2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES) | 978-1-6654-4499-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/INES52918.2021.9512917

Surgical Skill Assessment Automation Based on Sparse Optical Flow Data

Gábor Lajkó
EIT Digital Master School
Autonomous Systems track
Technische Universität Berlin
Berlin, Germany &
Eötvös Loránd University
Budapest, Hungary
gabor.lajko@
masterschool.eitdigital.eu

Renáta Nagyné Elek
Antal Bejczy Center for Intelligent
Robotics (IROB), EKIK &
Doctoral School of Applied Informatics &
John von Neumann Faculty of Informatics
Óbuda University
Budapest, Hungary
renata.elek@irob.uni-obuda.hu

Tamás Haidegger
Antal Bejczy Center for Intelligent
Robotics (IROB), EKIK &
John von Neumann Faculty of Informatics
Óbuda University
Budapest, Hungary
haidegger@irob.uni-obuda.hu

***Abstract*—Objective skill assessment based personal feedback is a vital part of surgical training. Automated assessment solutions aim to replace traditional manual (experts' opinion-based) assessment techniques, that predominantly requires the most valuable time commitment from senior surgeons. Typically, either kinematic or visual input data can be employed to perform skill assessment. Minimally Invasive Surgery (MIS) benefits the patients by using smaller incisions than open surgery, resulting in less pain and quicker recovery, but increasing the difficulty of the surgical task manyfold. Robot-Assisted Minimally Invasive Surgery (RAMIS) offers higher precision during surgery, while also improving the ergonomics for the performing surgeons. Kinematic data have been proven to directly correlate with the expertise of surgeons performing RAMIS procedures, but for traditional MIS it is not readily available. Visual feature-based solutions are slowly catching up to the efficacy of kinematics-based solutions, but the best performing methods usually depend on 3D visual features, which require stereo cameras and calibration data, neither of which are available in MIS. This paper introduces a general 2D image-based solution that can enable the creation and application of surgical skill assessment solutions in any training environment. A well-established kinematics-based skill assessment benchmark's feature extraction techniques have been repurposed to evaluate the accuracy that the generated data can produce. We reached individual accuracy up to 95.74% and mean accuracy – averaged over 5 cross-validation trials – up to 83.54%. Additional related resources such as the source codes, result and data files are publicly available on Github (https://github.com/ABC-iRobotics/VisDataSurgicalSkill).**

***Index Terms*—Robot-Assisted Minimally Invasive Surgery, Surgical Skill Assessment, JIGSAWS, Optical Flow**

I. Introduction

The introduction of Minimally Invasive Surgery (MIS) has revolutionized operations more than 50 years ago [1]. 25 years ago, with the introduction of robotic tele-surgical systems, a new form of MIS was born: the Robot-Assisted Minimally Invasive Surgery (RAMIS) [2], [3]. Since traditional surgical skill assessment techniques often require the expertise of skilled surgeons, whose time is a valuable an scarce resource, the need for automatic skill assessment techniques was given [4], [5].

MIS requires multiple years of training, but even though its benefits are beyond doubt: frequent skill assessment of the trainees is still not a part of the clinical practice [6]. Standard evaluation techniques include checking the time trainees take to perform tasks, having an expert surgeon overseeing the exams, or provide expert rating based on pre-defined criteria, such as GEARS [7] or OSATS [8].

Automatized skill assessment aims to free up expert evaluators by reliably automating the process of assessment as much as possible. It can classify the expertise level of surgeons-in-training, or provide personalised feedback, precise insight on how to improve certain subtasks. In principle, they indirectly support patient care by supporting the training of surgeons.

The problem at hand is that no reliable automatic skill assessment technique has been proven to be applicable to both MIS and RAMIS. Either kinematic or visual data is used primarily [9]. Kinematic solutions are more commonly used [9], but they cannot be used for the training of traditional, manual MIS. Given the popularity of Deep Learning methods in this field of research, the issue of overfitting and thereby the lack of generalisation-capability is also an ever-present issue, often combated by cross-validation, regularization, or transfer learning [10].

This article introduces a widely applicable, scalable and practical assessment data generation method, aiming to enable the integration of visual-based surgical skill assessment solutions into the curriculum of MIS trainings, by providing generally accessible 2D visual features.

A. JIGSAWS

Created by the cooperation of the Johns Hopkins University (JHU) and the creators of the Da Vinci Surgical System [11], Intuitive Surgical, Inc. (ISI) — the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [12] is one of the most widely-used surgical skill assessment datasets — as evidenced by [9] too. It contains kinematic and video data on three basic surgical tasks (knot-tying, suturing and needle-passing) — essential in surgical training curricula — and comes annotated

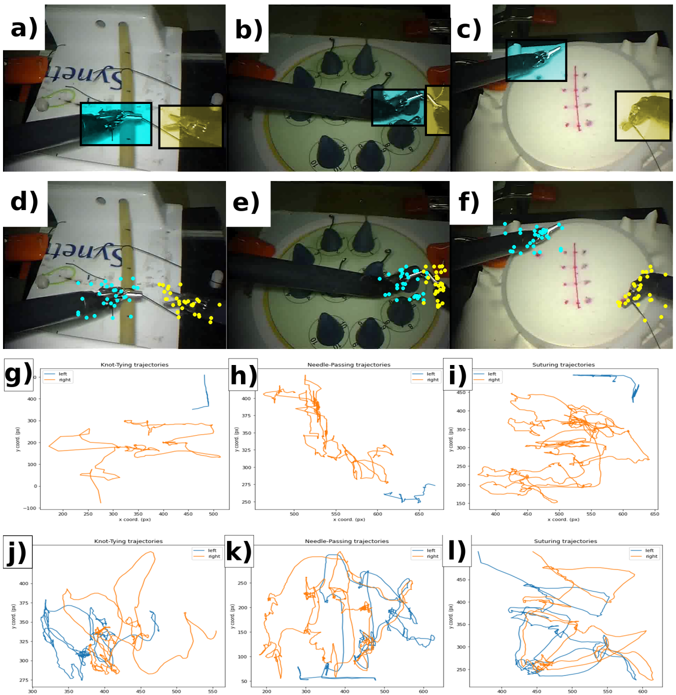

Fig. 1. The three surgical tasks: Knot-Tying (KT), Needle-Passing (NP) and Suturing (ST). The first row (a)-c)) illustrates the selection of Regions-Of-Interest (ROIs), the second shows the initial samples computed by the Shi-Tomasi method (d)-f)). The third an fourth row show one tracked points' trajectory from each tool (blue=left and orange=right). g) to i) the data is from a novice user, j) to l) it is an expert subject.

with expert rating scores by an experienced surgeon [13]. Figure 1 illustrates three steps of our workflow (selecting Regions-Of-Interest, saving initial samples, and tracking trajectories) on all three tasks.

B. State of the Art

The use of kinematic data predates visual data in the field of surgical skill assessment and offers very precise and useful assessment possibilities. In general kinematic data is the most suitable for surgical skill assessment in RAMIS procedures. Fawaz et al. has achieved 100% accuracy with both classification and regression tasks performed on the kinematics data of JIGSAWS, using a combination of Approximate Entropy and Fully Connected Convolutional Neural Networks [14], providing tangible, personalised feedback for surgeons in training. However kinematics cannot be considered superior to visuals anymore [15]. During manual MIS training, where assessment methods would be the most useful - there is no direct access to the kinematic data.

Ming et al. achieved a mean accuracy of 79.29% / 76.79%, 80.71% / 83.81% and 72.57% / 76.65% on the basis of Space Temporal Interest Points (STIP)/Improved Dense Trajectory (iDT) representation of the three subtasks of JIGSAWS respectively (see Table I). They found that although iDT produced better results, its use is way more memory-demanding, and therefore not practical. They proved that similarly to kinematic-based solutions, it is possible to distinguish novice and expert users based on visual features. Their solution relies on a histogram of features, processed by a support vector machine (SVM), to achieve classification.

C. Related Work

Frequency-based solutions such as the Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) are common [16] [17], as well as sequential motion texture (SMT), and approximate entropy (ApEn) [16].

Out of Deep Learning methods Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Fully-Connected Neural Networks (FCN) are the most frequently used, with other architectures succh as Long Short Term Memory (LSTM), VGG and Residual Neural Networks (ResNet) also being used for classification [9]. Out of CNN, RNN and FCN, CNN seems to be the superior technique, due to the lack of long sequences in surgical data, which recurrent networks need, and that the generally employed statistical features eliminate the use of temporal data that methods such as LSTM would require [9].

Visual-based methods are more practical in surgeon training, as kinematic data is only available in RAMIS(or with the use of external sensor system), while visual data is accessible in MIS as well, and they require no explicit preparation once they are trained and cross-validated [9].

II. Materials and Methods

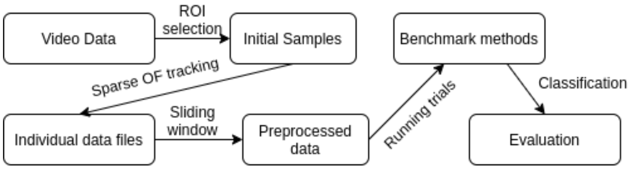

Fig. 2. The proposed workflow. First we compute and save the initial samples from each video input using user-selected Regions-Of-Interest, then by tracking them with sparse Optical Flow we process the videos, creating data files, that are further processed by a sliding window method, outputting the final input data. Using the benchmark methods of Anh et al. [18] we classify the users' expertise.

A. Optical Flow

Optical flow is a fundamental algorithm for movement detection in video analysis. It estimates motion between two consecutive video frames by calculating a shift vector to quantify the displacement difference [19]. The Lucas—Kanade method is commonly used to calculate the Optical Flow for a sparse feature set. The main idea of this method is based on a local motion constancy assumption, where nearby pixels have the same displacement direction. Our data generation method (see II-D) uses the Pyramidal Implementation of the Lucas Kanade algorithm [20].

B. Benchmark methods

For evaluation the codebase of Anh et al was used [18], [21]. Taking the kinematic data of the JIGSAWS set as input [12], they implemented 9 different evaluation methods for the classification of surgical skill into three categories: expert, intermediate and novice. To counteract overfitting, the LOSO (Leave-One-Super-Trial-Out) cross-validation technique was used for each of the benchmark methods introduced below.

1) Convolutional Neural Network (CNN): Commonly used for classification, segmentation and image processing CNNs ensure translation invariance and parameter sharing through convolution [22], [23]. They are based on the assumption that nearby data points are more closely correlated than further ones. In our data (see II-D) the two surgical tools are separated, therefore data-points closer to each other correlate more, as they are more likely to belong to the same tool.The local dependencies CNN relies on are further supported by the format, that Optical Flow and Position data are side-by-side, and the rows correspond to timestamps.

2) Long Short Term Memory (LSTM): LSTMs are specialised Neural Networks by design ideal for time series data analysis [24]. Using the combination of three types of gates (input, output and forget) and a dedicated memory cell – storing the internal representation of the learned information – they are able to record long term data representations. Using time series data and the assumption that the movement of surgical tools have detectable patterns, that can correlate with surgical expertise, LSTM-based solutions are suitable for evaluation of expertise.

3) CNN + LSTM: A straightforward combination of the previous two models. Two convolutional layers - batch optimization and ReLu activation functions, with same padding - followed by two LSTM blocks. The temporal information of the convolutional layers' outputs is processed by the LSTM blocks, in order to learn contextually, but from an already processed information source. Among others, Li et al. [25] have demonstrated high accuracy predictions using the combination of these methods.

4) Residual Neural Network (ResNet): Primarily used for classification tasks, using so-called skip connections as shortcuts to solve the degradation problem [26] — essentially short circuiting shallow layers to deep layers — it enables the creation of deeper networks without loss of performance. It is a reliable technique even within smaller networks. The model of the benchmark also consists of only 3 blocks, meaning that it does not utilize its strength to the fullest, but it still can perform accurate classification.

5) Convolutional Autoencoder (convAuto): Traditionally used for data compression, dimensionality reduction and the denoising of data without significant information loss, an autoencoder — often symmetric, built up of two blocks: an encoder and a decoder — is an unsupervised machine learning model [27]. It aims to create a copy of its input as an output, reverse engineering the problem by trying to find the right filter. It compresses the input time series into a latent space representation, then the network tries restructuring it into the original input data in the decoder. For classification, after the network has been fully trained, the encoder's output is fed to an SVM classifier.

6) DFT & DCT: The Discrete Fourier Transformation and the Discrete Cosine Transformation are traditionally used to transform time series data from the time domain to the frequency domain. The use of frequency features in surgical skill assessment has been proven to perform well by Zia et al. [17].

C. Setup

The evaluation of results has been done in a 64 bit Ubuntu 18.04.5 LTS environment, using jupyter-notebooks with Python 3.6.9, and the following packages: opencv-python 4.2.0.34, scikit-learn 0.23.2 and Keras 2.1.5. The computer's CPU is a twelve core Intel® Core™ i7-8750H CPU, running on 2.20GHz with 8 GB RAM and a GeForce GTX 1050 Ti/PCIe/SSE2 GPU.

D. Data generation

The Kanade—Lucas Optical Flow requires the output of the Shi—Tomasi Corner Detection [28] — published in 1994 — which is built on the Harris corner detection [29] from 1988, with the additional criteria of filtering for "good features to track", and since then, became one of the most widely used corner detection methods over the world.

First a suitable initial frame is selected, where both surgical tools are visible, then user-selected Regions-Of-Interest (ROIs) are preprocessed and saved. The frames are first turned grayscale, then blurred by a median filter, and run through a binary filter using adaptive thresholds, in order to denoise the frames and enable the better detection of features, using the Shi—Tomasi detector on the respective ROIs. If the number of non-zero samples is not the maximum possible amount: a warning comes up and the ROI selection has to be repeated. This is performed for both tools separately, which ensures that the initial corner points are fully sampled.

The resulting output constitutes the initial features to be tracked. Each video is traversed, set to the initial frame, the features of which are then tracked by the Kanade—Lucas OF. The features are extracted from each frame of the video, and collected in a list, with the dimensions: $frame\_number \times sample\_size \times 2 \times 2$. This list is then iterated through, and for each frame's data a row of 240 features - made up of the respective Optical Flow and Positional information of each tool - is added to the output.

Given the ROI data and the generated output, the final output needs to be created by grouping the data according to the surgical tasks and the expertise level of users. This is accomplished using the sliding window preprocessing method implemented by Anh et al. for their benchmark [18], to process the multivariate time series and separate chunks of the data into uniformly sized local windows, thereby enabling the evaluation of our data with the same networks originally designed for the kinematic data of JIGSAWS.

III. Results

Cross-validation is a strong tool against overfitting. To evalute methods more precisely, it is customary to take the average of multiple cross-validated executions. As chosen by Anh et al. [18] each method is run 5 times for each generated input file. Within each run there are 5 trials, using the LOSO cross-validation method, then the mean accuracy is calculated.

The intermediate class is prone to misclassification [9]. Funke et al's hybrid 3D network has misclassified every intermediate surgeon into either expert or novice, using optical flow data for knot tying [30]. This may partially be due to data disparity, as intermediate and expert subjects are underrepresented in comparison to novices [12]. Funke's solution relied on a Leave-One-User-Out (LOUO) cross-validation, which is less robust to such disparity. Anh et al.'s benchmark also faced issues with the classification of intermediate users [18]. Given this, we have decided to omit intermediate subjects from the main evaluation, and only comparatively analysed 3-class classification, the result of which is presented in Table III.

A. *Results by methods*

The following results have been obtained without the data of intermediate subjects.

1) CNN: With a minimum standard deviation of 1.43% and best mean accuracy of 80.72%, CNN has responded well to our data.

2) LSTM: Although its best mean accuracy of 80.44% is promising, its standard deviation ranging from 5.97% to 15.49%, as well as a high number of trials with zero true positives for experts show that LSTM by itself is not complex enough to fit to our data.

3) CNN+LSTM: Able to counteract the shortcomings of the simple LSTM, it improved on the results of both models, with 83.19% maximum mean accuracy and consistently less than 2% standard deviation. Its highest mean accuracy was 79.19%, even outperforming the more complex ResNet model (78.65%).

4) ResNet: ResNet modelled our data the best, with 84.23% highest mean accuracy and 1.36% standard deviation. Having the highest number of layers, it suggests that our data scales in performance with model complexity.

5) convAuto: The only unsupervised method in the benchmark, convAuto reached a maximum of 79.77%, with an average of 5% standard deviation. Its highest mean of 67.78% is among the lower ones, but given that it relies on the SVM classifier, its efficacy could be improved by tuning its hyperparameters, or employing a kernel trick.

B. *Results by surgical tasks*

1) Knot-Tying: Ming et al. found Knot-Tying to be the easiest surgical task to assess with both STIP and iDT [31]. Our method is similar in principle to their STIP method, as it also tracks the movement of interest points/features over time.

When it comes to the Model evaluation, even the worst average accuracy (produced by LSTM) was at 74.75%. Regarding SVM evaluation the same value was 62.1%. Coincidentally the highest SVM evaluated accuracy (80.43%) and the highest average SVM accuracy (76.07%) were also by LSTM. Given that model-evaluated LSTM results had many outliers, SVM could improve the recall and precision. For the best performing configuration of LSTM, 8 out of 25 trials (32%) had 0 expert true positives. To measure the efficacy specifically for expert classification, the Recall metric needs to be used.

$$ExpertRecall = \frac{ExpertTruePositive}{NumOfExpertsInTrial} \tag{1}$$

The highest individual expert recall of LSTM was 89.7%, but given the 8 cases where the value is 0, its mean is 24.36%.

Overall ResNet, CNN+LSTM and convAuto all performed well for the Knot-Tying task. LSTM has done well, but its low mean expert recall leaves the need for further investigation before it could be deemed as reliable, and CNN - considered to be a top-performer [9] - has fallen behind. Its highest accuracy was 80.42%, and highest mean accuracy 78.38%. SVM predictions dropped its accuracy below 70%. ResNet's model-based evaluation resulted in the highest accuracy (83.54%), while the highest mean accuracy (80.11%) came from the model-predictions of CNN+LSTM. Figure 3 shows the ranges of accuracies for each used method with Knot-Tying skill data.

2) Suturing: The same observations apply to Suturing as to Knot-Tying: ResNet, CNN+LSTM and convAuto performed well, LSTM has shown high results in some cases, accompanied by confusion matrix anomalies (the maximum Expert Recall only being 3.33%, with an overall average of 0.22%), and CNN seemingly performed the worst, even though it still did so above 75% on Model average accuracies. Even though Ming et al. [31] also found that Suturing is harder to classify than Knot-Tying, CNN has been found to be one of the most reliable methods by the review of Yanik et al. [9]. It is possible that the CNN model of the benchmark is too small, and it would perform better with higher complexity and more layers. Figure 4 illustrates the accuracy of each applied method, given the Suturing task.

3) Needle-Passing: Ming et al. claimed that Needle-Passing was the hardest skill to perform classification for, because they did not find significant differences between the trajectories of expert and novice users' left hand movements [31]. With our data generation method only LSTM dropped significantly in efficacy in comparison with its performance on the other skills.

The highest Model average precision (79.74%) and the highest SVM accuracy (71.58%) were both achieved by CNN+LSTM, making it the overall best for the skill of Needle-Passing. The range of accuracies given all the Needle-Passing data for each method is illustrated in Figure 5.

C. *Performance analysis*

Our generated data combined with the benchmark of Anh et al. [18] has successfully outperformed the solutions of Ming et al. [31] in both Suturing and Knot-Tying, and only slightly fell short of their results in Needle-Passing. Our goal was to find a method that can achieve similar results to the state of the art in the field of surgical skill assessment, while keeping generalisation and practicality in mind, in order to keep it relevant for manual MIS training, where 3D and pose information is not available. Table I presents the detailed comparison of these methods.

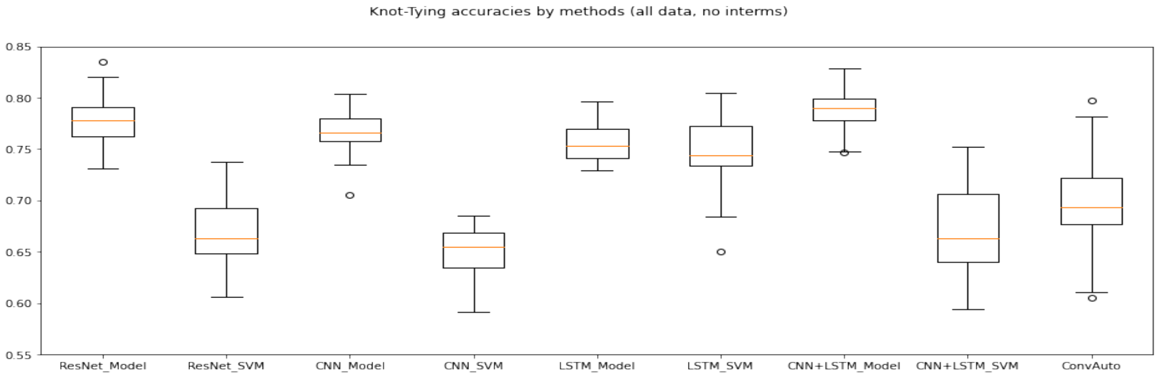

Fig. 3. Knot-Tying accuracies without intermediates. The best perfoming methods were: ResNet and CNN+LSTM.

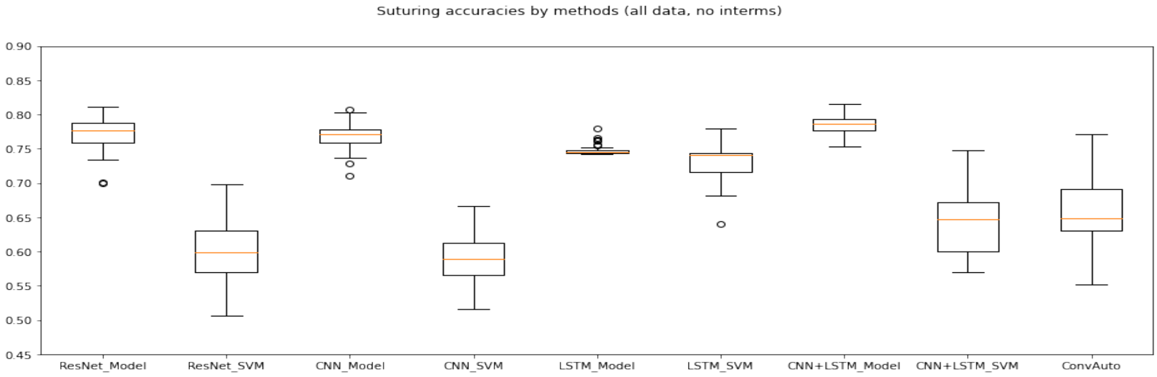

Fig. 4. Suturing accuracies without intermediates. The best performing methods are ResNet, CNN and CNN+LSTM.

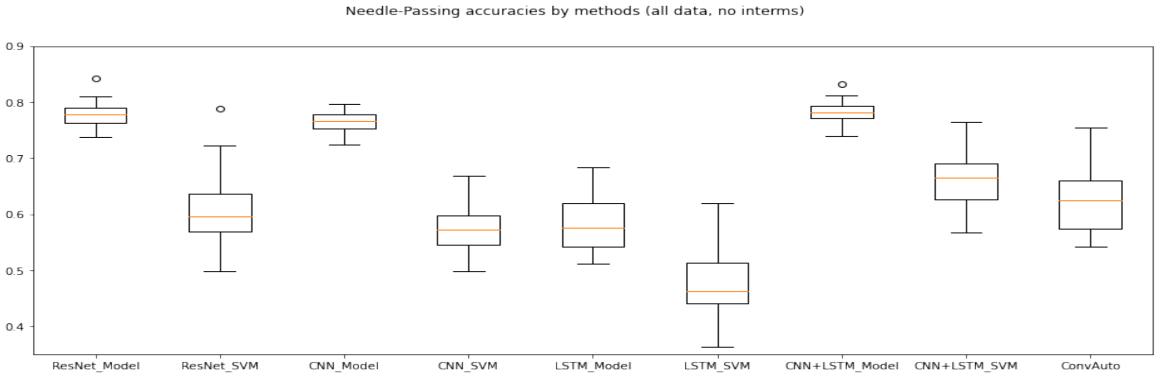

Fig. 5. Needle-Passing accuracies without intermediates. ResNet, CNN and CNN+LSTM outperform the other methods.

Residual Neural Network performed the best, closely followed by the combined model of CNN and LSTM. DFT and DCT have performed too poorly to be able to rely on them for classification. Even though LSTM's accuracy is high, its confusion matrices showed it to be unreliable. It has been observed, that although the SVM classification decreases the overall accuracy, it performs more consistently well overall. This is likely due to the fact that SVMs are suspectible to outliers [32]. They perform well in high-dimensional feature spaces, and although we have 240 features, the number of rows greatly outnumber this.

IV. Conclusion & Discussion

A. *Kinematics vs. Visual*

Kinematic data is highly precise, describes the motions of each relevant joint with the combination of rotation matrices, linear and angular velocities. However, visual solutions have the potential to replace kinematic ones due to the fact that theoretically it is possible to calculate the approximation of the same kinematic data from visual input [33].

Visual methods are inherently restricted to two dimensions. This can be counteracted by 3D visual methods, but they

TABLE I
The comparison of our results to that of the state of the art. Every method listed here used the JIGSAWS dataset as visual input data source.

Author (Year)	Method	ST	NP	KT
Funke et al. (2019)	3D ConvNet +TSN [30]	100%	96.4%	95.8%
Ming et al. (2021)	STIP [31]	79.29%	87.01%	72.57%
Ming et al. (2021)	iDT [31]	76.79%	83.81%	76.65%
Our solution	CNN [18]	80.72%	79.66%	80.41%
Our solution	CNN+LSTM [18]	81.58%	83.19%	82.82%
Our solution	ResNet [18]	81.89%	84.23%	83.54%

TABLE III
The performance overview of each Deep Learning methods in case of 3 and 2 classes respectively. Eval. stands for the evaluation type, meaning either directly Model-based predictions or SVM classifier. High and Low Avg. are the best and worst mean accuracies, given the 5 trials in each run, while Abs. High and Low are from the accuracies of individual trials.

Method	Eval.	class num.	Low. Avg.	High. Avg.	Abs. Low.	Abs. High.
CNN	Model	2	73.699%	79.12%	54.34%	93.67%
		3	49.595%	66.83%	35.69%	75.44%
CNN	SVM	2	53.97%	66.53%	29.45%	94.68%
		3	28.15%	56.797%	43.65%	70.59%
LSTM	Model	2	51.16%	79.62%	24.44%	90.64%
		3	40.48%	62.19%	15.25%	69.63%
LSTM	SVM	2	40.95%	80.44%	22.22%	91.94%
		3	30.198%	60.87%	15.25%	70.86%
CNN+ LSTM	Model	2	74.69%	83.19%	56.42%	93.65%
		3	56.12%	73.09%	46.37%	82.65%
CNN+ LSTM	SVM	2	58.987%	73.44%	28.27%	93.23%
		3	33.12%	68.57%	5.15%	82.94%
ResNet	Model	2	73.75%	83.54%	54.55%	95.74%
		3	47.93%	70.25%	33.46%	80%
ResNet	SVM	2	54.21%	73.64%	23.3%	93.04%
		3	25.92%	61.496%	7.15%	79.17%
convAuto	SVM	2	58.58%	75.52%	33.33%	93.21%
		3	30.82%	52.57%	16.14%	77.36%

require camera calibration information and stereo recording, which are not available during traditional MIS trainings, where skill assessment is of the highest importance. However, given their accessibility in RAMIS skill assessment, and their higher performance, their application is justified. In 2019 Funke et al. [30] have achieved close to 100% accuracy on all three surgical skills in the JIGSAWS set, using 3D visual features, something previously only possible with kinematic solutions. For the goals we have set for ourselves, however, 3D approaches are not feasible.

B. Intermediate users

Intermediate users are prone to misclassification, to the point of not being able to identify any of them correctly [30]. This supposedly stems from the uneven distribution of the dataset, resulting in the DNN failing to generalize the features. Table II shows the corresponding analysis of 3-class classification confusion matrices.

TABLE II
The Recall of Intermediate users with each method

Method	# of zero true positive cases	Recall
ResNet	42 (out of 150)	28%
CNN	47 (out of 150)	31.33%
CNN+LSTM	55 (out of 150)	36.66%
LSTM	117 (out of 150)	78%
convAuto	12 (out of 75)	16%

LSTM has struggled with the classification accuracy even in the 2 class version. With 78% of times not being able identify a single intermediate user, it is the worst performing method.

The Convolutional Autoencoder performed the best, with the rest of the methods forming a relatively balanced middle range. Every analyzed method other than LSTM could classify intermediate users with at least 66% accuracy. Table III shows the highest and lowest mean accuracy of each investigated method.

C. Conclusion

This paper aimed to create a practical, generally applicable solution that can enable the creation of visual-based surgical skill assessment methods, and can potentially lead to the inclusion of automated skill assessment in the curriculum of minimally invasive surgical training, introducing benefits such as objectivity, reproducibility, and the fact that it won't require human expertise.

The proposed method has measured up to the state of the art in 2D visual-based skill assessment, with more than 80% accuracy for all three surgical subtasks available in JIGSAWS (knot-tying, suturing and needle-passing). By introducing new visual features - such as image-based orientation and image-based collision detection - or from the evaluation side: utilizing other SVM kernel methods, tuning the hyperparameters, or using boosted tree algorithms instead, classification accuracy can be further improved.

We have shown the potential use of optical flow as an input for RAMIS skill assessment, highlighting the maximum accuracy achievable with this data by evaluating it with an established skill assessment benchmark, by evaluating its methods independently. The highest performing method, the Residual Neural Network reached 81.89%, 84.23% and 83.54% accuracy for the skills of Suturing, Needle-Passing and Knot-Tying respectively.

Acknowledgment

This research was performed as part of project no. 2019-1.3.1-KK-2019-00007 that has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme. T. Haidegger and R. Nagyné Elek are supported through the New National Excellence Program of the Ministry of Human Capacities. T. Haidegger is a Bolyai Fellow of the Hungarian Academy of Sciences. G. Lajkó holds a scholarship from the European Institute of Innovation & Technology (EIT) for his master studies in Autonomous Systems.

REFERENCES

[1] B. Radojcić, R. Jokić, S. Grebeldinger, I. Meljnikov, and N. Radojić, "[History of minimally invasive surgery]," *Medicinski Pregled*, vol. 62, no. 11-12, pp. 597–602, Dec. 2009.

[2] L. Weinberg, S. Rao, and P. F. Escobar, "Robotic Surgery in Gynecology: An Updated Systematic Review," *Obstetrics and Gynecology International*, vol. 2011, pp. 1–29, 2011. [Online]. Available: http://www.hindawi.com/journals/ogi/2011/852061/

[3] D. Serikbayev East Kazakhstan State Technical University and G. K. Shadrin, "Application of Compensation Algorithms to Control the Movement of a Robot Manipulator," *Acta Polytechnica Hungarica*, vol. 17, no. 1, pp. 191–214, 2020. [Online]. Available: http://uni-obuda.hu/journal/Shadrin_Alontseva_Kussaiyn-Murat_Kadyroldina_Ospanov_Haidegger_98.pdf

[4] R. Nagyné Elek and T. Haidegger, "Non-Technical Skill Assessment and Mental Load Evaluation in Robot-Assisted Minimally Invasive Surgery," *Sensors*, vol. 21, no. 8, p. 2666, Apr. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/8/2666

[5] T. Haidegger, "Autonomy for Surgical Robots: Concepts and Paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 2, pp. 65–76, May 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8698847/

[6] T. H. Renáta Elek, "Robot-Assisted Minimally Invasive Surgical Skill Assessment—Manual and Automated Platforms," *Acta Polytechnica Hungarica*, vol. 16, no. 8, Sep. 2019. [Online]. Available: http://uni-obuda.hu/journal/Nagyne-Elek_Haidegger_95.pdf

[7] "Global Evaluative Assessment of Robotic Skills (GEARS)." [Online]. Available: https://www.csats.com/gears

[8] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents: OBJECTIVE STRUCTURED ASSESSMENT OF TECHNICAL SKILL," *British Journal of Surgery*, vol. 84, no. 2, pp. 273–278, Feb. 1997. [Online]. Available: http://doi.wiley.com/10.1046/j.1365-2168.1997.02502.x

[9] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Miller, B. Van Voorst, B. Makled, J. Norfleet, and S. De, "Deep Neural Networks for the Assessment of Surgical Skills: A Systematic Review," Mar. 2021, arXiv: 2103.05113. [Online]. Available: http://arxiv.org/abs/2103.05113

[10] D. Zhang, Z. Wu, J. Chen, A. Gao, X. Chen, P. Li, Z. Wang, G. Yang, B. Lo, and G.-Z. Yang, "Automatic Microsurgical Skill Assessment Based on Cross-Domain Transfer Learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4148–4155, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9072568/

[11] "da Vinci Surgical System, Intuitive Surgical, Inc." [Online]. Available: https://www.davincisurgery.com/

[12] A. K. Lefor, K. Harada, A. Dosis, and M. Mitsuishi, "Motion analysis of the JHU-ISI Gesture and Skill Assessment Working Set using Robotics Video and Motion Assessment Software," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 12, pp. 2017–2025, Dec. 2020. [Online]. Available: http://link.springer.com/10.1007/s11548-020-02259-z

[13] D. El-Saig, R. N. Elek, and T. Haidegger, "A Graphical Tool for Parsing and Inspecting Surgical Robotic Datasets," in *2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI)*. Budapest, Hungary: IEEE, Nov. 2018, pp. 000 131–000 136. [Online]. Available: https://ieeexplore.ieee.org/document/8928222/

[14] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 9, pp. 1611–1617, Sep. 2019. [Online]. Available: http://link.springer.com/10.1007/s11548-019-02039-4

[15] D. Lee, H. W. Yu, H. Kwon, H.-J. Kong, K. E. Lee, and H. C. Kim, "Evaluation of Surgical Skills during Robotic Surgery by Deep Learning-Based Multiple Surgical Instrument Tracking in Training and Actual Operations," *Journal of Clinical Medicine*, vol. 9, no. 6, p. 1964, Jun. 2020. [Online]. Available: https://www.mdpi.com/2077-0383/9/6/1964

[16] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 5, pp. 731–739, May 2018. [Online]. Available: http://link.springer.com/10.1007/s11548-018-1735-5

[17] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, T. Ploetz, M. A. Clements, and I. Essa, "Automated video-based assessment of surgical skills for training and evaluation in medical schools," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 9, pp. 1623–1636, Sep. 2016. [Online]. Available: http://link.springer.com/10.1007/s11548-016-1468-2

[18] N. X. Anh, R. M. Nataraja, and S. Chauhan, "Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques," *Computer Methods and Programs in Biomedicine*, vol. 187, p. 105234, Apr. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260719313586

[19] A. I. Karoly, R. N. Elek, T. Haidegger, K. Szell, and P. Galambos, "Optical flow-based segmentation of moving objects for mobile robot navigation using pre-trained deep learning models*," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. Bari, Italy: IEEE, Oct. 2019, pp. 3080–3086. [Online]. Available: https://ieeexplore.ieee.org/document/8914359/

[20] J. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," 1999.

[21] S. C. Nguyen Xuan Anh, Ramesh Mark Nataraja, "Feature extraction benchmark repository," https://github.com/SimonNgj/compssa, 2019.

[22] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018. [Online]. Available: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9

[23] N. Yusupova, G. Shakhmametova, and R. Zulkarneev, "Complex analysis of medical data with data mining usage," *Acta Polytechnica Hungarica*, vol. 17, pp. 75–93, 2020.

[24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: https://direct.mit.edu/neco/article/9/8/1735-1780/6109

[25] F. Li, K. Shirahama, M. Nisar, L. Köping, and M. Grzegorzek, "Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors," *Sensors*, vol. 18, no. 3, p. 679, Feb. 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/2/679

[26] R. P. Monti, S. Tootoonian, and R. Cao, "Avoiding Degradation in Deep Feed-Forward Networks by Phasing Out Skip-Connections," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds. Cham: Springer International Publishing, 2018, vol. 11141, pp. 447–456, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-01424-7_44

[27] Q. V. Le, G. Brain, and G. Inc, "A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks," 2015.

[28] Jianbo Shi and Tomasi, "Good features to track," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*. Seattle, WA, USA: IEEE Comput. Soc. Press, 1994, pp. 593–600. [Online]. Available: http://ieeexplore.ieee.org/document/323794/

[29] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Procedings of the Alvey Vision Conference 1988*. Manchester: Alvey Vision Club, 1988, pp. 23.1–23.6. [Online]. Available: http://www.bmva.org/bmvc/1988/avc-88-023.html

[30] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1217–1225, Jul. 2019. [Online]. Available: http://link.springer.com/10.1007/s11548-019-01995-1

[31] Y. Ming, Y. Cheng, Y. Jing, L. Liangzhe, Y. Pengcheng, Z. Guang, and C. Feng, "Surgical skills assessment from robot assisted surgery video data," in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*. Shenyang, China: IEEE, Jan. 2021, pp. 392–396. [Online]. Available: https://ieeexplore.ieee.org/document/9362525/

[32] T. Kanamori, S. Fujiwara, and A. Takeda, "Breakdown Point of Robust Support Vector Machines," *Entropy*, vol. 19, no. 2, p. 83, Feb. 2017. [Online]. Available: http://www.mdpi.com/1099-4300/19/2/83

[33] M. K. Hasan, L. Calvet, N. Rabbani, and A. Bartoli, "Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry," *Medical Image Analysis*, vol. 70, p. 101994, May 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841521000402