Literature review of papers dealing with Surgical skill assessment from video

1.) 2018_Tool detection and Operative Skill Assessment in surgical videos using Region Based Convolution Neural Networks

Methodology:

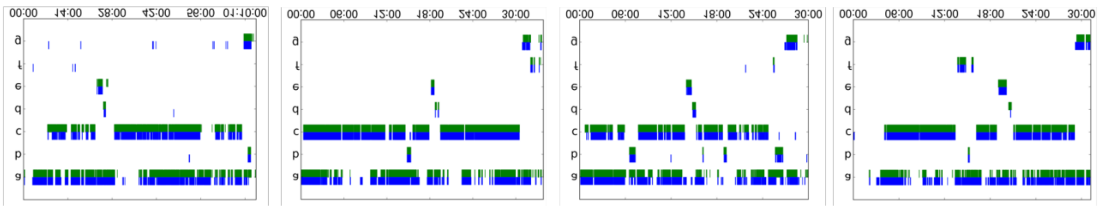
The author proposes a region-based convolutional neural network in which the input is a video frame, and the output is the spatial coordinates of bounding boxes around any of the seven surgical instruments (namely Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator & Specimen Bag). The output is then used to perform analyses of tool movements, from tracking tool usage patterns to evaluating motion economy, and to correlate these measures with surgical skill.

The base network is a VGG-16, which is used to extracts visual features. On top of this network is a region proposal network (RPN) that shares convolutional features with object detection networks. For each input image, the RPN generates region proposals likely to contain an object, and features are pooled over these regions before being passed to a final classification and bounding box refinement network. The network is pre-trained on the ImageNet dataset and then fine-tuned on m2cai16-tool-locations dataset which contains spatial tool annotations and surgical instrument classes.

To assess instrument usage patterns, the author generates timelines displaying tool usage over the course of each of the testing videos.

Surgical skills are evaluated **manually** based upon the total distance travelled, timelines and other metrics.

A new dataset has also been introduced, m2cai16-tool-locations, which extends the already existing m2cai16-tool dataset with spatial annotations of tools. This dataset is publicly available.



Reasons behind this approach:

- It broadens the scope of skill evaluation because manually performed surgery can now be evaluated.
- Unlike all other previous works, it is not task-specific, a limitation to task-specific studies is that surgical tasks and gestures differ substantially from and do not accurately reflect surgical performance in real-world surgeries.

- Instead of using short segments of procedures or of surgeons performing simulated training tasks to analyse surgical skill, it leverages unedited, full-length surgical operations. This differentiation is crucial for performance assessment, since real-time operations include smoke, lens fogging, variable anatomy, and different usage patterns not found in simulation scenarios. Limited segments of real operations may not give a comprehensive assessment of surgical performance.
- The use of the RPN enables significant computational gains over related previous work.

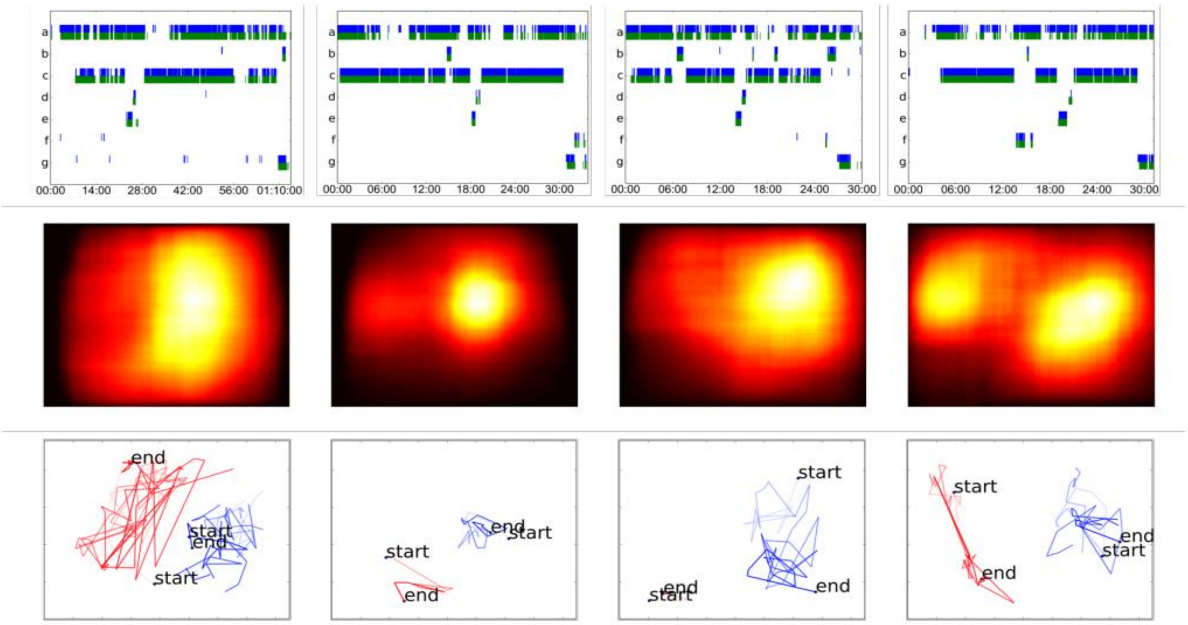
Research gaps:

- Surgical skills are evaluated **manually** based upon the total distance travelled, timelines and other metrics.
- There is no provision for giving feedback to the surgeon, which can be made possible once you have the spatial coordinates of surgical tools.

Potential fixes:

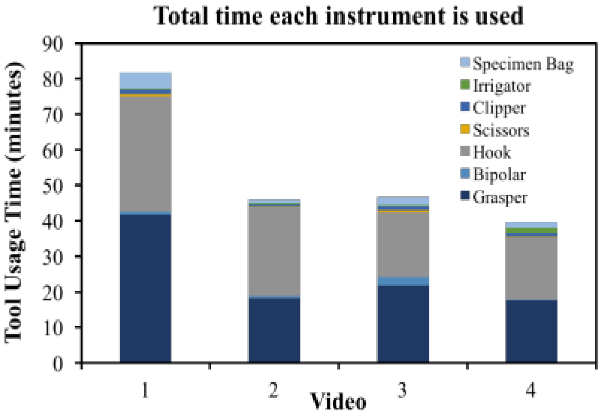
- The system of evaluating skills can be made automatic.
- A feedback system can be built using CAM (class activation map), once we have the spatial coordinates of the tools.

Results:

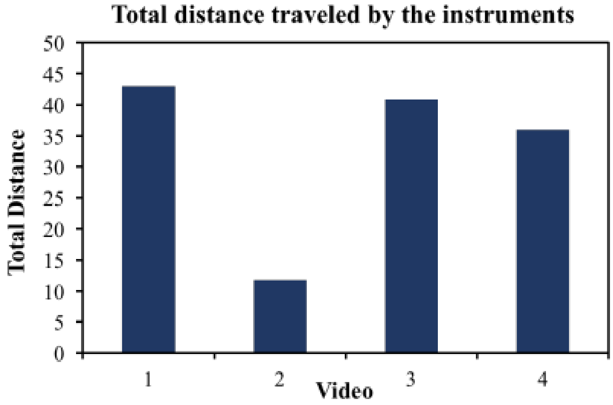


(1) (2) (3) (4)

Figure 6: Timelines (top), heat maps (middle), and trajectories (bottom) of tool usage for the testing videos 1 through 4 in m2cai16-tool. In the timelines, (a)-(g) correspond to Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator, and Specimen Bag, respectively. These metrics effectively measure bimanual dexterity, efficiency, and overall operative skill and enable us to efficiently examine back and forth switching of instruments, movement range, and motion patterns of tools. We find that testing video 2 correlates with the most well-executed surgery, reflecting focused and skillful execution of each step of the surgical procedure. In contrast, the surgeons in the other testing videos have much less economy of motion, handle the instruments with less dexterity, and struggle with certain parts of the procedure.



Total instrument usage times, by video. Reflecting level of skill in handling tissue, the longer presence of the bipolar in testing video 3 indicates tissue damage, as the bipolar is used to stop bleeding.



2.) 2019_Video-based surgical skill assessment using 3D convolutional neural networks

Methodology:

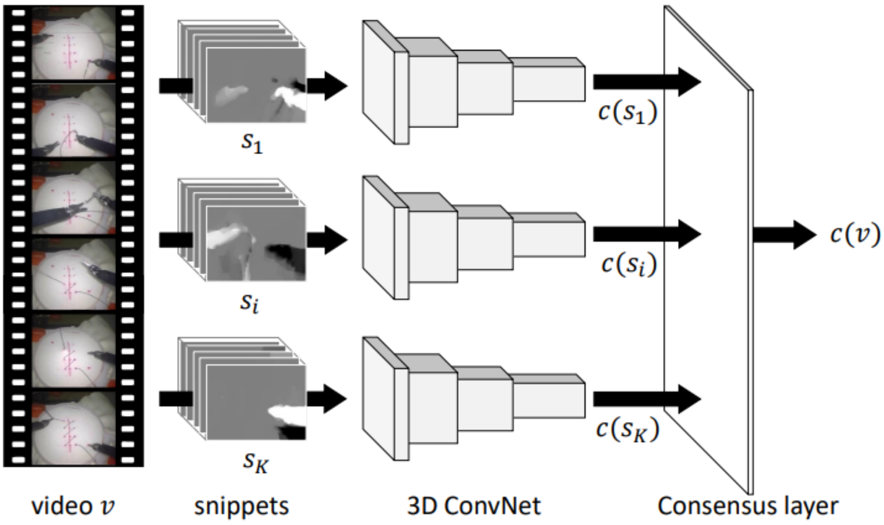
Video is divided into k segments & n consecutive random frames are chosen from each segment, which are then passed to Inception-v1 I3D model to get spatial as well as temporal features from stack of frames. The final dense layer of Inception v1 I3D is changed (so that model can adapt to our problem) & now has three neurons (each representing a skill level). The individual classification result from each segment is obtained and all the results are averaged to get the final skill level.

Reasons behind this approach:

- To avoid extracting temporal and spatial features differently.
- Dealing with short snippets instead of complete videos is an effective strategy to reduce the size and complexity.

- Segment-based snippet sampling scheme at training time already realizes some kind of data augmentation. This is appealing because the available training data for the problem of surgical skill assessment is still very limited.

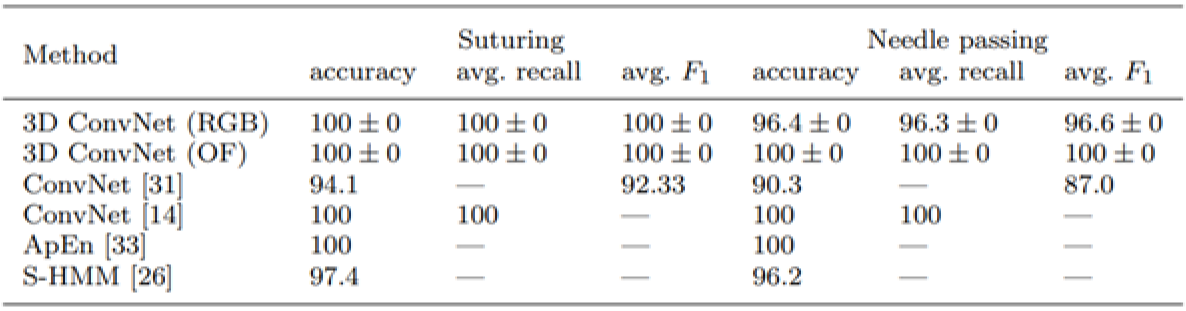
Architecture:



Research gaps:

A 3d convolution layer requires a high computational cost and consumes a lot of memory.

Results:



Method	Suturing			Needle passing		
	accuracy	avg. recall	avg. F_1	accuracy	avg. recall	avg. F_1
3D ConvNet (RGB)	100 ± 0	100 ± 0	100 ± 0	96.4 ± 0	96.3 ± 0	96.6 ± 0
3D ConvNet (OF)	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
ConvNet [31]	94.1	—	92.33	90.3	—	87.0
ConvNet [14]	100	100	—	100	100	—
ApEn [33]	100	—	—	100	—	—
S-HMM [26]	97.4	—	—	96.2	—	—

Potential fixes:

A simpler & less memory extensive 3d convolution is possible with a competitive accuracy of 100% on knot tying and 97.2% on Suturing Task.

Code at can be found at my git(https://github.com/Bhavya-2k03/SurgicalSkillClassificationFromVideo)

3.) 2022_Surgical Skill Evaluation from Robot-Assisted Surgery Recordings

Methodology:

The task of feature extraction is decomposed into 2 phases: transfer-learning based intra-frame local feature extraction and an end-to-end inter-frame temporal feature learning model.

Spatial feature extraction is done by passing p frames into the pretrained model (here ResNet 50) and temporal features extraction through 1-D CNN.

The spatial features are then incorporated into the model design by using FFT.

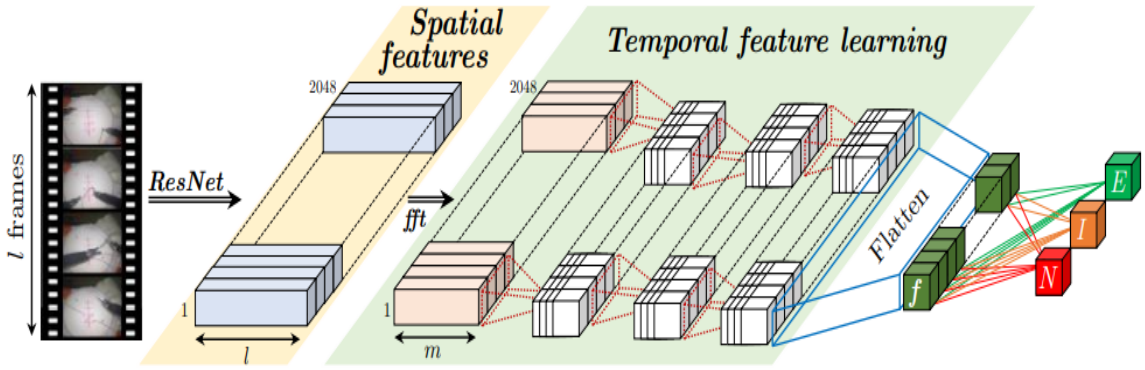
Input to ResNet 50 is a video with p frames. The output is a 2048×p feature tensor. To learn the inter-frame temporal features associating with the surgeon's gestures and maneuvers, 2048×p FFT frequency tensor is truncated and the first m FFT coefficients are kept for every 2048 features. The new data tensor of 2048×m is then fed to a 1-D CNN model for temporal feature learning.

Finally, the obtained deep representations of surgical dexterity levels are passed to a classification head for skill assessment.

Reason behind this approach:

- Basic surgical tasks are inherently repetitive and sequential. Encoding the time-series actions into the frequency domain essentially facilitates retrieving information (e.g., smoothness, jittery motions, abrupt movements, etc.) that differentiates skill levels of surgeons.
- Additionally, with the prior domain knowledge on surgical activity frequency range, FFT helps to address the issue that surgery recording may have different frame lengths.
- 3D CNN is memory extensive and thus require a lot of computational resources, therefore this model acts as an alternative.
- The two-stage learning in this method reduced the learning complexity and led to a lightweight model over 2-D CNN+RNN models in prior arts.

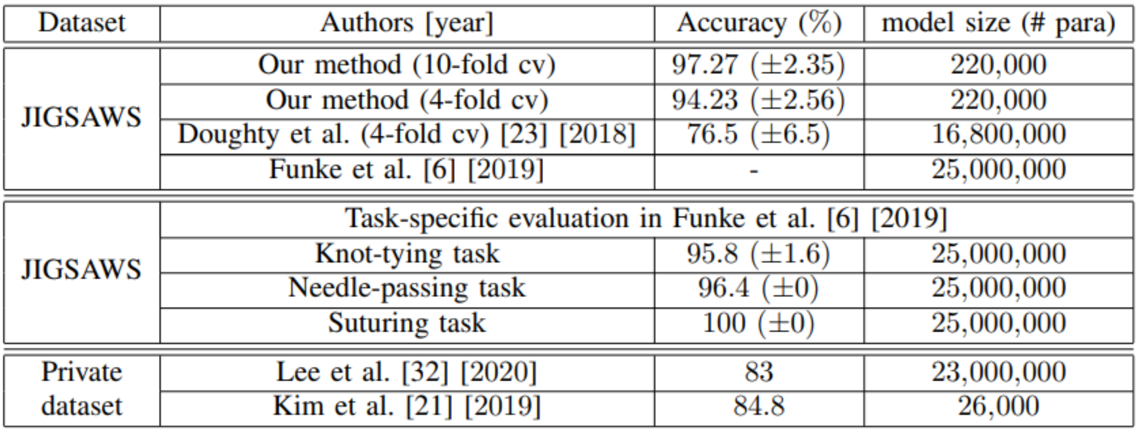
Architecture:



Research gaps:

Lower in accuracy in needle passing task (96.4 %), when compared to 3D CNN (100%).

Results:



Dataset	Authors [year]	Accuracy (%)	model size (# para)
JIGSAWS	Our method (10-fold cv)	97.27 (±2.35)	220,000
	Our method (4-fold cv)	94.23 (±2.56)	220,000
	Doughty et al. (4-fold cv) [23] [2018]	76.5 (±6.5)	16,800,000
	Funke et al. [6] [2019]	-	25,000,000
JIGSAWS	Task-specific evaluation in Funke et al. [6] [2019]		
	Knot-tying task	95.8 (±1.6)	25,000,000
	Needle-passing task	96.4 (±0)	25,000,000
	Suturing task	100 (±0)	25,000,000
Private dataset	Lee et al. [32] [2020]	83	23,000,000
	Kim et al. [21] [2019]	84.8	26,000

4.) 2022_Surgical Skill Assessment via Video Semantic Aggregation

Methodology:

The author has proposed a Video Semantic Aggregation (ViSA) framework, for surgical skill assessment. Frames from a fixed number of timesteps are the inputs to the model.

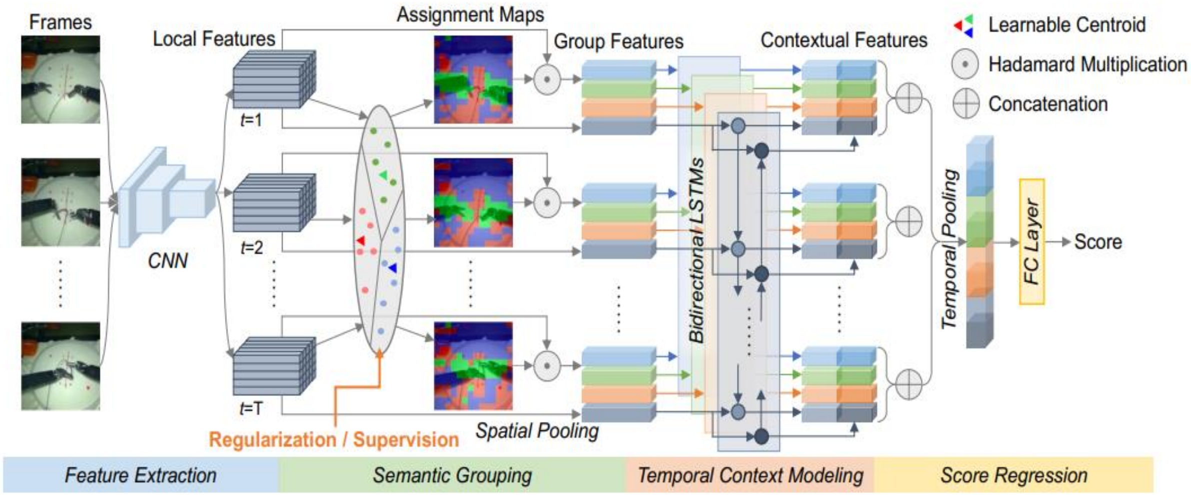
Firstly, spatial features are extracted from the frames through CNN. Then the similar local semantic features are aggregated through clustering and the abstract features for each semantic group is generated (known as the semantic grouping stage).

The features corresponding to the same semantic (across time) are aggregated and their temporal relationship is modelled through bidirectional LSTMs. The final score is then predicted based on the obtained spatiotemporally aggregated features.

Reason behind this approach:

Existing works proposing a CNN-LSTM joint framework miss the point that modelling long-term relationships by LSTMs on spatially pooled CNN features would neglect the difference among semantic concepts such as tools, tissues, & background in the spatial dimension. Hence consequent networks could hardly model the temporal relationship of the local features in different spatial parts separately, e.g., the movements of different tools and the status changes of tissue.

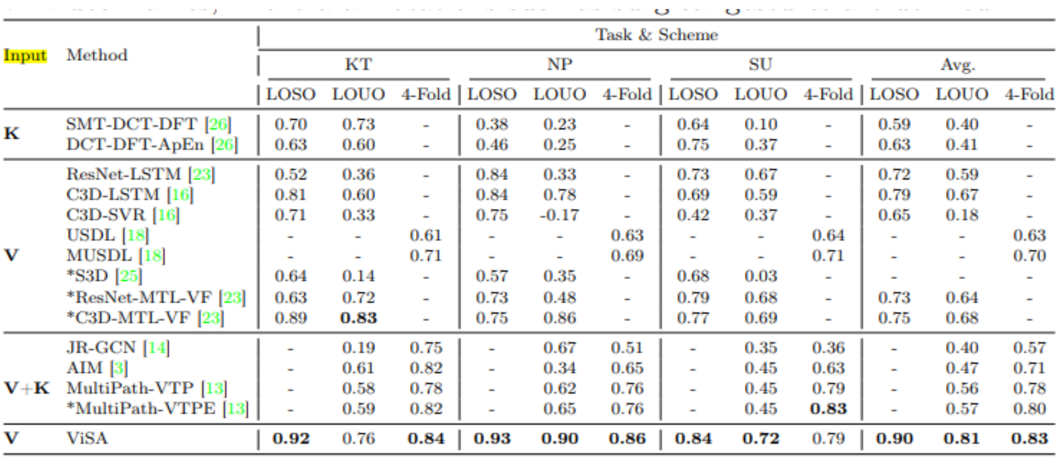
Architecture:



Results:

Spearman’s Rank Correlation coefficient (-1 to 1)

- 0.3 - 0.5 → Low correlation
- 0.5 - 0.7 → Moderately correlated
- 0.7 – 1.0 → Highly correlated



Input	Method	Task & Scheme											
		KT			NP			SU			Avg.		
		LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold
K	SMT-DCT-DFT [26]	0.70	0.73	-	0.38	0.23	-	0.64	0.10	-	0.59	0.40	-
	DCT-DFT-ApEn [26]	0.63	0.60	-	0.46	0.25	-	0.75	0.37	-	0.63	0.41	-
V	ResNet-LSTM [23]	0.52	0.36	-	0.84	0.33	-	0.73	0.67	-	0.72	0.59	-
	C3D-LSTM [16]	0.81	0.60	-	0.84	0.78	-	0.69	0.59	-	0.79	0.67	-
	C3D-SVR [16]	0.71	0.33	-	0.75	-0.17	-	0.42	0.37	-	0.65	0.18	-
	USDL [18]	-	-	0.61	-	-	0.63	-	-	0.64	-	-	0.63
	MUSDL [18]	-	-	0.71	-	-	0.69	-	-	0.71	-	-	0.70
	*S3D [25]	0.64	0.14	-	0.57	0.35	-	0.68	0.03	-	-	-	-
	*ResNet-MTL-VF [23]	0.63	0.72	-	0.73	0.48	-	0.79	0.68	-	0.73	0.64	-
	*C3D-MTL-VF [23]	0.89	**0.83**	-	0.75	0.86	-	0.77	0.69	-	0.75	0.68	-
V+K	JR-GCN [14]	-	0.19	0.75	-	0.67	0.51	-	0.35	0.36	-	0.40	0.57
	AIM [3]	-	0.61	0.82	-	0.34	0.65	-	0.45	0.63	-	0.47	0.71
	MultiPath-VTP [13]	-	0.58	0.78	-	0.62	0.76	-	0.45	0.79	-	0.56	0.78
	*MultiPath-VTPE [13]	-	0.59	0.82	-	0.65	0.76	-	0.45	**0.83**	-	0.57	0.80
V	ViSA	**0.92**	0.76	**0.84**	**0.93**	**0.90**	**0.86**	**0.84**	**0.72**	0.79	**0.90**	**0.81**	**0.83**

5.) 2022_Video-based Surgical Skills Assessment using long term Tool Tracking

Methodology:

Frames are the input to the model. Tools in the frame are detected (independent of other frames) and tracks are then created by linking the corresponding detections across time. A similarity score is calculated for every pair of new detections and active tracks, to match new data to tracks through data association.

At each frame, YOLOv5 (trained on a subset of CholecT50) is used to detect all tools present in the frame. **The dataset has been annotated with the bounding box location and the class of each surgical instrument present in the scene.**

Trajectories are then decomposed and their variations over time is used in the calculation of motion metrics (such as distance (path length), velocity, acceleration, jerk, curvature etc.) that describe the path of the tool. These features are then used in a random forest model to classify surgeons into the high and low efficiency classes.

Reason behind this approach:

- It broadens the scope of skill evaluation because manually performed surgery can now be evaluated.
- Unlike all other previous works, it is not task-specific, a limitation to task-specific studies is that surgical tasks and gestures differ substantially from and do not accurately reflect surgical performance in real-world surgeries.

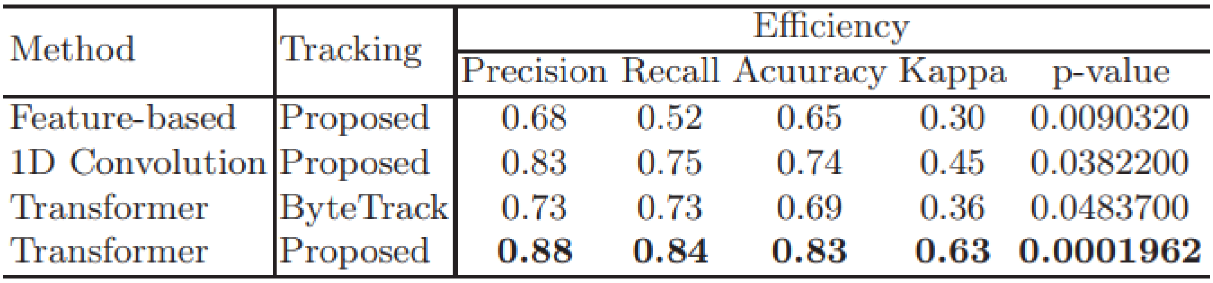
Research gaps:

Dataset needs to be annotated for training YOLOv5 for tools detection.

Potential fixes:

Instead of manually annotating a new dataset, we can use Dm2cai16-tool-location dataset which can be used to train yolov5 for surgical tool classification and detection.

Results:



Method	Tracking	Efficiency				
		Precision	Recall	Acuuracy	Kappa	p-value
Feature-based	Proposed	0.68	0.52	0.65	0.30	0.0090320
1D Convolution	Proposed	0.83	0.75	0.74	0.45	0.0382200
Transformer	ByteTrack	0.73	0.73	0.69	0.36	0.0483700
Transformer	Proposed	**0.88**	**0.84**	**0.83**	**0.63**	**0.0001962**

6.) 2020_Surgical Skill Assessment on In-Vivo Clinical Data via the Clearness of Operating Field

Methodology:

The author proposes to utilize the clearness of operating field (COF) for skill assessment. The COF reflects the amount of bleeding and the visibility of anatomy landmarks.

First of all, Colour features are extracted to describe the severity of bleeding and semantic features to provide anatomy information. For colour features, colour histograms in RGB space, HSV space, and Red-ratio space (R/G and R/B) are computed in every video frame. As for the semantic features, the ResNet-101 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.

After feature extraction, the video is transformed into a feature sequence and fed to model which consists of a score branch to evaluate frame quality and a weight branch to provide frame importance. Given the feature of a video frame, the score branch produces a score of this frame, while the weight branch outputs a frame weight. The weight branch is then followed by a SoftMax function so that the weights of all frames sum to one. Then the video-level score is obtained by the weighted sum of frame-level scores.

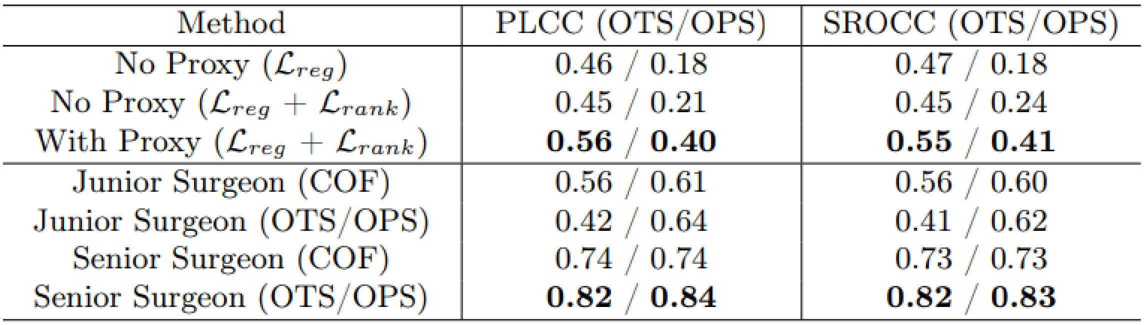
Reason behind this approach:

The author has argued that visual tracking is not robust enough and often involves manual corrections.

Research gaps:

Extra-abdominal views need to be removed manually.

Results:



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**

7.) 2021_Towards unified surgical skill Assessment

Methodology:

The author identifies three aspects from the medical literature that are likely to characterize surgical skills and also suitable for automatic assessment, i.e., surgical tool usage, surgical field clearness, and surgical event pattern. The first aspect is the movement of surgical tools, which could reflect the instrument handling proficiency and motion efficiency of the surgeon. The second aspect is the clearness of the operating field as a skill proxy. Skill proxy means an indirect indicator that is statistically correlated to surgical skills. The third aspect is the workflow of surgical events or actions.

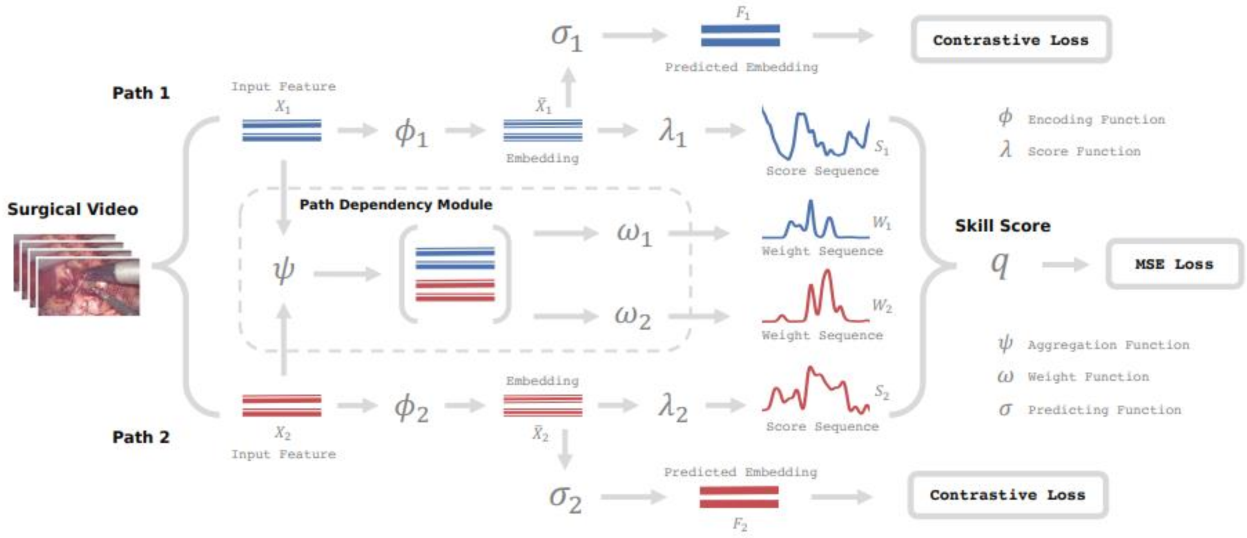
The framework comprises multiple paths in parallel, with each corresponding to a skill aspect. Aspect-specific feature sequences extracted from surgical videos are forwarded along each path, subsequently transformed into skill score sequences for each aspect.

Feature sequences are aggregated from all the skill aspects to provide temporal importance weights for the score sequences. Lastly, the weighted score sequences are pooled over time and fused across paths as the final assessment prediction.

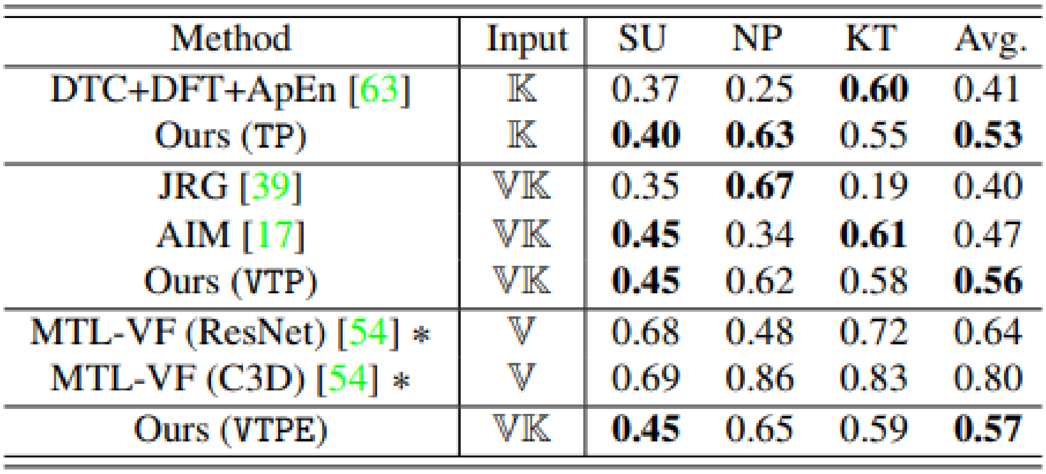
Reason behind this approach:

- Prior works on automated surgical skill assessment, mostly rely on only one of the aspects listed in methodology.
- Captures dependencies among aspects, which were not addressed in previous work.

Architecture:



Results:



Method	Input	SU	NP	KT	Avg.
DTC+DFT+ApEn [63]	𝕂	0.37	0.25	**0.60**	0.41
Ours (TP)	𝕂	**0.40**	**0.63**	0.55	**0.53**
JRG [39]	𝕍𝕂	0.35	**0.67**	0.19	0.40
AIM [17]	𝕍𝕂	**0.45**	0.34	**0.61**	0.47
Ours (VTP)	𝕍𝕂	**0.45**	0.62	0.58	**0.56**
MTL-VF (ResNet) [54] *	𝕍	0.68	0.48	0.72	0.64
MTL-VF (C3D) [54] *	𝕍	0.69	0.86	0.83	0.80
Ours (VTPE)	𝕍𝕂	**0.45**	0.65	0.59	**0.57**

8.) 2020_Multitask Learning for Video-based Surgical Skill Assessment

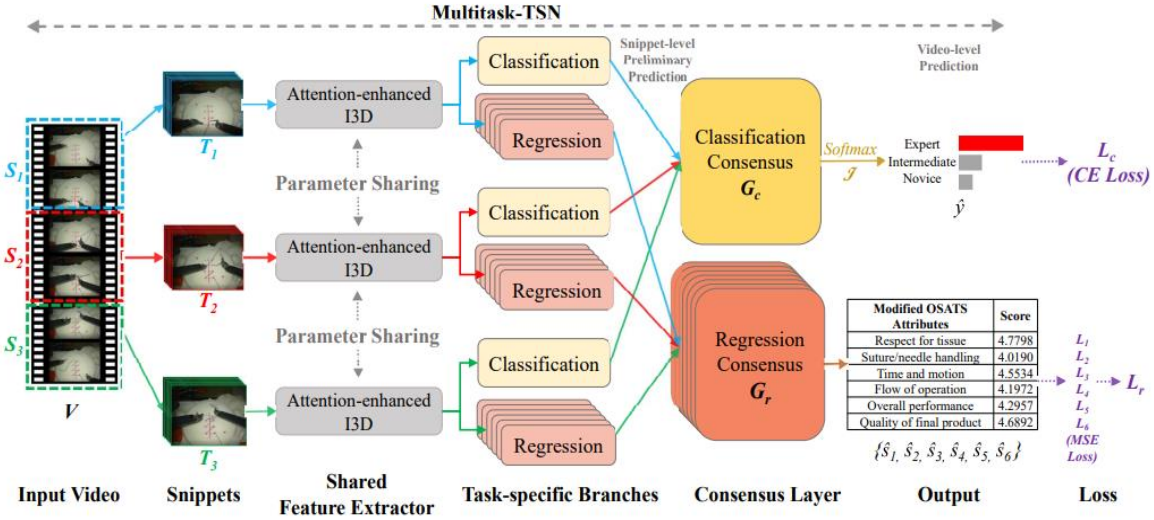
Methodology:

The author proposes a Multitask-Temporal Segment Network (Multitask TSN) framework and form a First, Multitask-TSN divides the input video sequence uniformly into segments and samples a short snippet with a predefined length randomly from each segment, allowing us to extract content covering the entire video with relatively low computational overhead. Then, each snippet is fed to the feature extractor (Attention enhanced I3D) shared for all tasks, followed by multiple task specific Classification and Regression branches. This generates snippet-level preliminary predictions for the target tasks. Finally, in the consensus layer, the snippet-level information for each task is aggregated across all the snippets using the classification aggregation function and the regression aggregation function to form video-level multitask predictions, and to produce our final output.

Reason behind this approach:

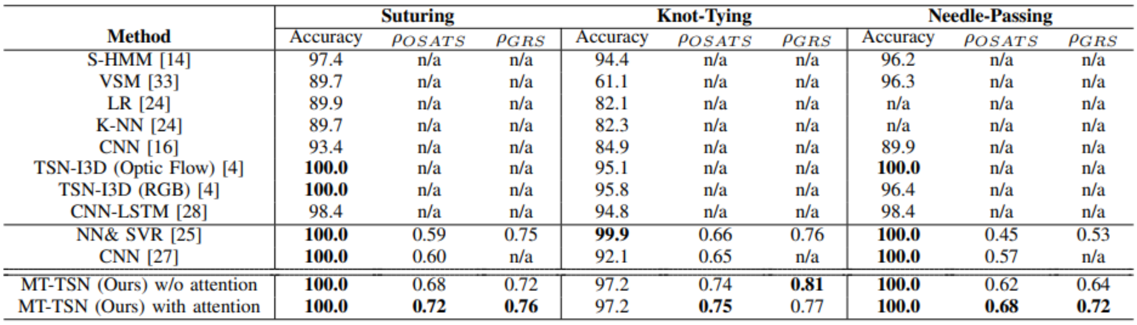
- The author argues that existing studies implement attention mechanisms in one single dimension, either spatial or temporal. This is insufficient to handle the heavy redundancies in long and cluttered surgical videos.
- Previous works conducted skill classification and OSATS score regression separately whereas the author proposes a joint approach.

Architecture:



Modified OSATS Attributes	Score
Respect for tissue	4.7798
Suture/needle handling	4.0190
Time and motion	4.5534
Flow of operation	4.1972
Overall performance	4.2957
Quality of final product	4.6892

Results:



Method	Suturing			Knot-Tying			Needle-Passing		
	Accuracy	ρ_{OSATS}	ρ_{GRS}	Accuracy	ρ_{OSATS}	ρ_{GRS}	Accuracy	ρ_{OSATS}	ρ_{GRS}
S-HMM [14]	97.4	n/a	n/a	94.4	n/a	n/a	96.2	n/a	n/a
VSM [33]	89.7	n/a	n/a	61.1	n/a	n/a	96.3	n/a	n/a
LR [24]	89.9	n/a	n/a	82.1	n/a	n/a	n/a	n/a	n/a
K-NN [24]	89.7	n/a	n/a	82.3	n/a	n/a	n/a	n/a	n/a
CNN [16]	93.4	n/a	n/a	84.9	n/a	n/a	89.9	n/a	n/a
TSN-I3D (Optic Flow) [4]	**100.0**	n/a	n/a	95.1	n/a	n/a	**100.0**	n/a	n/a
TSN-I3D (RGB) [4]	**100.0**	n/a	n/a	95.8	n/a	n/a	96.4	n/a	n/a
CNN-LSTM [28]	98.4	n/a	n/a	94.8	n/a	n/a	98.4	n/a	n/a
NN& SVR [25]	**100.0**	0.59	0.75	**99.9**	0.66	0.76	**100.0**	0.45	0.53
CNN [27]	**100.0**	0.60	n/a	92.1	0.65	n/a	**100.0**	0.57	n/a
MT-TSN (Ours) w/o attention	**100.0**	0.68	0.72	97.2	0.74	**0.81**	**100.0**	0.62	0.64
MT-TSN (Ours) with attention	**100.0**	**0.72**	**0.76**	97.2	**0.75**	0.77	**100.0**	**0.68**	**0.72**