ± 0.88	86.72 ± 0.85	81.12 ± 0.93	72.32 ± 1.04	74.89 ± 1.37	60.12 ± 1.55	56.36 ± 1.24	42.06 ± 1.51
ViT [8]	84.56 ± 1.72	78.51 ± 1.42	75.62 ± 1.83	64.77 ± 1.97	72.45 ± 1.91	55.15 ± 2.42	53.60 ± 2.63	38.18 ± 2.79
TimesFormer [2]	90.76 ± 1.05	85.38 ± 0.93	84.47 ± 0.95	75.97 ± 1.26	77.83 ± 0.96	64.24 ± 1.20	55.17 ± 1.26	42.69 ± 1.34
GLSFormer	**92.91 ± 0.67**	**90.04 ± 0.71**	**89.45 ± 0.79**	**81.89 ± 0.92**	**80.24 ± 1.02**	**69.98 ± 1.09**	**56.07 ± 1.12**	**48.35 ± 1.22**