sensors

MDPI

*Article*

# An Automated Skill Assessment Framework Based on Visual Motion Signals and A Deep Neural Network in Robot-Assisted Minimally Invasive Surgery

**Mingzhang Pan [1], Shuo Wang [1], Jing Li [1], Xiuze Yang [1] and Ke Liang [1,2,*]**

[1] College of Mechanical Engineering, Guangxi University, Guangxi 530004, China

[2] Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology, School of Mechanical Engineering, Guangxi University, Guangxi 530004, China

* Correspondence: 20200035@gxu.edu.cn

**Abstract:** The surgical skill assessment can quantify the quality of the surgical operation by the motion state of the surgical instrument tip (SIT), which is considered one of the effective primary means to improve the accuracy of surgical operation. Traditional methods have displayed promising results in skill assessment. However, this success is predicated on the SIT sensors, making these approaches impractical in the minimally invasive surgical robot with such a tiny end size. To address the assessment issue of operation quality for Robot-Assisted Minimally Invasive Surgery (RAMIS), this paper proposed a new automatic assessment Framework of surgical skills based on visual motion tracking and deep learning. The new method innovatively combines vision and kinematics. The Kernel Correlation Filter (KCF) is introduced to get the key motion signals of SIT and classify by using the Residual Neural Network (ResNet), realizing automated skill assessment in RAMIS. To verify its effectiveness and accuracy, the proposed method is applied to the public minimally invasive surgical robot dataset, the JIGSAWS. The results show the method based on visual motion tracking technology and a deep neural network model can effectively and accurately evaluate the skill of robot-assisted surgery in near real-time. In a fairly short computational processing time of 3 to 5 seconds, the average accuracy of the assessment method is 92.04% and 84.80% in distinguishing two and three skill levels. This study makes an important contribution to the safe and high-quality development of RAMIS.

**Keywords:** robot-assisted minimally invasive surgery; surgical skill assessment; visual motion tracking; kernel correlation filter; residual neural network

**Citation:** To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

## 1. Introduction

Recent years have witnessed the remarkable processes of the RAMIS in general surgery, gastrointestinal surgery, urology, and gynecology due to the advantages of 3D vision, motion scaling, and tremor filtering [1, 2]. The quality of these surgeries is related to the skill of the operation, which will significantly affect the patient's health and safety [3, 4]. Therefore, surgeons must reach the needed surgical operation skills before surgery. Improving surgical skills largely depends on accurate skill evaluation methods [5]. So, the surgical skill assessment methods research is important in RAMIS.

Most studies are carried out the attention to analyzing the motion signals of SIT. Farcas et al. [6] used a traditional laparoscopic box trainer to install a customized motion tracking system to analyze and study the instrument motion at the stage of suture task in vivo determined in the simulator, providing an assessment of velocity and acceleration. One purpose of these simulators is to reduce the subjective reliance on experts and observers when evaluating performance or technical skills [7]. In surgical skill training, key motion signals of SIT have provided objective and accurate skill assessment [8]. Therefore, getting key motion signals of SIT has important research significance. Jiang et al. [9] analyzed the key motion features, such as the SIT's trajectory, and distinguished operators' motion control skills with different skill levels based on the Dynamic Time Warping (DTW) algorithm. Oquendo et al. [10] designed a magnetic induction motion tracking system and algorithm. The algorithm can automatically track the suture trajectory to evaluate the suture skills of the trainees in pediatric laparoscopy. However, introducing these sensors, data gloves, and other extra tools [11] dramatically reduces training efficiency and increases the burden and cost of surgical skill assessment. In addition, the software-based motion tracking system has also been used for surgical skill assessment. However, these methods suffer low tracking accuracy [12, 13]. Overall, although the above assessment methods prove that the motion tracking system is effective in surgical skill evaluation, it is currently difficult to apply to the training and assessment in RAMIS due to the problems such as low efficiency and poor accuracy. Therefore, it is urgent to study a genuinely efficient and accurate assessment method suitable for RAMIS.

Based on the above needs, the automatic assessment of surgical skills using deep-learning neural networks has become a hot research topic. The application of deep neural networks needs to be based on the data sets. Thus, many scholars have studied the RAMIS surgical skill assessment dataset. Rivas-Blanco et al. [14] explored the dataset that could be used to automate surgical robotic tasks, surgical skill assessment, and gesture recognition. In addition, the JIGSAWS [15] is one of the most widely available datasets for technical skill assessment in surgical robots. These large amounts of data can promote the development of surgical robot skill assessment toward automation. Kitaguchi et al. [16] proposed a deep learning method based on a Convolutional Neural Network (CNN). It can achieve high-precision automatic recognition of surgical actions with an accuracy rate of 91.9%. The Long Short-Term Memory (LSTM) model [17] and a symmetric dilated convolutional neural network model, SD-Net [18], have also been used for the automatic assessment of surgical skills. Nguyen et al. [19] described an automated assessment system using a CNN-LSTM neural network model and IMU sensors. This model performed classification and regression tasks for kinematic data in JIGSAWS, achieving over 95% accuracy. Wang et al. [20] proposed an analytical deep learning framework for surgical training skill assessment based on sensor data and CNN, implementing deep convolutional neural networks to map multivariate time series data of kinematics to individual skill levels. Although their research has achieved promising results, the experiments are based on existing datasets or sensor data. This was valuable for laboratory research, but it is still a long way from being practically applied to RAMIS. Our aim is therefore to develop a broadly applicable, scalable evaluation method that can be easily integrated into surgical robots.

In RAMIS, the endoscope can provide visual field information, which we believe can play a role in surgical skill assessment. If vision can be used instead of sensors to obtain signals, no additional equipment is required, which meets the needs of practical applications. The motion signals are used as the input features of neural networks in RAMIS skill assessment [21]. Traditional kinematic data are no longer superior to visual data completely in surgical skill assessment [22]. Funke et al. [23] achieved nearly 100% classification accuracy using 3D visual features. Evaluation methods based on 3D visual features tend to outperform 2D methods, but they have limited utility and are not suitable for RAMIS training. To help integrate automated skill assessment into surgical training practice, our proposed solution, therefore, relies on 2D visual features. Ming et al. [24] obtained over 70% accuracy based on 2D videos in surgical skill assessment, which represent the motion dynamics via improved dense trajectory (IDT) features and space temporal interest points (STIP). Lajkó et al. [25] demonstrated the potential application of optical flow for skill assessment based on the 2D vision in RAMIS and achieved an assessment accuracy of over 80%. The accuracy of 2D vision is not as good as that of 3D vision, but it is lower training costs and more efficient to apply to the automatic skill assessment in

RAMIS. So, this paper studies the intuitive and efficient assessment method using endoscopic 2D visual motion signals in RAMIS.

Based on the above problems for surgical skill assessment in RAMIS, this study proposed a new automated surgical skill assessment framework based on visual motion tracking technology, and a deep neural network model is proposed. It can be applied to real-time stage identification and online assessment. The new method studied a KCF algorithm [26] that can realize the motion tracking of SIT. It establishes key motion signal features in the video. Meanwhile, the method shows a ResNet [27] model. It uses the visual motion signals as input to improve the classification efficiency of surgical skills and realize the efficient assessment of surgical skills. Also, this method effectively considers the advantages of visual efficiency and accuracy of motion signals, improving the assessment accuracy of surgical skills. Finally, the JIGSAWS is used to corroborate the effectiveness of the proposed method. The result shows the classification of this method is better than that of other models. In this paper, a practical framework is provided for the automatic online assessment of objective skills in RAMIS.

To sum up, the innovations and contributions of this paper are as follows:

- A novel end-to-end analytical framework with visual tracking and deep learning for skill assessment based on high-level analysis of surgical motion.
- Visual technology is used to replace traditional sensors to obtain motion signals in RAMIS.
- The proposed model was verified in the JIGSAWS dataset and exploration of validation schemes applicable for the development of surgical skills assessment in RAMIS.

## 2. Materials and Methods

The surgical skill assessment framework based on visual tracking and deep learning in RAMIS is shown in Figure 1. The endoscope at the end of the surgical robot is used to provide visual information, and the required motion signals of SIT is recorded by the KCF, which is a multivariate time series (MTS) including [$x$, $y$, $t$, $v$, $a$, $MJ$] (section 3.2.2). The recorded MTS is input into ResNet for classification. It outputs a discriminative assessment of surgical skills through a deep learning architecture, finally, the results are fed back to the operator. This chapter introduces the principles of the relevant models in detail.

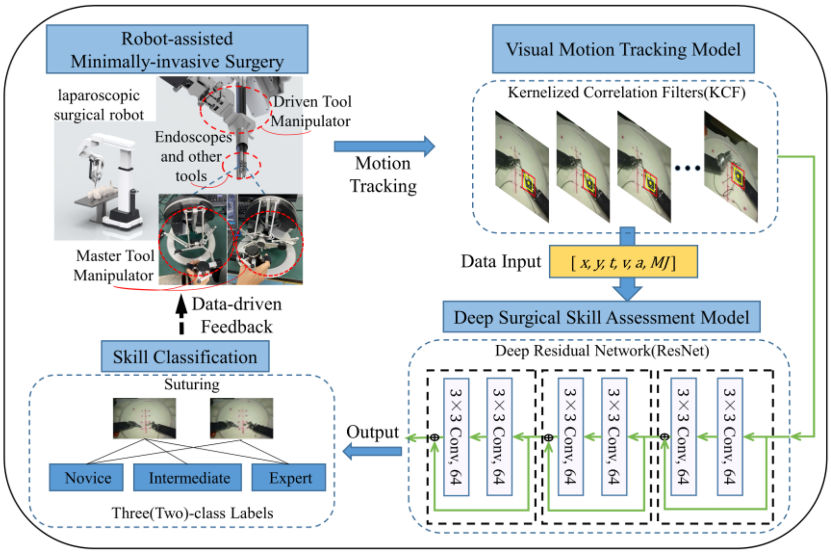

**Figure 1.** A framework for RAMIS based on visual motion tracking and deep learning neural network.

### 2.1. KCF

The core part of most current trackers is a classifier, whose task is distinguishing the goals from the surroundings. In this study, the tracking model needs to accurately identify the SIT and capture their movements from the surroundings. The SIT moves at a relatively high speed when doctors perform surgical tasks, which is a great challenge for the tracking models.

The KCF is a high-speed and accurate motion-tracking algorithm, which has proven to be a very accurate tracking tool [28]. It is a kernel-based ridge regression classifier [29] that uses the cyclic matrix gained by cyclic displacement to collect positive and negative samples. The matrix operation is transformed into the point multiplication of the elements by using the diagonalization property of the cyclic matrix in the Fourier domain. The efficiency of calculation is improved. Meanwhile, the multi-channel Histogram of Oriented Gradient (HOG) replaces the single-channel gray features and extends to multi-channel linear space to achieve higher robustness and accuracy.

As shown in Figure 2, the KCF mainly includes two stages, training, and detection. In this study, the Spatio-Temporal Context model [30] is referred to learn about this framework. In the training stage, the features of the target region are extracted. Then, the kernel function is used to calculate the generation vector of the kernel matrix of the current regional features.

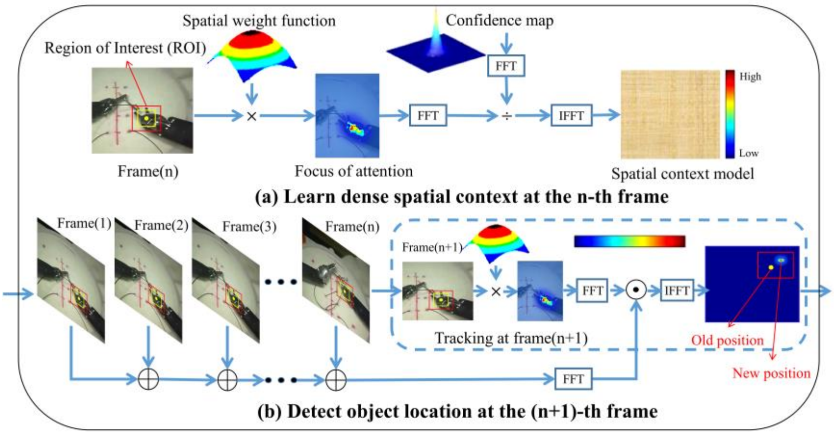

**Figure 2.** The Basic framework of the kernelized correlation filter algorithm in this study: (a) is the training stage, and (b) is the detection stage. FFT means the Fast Fourier Transform, and IFFT means the inverse FFT.

The KCF uses the multi-channel HOG features, which need to add vectors of different channel features. Taking the Gaussian kernel function as an example, as shown in Equation (1):

$$k^{xx'} = exp(-\frac{1}{\sigma^2}(\|x\|^2 + \|x'\|^2 - 2f^{-1}(\sum_{c} \hat{x}_c^* \otimes \hat{x}'_c))) \tag{1}$$

where $x$ is each sample in the circular matrix $X$, $f^{-1}$ is the inverse Fourier transform, $x^*$ is the complex-conjugate of $x$, $\hat{x}^*$ is the discrete Fourier transform of $x^*$, and $k^{xx}$ is the first-row element of kernel function $k = C(k^{xx})$.

Then, the filter template's size is obtained using the kernel matrix and the ideal Gaussian output response. In the calculation, the kernel matrix is a cyclic matrix. Because of the large amount of data in the image, the kernel function can be diagonalized in the frequency domain to speed up the algorithm's calculation. The kernelized ridge regression classifier weights are shown in Equation (2):

$$\hat{\alpha} = \frac{\hat{y}}{\hat{k}^{xx} + \lambda} \tag{2}$$

where $y$ is the output expectation and $\lambda$ is the regularization coefficient of the filter template.

In the detection stage, the features of the candidate regions are first extracted, and then the current regional features are calculated using the kernel function. The rapid detection is shown in Equation (3):

$$\hat{f}(z) = \hat{k}^{xz} \otimes \hat{\alpha} \tag{3}$$

The ideal regression expectation is the Gaussian, and the more like the tracking result of the previous frame, the greater the chance it is the tracking result of this frame. The center point in the next frame is more likely around the yellow point (inside the yellow box) in the region of interest (ROI), so the ideal regression is more likely to be the center than around in Figure 2. The box's position has changed, showing the SIT has moved.

### *2.2. ResNet*

The ResNet is mainly for classification tasks [27]. The so-called skip connection is used to solve the degradation problem in ResNet. Essentially, it directly connects the shallow network to the deep one and can create a deeper one without losing performance. Even in a smaller network, it is also a reliable method. The overall network structure of the ResNet classification model in this study is shown in Figure 3. The features are fed into a convolution layer, followed by three residual building blocks. Finally, the results of classification are output. It should be emphasized that the model is selected after repeated tests during training and validation.

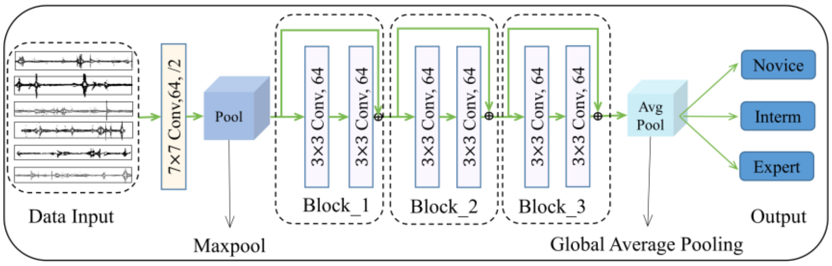

**Figure 3.** A 34-layer neural network structure with three residual building blocks.

The ResNet is composed of a series of residual building blocks. A block model is shown in Figure 4, and it can be expressed as Equation (4):

$$x_{l+1} = x_l + \mathcal{F}(x_l, W_l) \tag{4}$$

The residual building blocks contain two mappings: (1) the identity mapping, represented by $h(x_l)$, which is the right curve in Figure 4 (a); (2) The residual mapping. Residual refers to the $\mathcal{F}(x_l, W_l)$, generally consists of two or three convolutions, which is the left part in Figure 4 (a). In the convolution network, the number of Feature Maps in $x_l$ and $x_{l+1}$ may be different, and then the $1 \times 1$ Conv is needed to increase or reduce the dimension, which is shown in Figure 4 (b). The weight corresponds to $3 \times 3$ Conv, 64 shown in Figure 3. It can be expressed as Equation (5):

$$x_{l+1} = h(x_l) + \mathcal{F}(x_l, W_l) \tag{5}$$

Where $h(x_l) = W_l'x$, $W_l'x$ is the $1 \times 1$ conv.

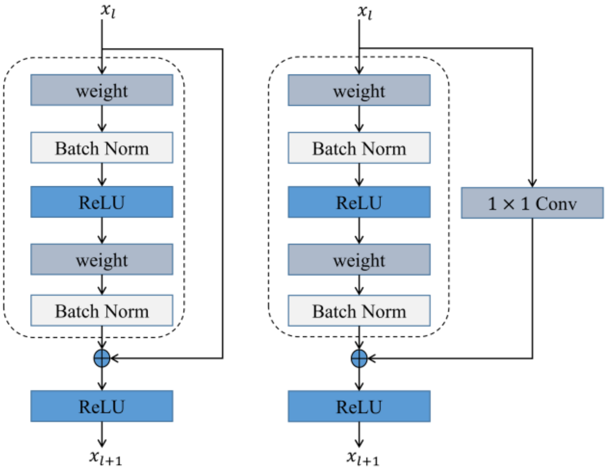

**Figure 4.** The residual building block in ResNet.

## 3. Experimental and Results

### 3.1. Dataset

We use the video collection in JASMAS to simulate the manipulation motion of the surgical robot. The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [31] was produced by Johns Hopkin University and Intuitive Surgery [32]. The JIGSAWS contain kinematic, video, and gesture data in three basic surgical tasks (Suturing, knot-tying, and needle-passing). Meanwhile, a Global Rating Score (GRS) assigned using the improved Objective Structured Assessment of Technical Skills is contained in JIGSAWS [33]. The data is collected from eight participants (B, C, D, E, F, G, H, and I), from the novice to expert at three levels. As shown in Figure 5, the participants performed each task five times by controlling the da Vinci surgical robot. These three tasks are standard parts of the surgical skills training curriculum [15]. Two skill labels are recorded in JIGSAWS: (1) The self-proclaimed skill labels, which are based on surgical robot practice time. The experts report more than 100 hours, the intermediates report between 10 and 100 hours, and the novices report less than 10 hours; (2) The labels based on GRS (scores range from 6 to 30). This is done manually by experienced surgeons. The higher the score, the higher the skill level. This study used the self-proclaimed skill level as the true label for the trial and compared it with the skill levels based on GRS.

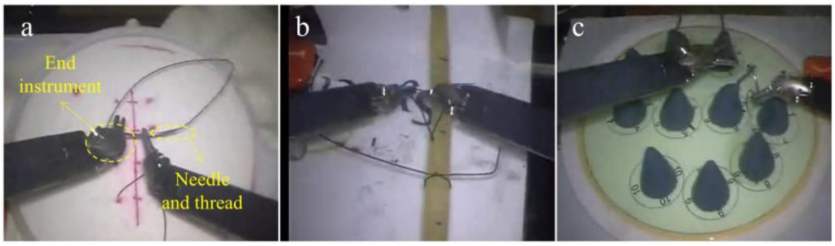

**Figure 5.** Three basic surgical tasks in JIGSAWS. (a) suturing, (b) knot-tying, and (c) needle-passing.

This study focuses on the suturing videos because it has a longer execution time and more complex actions in JIGSAWS. Only the twenty-four suturing videos selected to ensure the same quantity of input from the novices, intermediates, and experts are used as the experimental object. These videos are recorded at a 30 Hz sampling frequency. Table 1 shows more details. By the way, the other two tasks are the same experimental methods in this study, and we do not repeat them.

**Table 1.** The needed details of the suturing tasks in this experiment.

| Self-proclaimed skill labels | Name | Number of videos | Time (s) | The GRS |
|---|---|---|---|---|
| Novice | B, G, H, I | 8 | 172.5±58.3 | 14.5±2.9 |
| Intermediate | C, F | 8 | 90.8±15.1 | 24.0±3.8 |
| Expert | D, E | 8 | 83±13.3 | 17.3±2.5 |

Some values shown are the mean±standard deviation.

## *3.2. Experimental Setup*

### 3.2.1. Process of Visual Motion Tracking

A tracking program is designed based on KCF and runs in python. This program is used to automatically identify and track the ROI of the visible part of SIT in the 2D continuous video frames and record the key motion signals. The quality of surgical operation in RAMIS is presented by evaluating the motion mode of SIT. Such tracking methods have also been used to study the differences in physician hand movements in routine surgery [34, 35]. The center pixel position of the ROI in each frame (every thirtieth of a second) in the videos will be identified and tracked. Then the position coordinates ($x$, $y$) and their running time ($t$) will be automatically recorded. The KCF can overcome some short-time accidents, such as the instrument being blocked and covering the other and motion mutation. But the ROI position sometimes needs to be corrected, so we set the ROI can be manually selected box. As shown in Figure 6, the red box is the ROI selected manually. The minor differences in the box's size and position are ignored as long as the instrument is included.

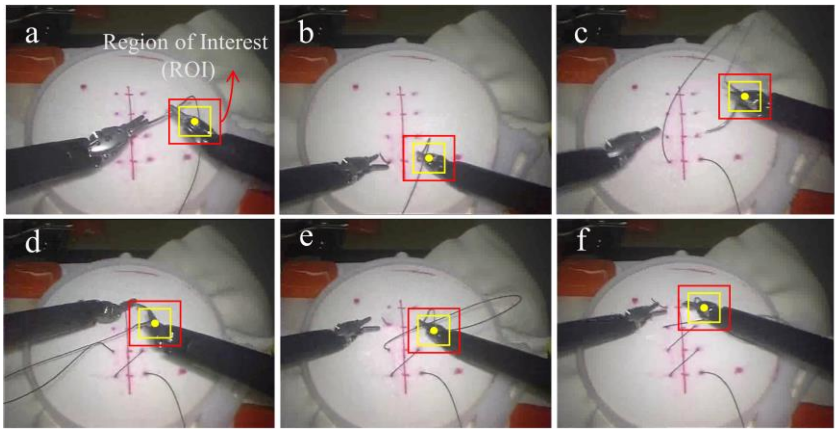

**Figure. 6.** The process of suturing is shown from a to f.

The trajectory of SIT is shown in Figure 7. The light blue part is the course projection in the X-Y plane. The length of the trajectory is 10105 px, 8078.4 px, and 4317.5 px, respectively, which can be calculated by $d_{n+1}$ in Table 2. The trajectory curve of the novice is the most complicated, and the expert is the smoothest in the same suturing task. The novice has more redundant actions, resulting in more than 80 seconds than experts and intermediates to complete this suture task. Consequently, the distinction in the suturing skill of different operators can be seen clearly from the trajectory curve.

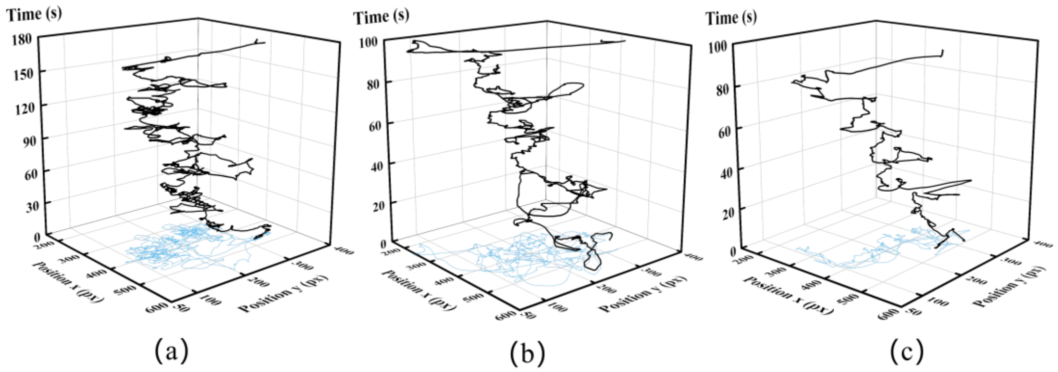

**Figure 7.** The SIT's trajectory in a group of suturing. (a) is from novices, (b) is from intermediates, and (c) is from experts.

### 3.2.2. Key Motion Futures

The tracking record for the position of SIT can quantify the instantaneous displacement, velocity, acceleration, velocity curvature, and motion jerk [36]. In this study, the key motion features in Table 2 are recorded as the input of the ResNet to evaluate the surgical skills. Motion data are captured and saved into CSV files on the PC according to the surgical tasks and the expertise level of users via software implemented in Python. Some features are obtained by calculating the difference by code.

**Table 2.** The specific key motion feature parameters.

| Symbol | Description | Formula |
|---|---|---|
| $t_n$ | The time recorded at frame n | / |
| $x_n$ | Position x-coordinate at frame n | / |
| $y_n$ | Position y-coordinate at frame n | / |
| $d_{n+1}$ | Distance moved between consecutive frames | $\sqrt{(x_{n+1}-x_n)^2+(y_{n+1}-y_n)^2}$ |
| $v$ | The mean velocity of the ROI in consecutive frames | $\sqrt{\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2}$ |
| $a$ | Mean acceleration of the ROI in consecutive frames | $\frac{dv}{dt}$ |
| $MJ$ | A parameter based on the cubic derivative of displacement with time, which means the change in the motion acceleration of the ROI to study motion smoothness | $\sqrt{\left(\frac{d^3x}{dt^3}\right)^2+\left(\frac{d^3y}{dt^3}\right)^2}$ |

The SIT's velocity, acceleration, and motion jerk curves are shown in Figure 8 as a quantitative performance of speed-stationarity-smoothness. These key signals are important features to measure surgical skills [36]. It can be seen that the three levels of operations show a linear trend. Still, the swings of the curves are different, reflecting the distinction in the actions of the three levels of operators. Compared to another two groups of operators, the curve of experts has less swing and less abnormal data, which shows the smoother suturing and the higher quality of the expert.

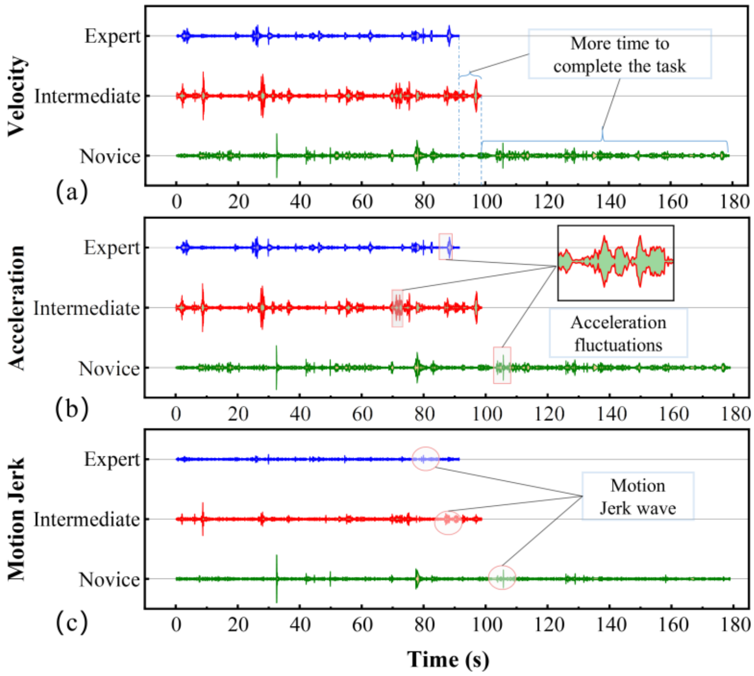

**Figure 8.** The signal graph of the motion from three-level operators in suturing. (a) is the velocity and time, (b) is the acceleration and time, and (c) is the motion jerk and time.

### 3.2.3. Implementation Details of Classification

This study’s assessment of surgical skills is formalized as a supervised classification problem. The input of the ResNet is the whole MTS of the kinematics of the end-effector in the surgical robot, which is recorded by the KCF tracking model, including [$x$, $y$, $t$, $v$, $a$, $MJ$]. Each feature represents a dimension of the ResNet input vector. The length of each input vector data depends on the motion’s time. This is accomplished using the benchmark’s sliding window preprocessing method implemented by Anh et al. [17]. The same padding is used in most places, to keep the dimensions of the output.

$$output\_width = \frac{W - F_w + 2P}{S_w} + 1 \tag{6}$$

$$output\_height = \frac{H - F_h + 2P}{S_h} + 1 \tag{7}$$

where $W$ and $H$ are the width and height of the input, $S$ is stride length, $F$ is filter dimensions and $P$ is the padding size (i.e., the number of rows or columns to be padded). In the case of the same padding, the following stands:

$$output\_width = ceil(\frac{H}{S_h}) \tag{8}$$

$$output\_width = ceil(\frac{W}{S_w}) \tag{9}$$

The output is a predicted label representing the corresponding professional level of the trainees, which can be encoded as 0: novice, 1 : intermediate, and 2: expert. The hyperparameters are selected empirically with a learning rate of 0.001 and a batch size of 24 and trained in a maximum of 100 epochs. To implement this network structure, the ResNet is trained from scratch without any pre-training model. It runs based on Python using the Keras library and TensorFlow on a computer with an Intel Core i5-10400F processor with 2.90 GHz and 16 GB RAM. To ensure that the results are more objective and accurate, as chosen by Anh et al. [17], each method is run five times for each generated input file.

Within each run, five trials use the Leave-one-super trial-out (LOSO) cross-validation method, and the mean accuracy is calculated.

#### 3.2.4. Modeling Performance Measures

In this study, four Common Indexes [37, 38]are applied to evaluate the performance of the classification model:

- *accuracy*, the ratio between the number of samples correctly classified and the total number of samples;

$$accuracy = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \tag{11}$$

- *precision*, the ratio between the correct positive predictions and the total positive results predicted by the classifier;

$$precision = \frac{T_p}{T_p + F_p} \tag{12}$$

- *recall*, the ratio between the positive predictions and the total positive results in the ground truth;

$$recall = \frac{T_p}{T_p + F_n} \tag{13}$$

- *F1-score*, a weighted harmonic average between precision and recall.

$$F1 - score = \frac{2 * (recall * precision)}{recsll + precision} \tag{14}$$

where $T_p$ and $F_p$ are the numbers of true positives and false positives, $T_n$ and $F_n$ are the numbers of true negatives and false negatives for a specific class.

### *3.3. Results*

In this study, a proposed automatic assessment framework of surgical skills in RAMIS based on endoscopic visual motion tracking technology and deep learning neural network is verified in JIGSAWS. Figure 9 shows the confusion matrix of classification results. Figure 9 (a) shows the complete three classifications, and Figure 9 (b) uses the results of two classifications without the intermediates. Specifically, the model's accuracy in this study reached 92.04% and 84.80% when the suturing task is divided into two and three classifications. The performance of fewer class classifications is naturally better than more class classifications, but the reason why the gap is so significant is worth analyzing and discussing (section 4). Among these three performance indicators, the 3-class accuracy is fairly poor. However, the novice is more accurate, reaching 96%. For experts, the worst accuracy of the classification is only 39%. The results are 3% and 53% higher than those of the three classifications when only labeled novice and expert.

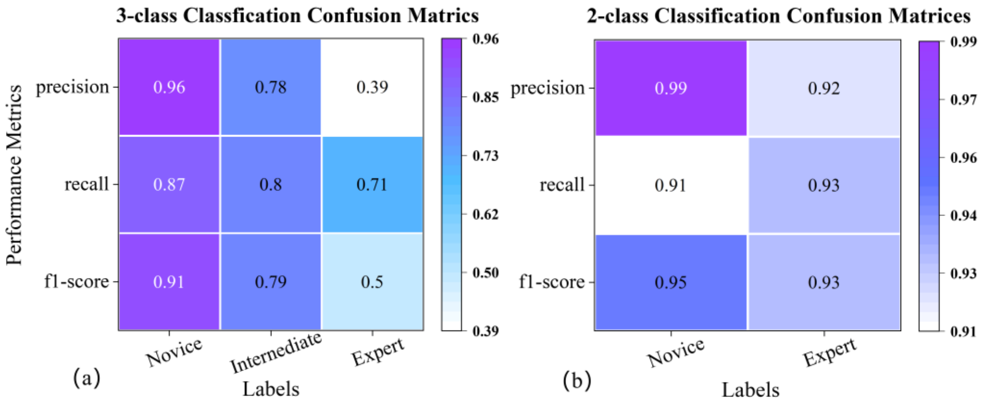

**Figure 9.** The confusion matrix for two and three classifications. (a) is the 3-class result; (b) is the 2-class result. The element value and color mean the probability of predicting skill labels, where the skill labels are self-proclaimed.

The results of this study are compared to the most advanced classifications in Table 3. They used JIGSAWS as a visual input source and performed experiments under the LOSO scheme. It can be seen the new model achieved fairly accurate results, which proves the skill assessment method for RAMIS proposed in this study is feasible.

**Table 3.** The results of this study are compared to those of the latest technology.

| Author (Year） | Method | Suture |
|---|---|---|
| Ming et al. (2021) [24] | STIP | 79.29% |
| Ming et al. (2021) | IDT | 76.79% |
| Lajkó G et al. (2021) [25] | CNN | 80.72% |
| Lajkó G et al. (2021) | CNN + LSTM | 81.58% |
| Lajkó G et al. (2021) | ResNet | 81.89% |
| Current Study | KCF + ResNet | 84.80% |

A different set of experiments that we performed with alternative architectures according to the same experimental arrangement and parameter configuration to support the ResNet better are carried out on LSTM, CNN, and CNN+LSTM. In Figure 10, the abscissa is the input features in the neural network, and the specific parameters are shown in Table 4. Firstly, the ResNet performs the best when the input features are 5. The accuracy of the ResNet in the case of three classes is 1.44%, 4.92%, and 6.76% higher than that of other networks. The accuracy of the ResNet in the case of two classes is 4.16%, 10.86%, and 11.8% higher than that of other networks. Secondly, the influence of input features on the results is based on trajectory data. With the increase of input data, the classification accuracy is higher. However, in Fig. 10 (b), the classification accuracy of the four input features decreased significantly, with a maximum reduction of 13.16% (ResNet). It can be seen that trajectory data have a significant impact on the results of the two classifications. The problem of accuracy is discussed in section 4. Due to the limitation of current technology, only these four key motion features can be recorded. The classification accuracy may be higher if more features can be obtained, and then more accurate feedback on skill assessment results can also be obtained.

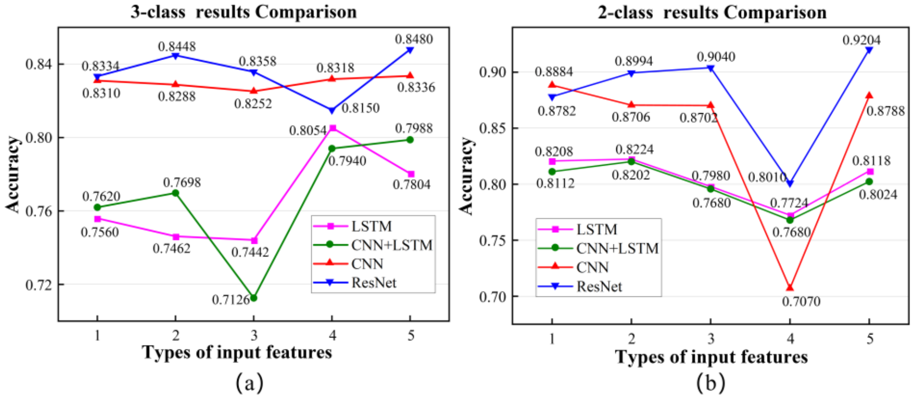

**Figure 10.** The comparison of classification results for different networks and different input features.

**Table 4.** The meaning of types of input features.

| Number | Input features |
|---|---|
| 1 | $[x, y, t]$ |

| | |
|---|---|
| 2 | $[x, y, t, v]$ |
| 3 | $[x, y, t, v, a]$ |
| 4 | $[v, a, MJ]$ |
| 5 | $[x, y, t, v, a, MJ]$ |

The ResNet not only has higher classification accuracy but also is competitive in computational efficiency with feedback of classification within 3 to 5 seconds in Table 5， thanks to the network structure of the jump connection. So, the ResNet is more suitable for the framework of surgical skill assessment than other networks in this study.

**Table 5.** The computational processing time of different neural networks.

| Input features | Method | Time |
|---|---|---|
| $[x, y, t, v, a, MJ]$ | CNN | 1~3 seconds |
| | ResNet | 3~5 seconds |
| | CNN+LSTM | 24~48 seconds |
| | LSTM | 16~68 seconds |

## 4. Discussion

### *4.1. Performance of the Framework*

The proposed surgical skill assessment framework has been effectively validated on the JIGSAWS. The new model's accuracy is 92.04% and 84.80% in the case of two and three classifications. It is proven the new method can effectively and accurately assess the quality of surgical operation and skill level in RAMIS. However, it is worth mentioning the intermediates and experts are prone to misclassification in the case of three classifications, and only 78% and 39% accuracy are achieved. These problems also appeared in the studies of Funk et al. [23], Anh et al. [17], and Lefor et al. [31]. To figure this out, we discussed the motion data gained by the KCF and the dataset.

### *4.2. Motion Features Assessment*

The motion features of SIT are analyzed from the results recorded by the KCF algorithm in this paper. Figure 11 shows the mean value of the three motion features distributed within a 99% confidence interval (CI). In Figure 11 (a), the operating speed of experts and intermediates is close and only differs by 0.004 px/s. As shown in Figure 11 (b), the intermediates got the maximum acceleration. Still, they had not taken the shortest time, which means many motion mutations of SIT are done during movement. Figure 11 (c) also shows that similar motion jerks happened to intermediates and experts and only differed by 0.002 px/s, but the novices performed best. The presence of deviating points in the graph may be due to the misclassification of the dataset itself. The mean square error of trajectory ($S$) shows a strong correlation with the operation ability of the instruments in the suturing task. Figure 11 (d) shows the deviation degree of the trajectory points relative to the trajectory center. The larger $S$ means exploratory or ineffective movement. Interestingly, what is reflected in Figure 11 (c) and (d) is that novices have the best results. Because novices are often cautious when performing because of inexperience, the same action will take more time and make more detailed actions. In Equation (10), owing to the more significant number of sampling points n at the same length (compared with the other two levels), the minor difference between continuous $x$ and $y$ results in a smaller $S$. The mean square deviation and motion jerk is hard to distinguish the detail skills accurately. So, the insignificant difference in motions between experts and intermediates causes the neural network not to indicate these two levels well.

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})^2 + (y_i - \bar{y})^2]} \quad (15)$$

where $x_i$ and $y_i$ are the two-dimensional coordinate values of the trajectory; $\bar{x}$ and $\bar{y}$ are the mean values of $x_i$ and $y_i$; n is the number of sampling points.

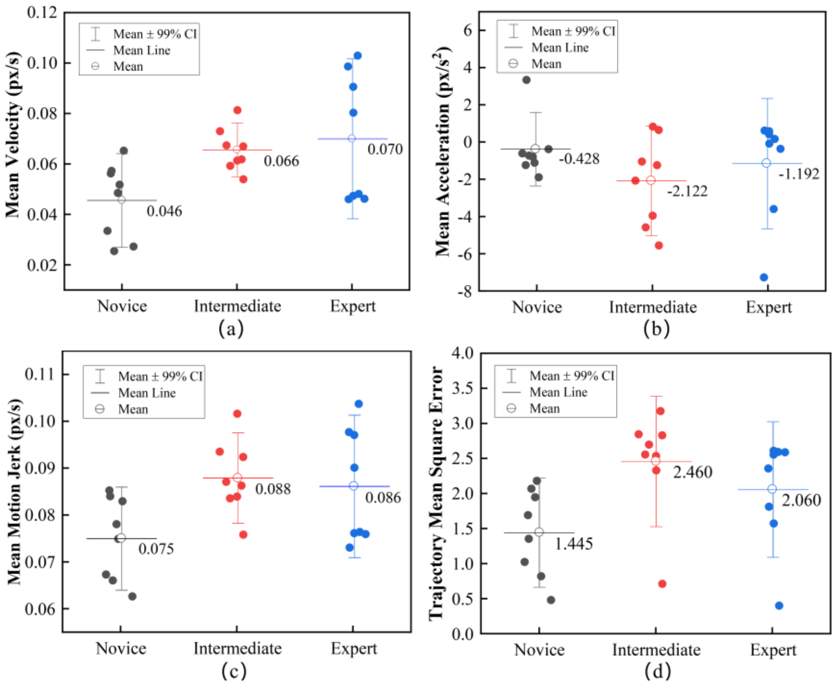

**Figure 11.** The motion features of the three-level performance in suturing. (a) is the mean velocity, (b) is the mean acceleration, (c) is the mean motion jerk, and (d) is the mean square error of trajectory.

### 4.3. Dataset Assessment

The GRS in JIGSAWS contains six scales scored from 1 to 5, including (1) Respect for tissue, (2) Suture handling, (3) Time and motion, (4) Flow of operation, (5) overall performance, and (6) Quality of final product15. Figure 12 shows the distribution of the GRS, thus reflecting the performance of the operation. The interquartile range (IQR) measures the degree of dispersion in the box plot. As can be seen the intermediates get the highest composite score with a median of 3.813, showing the best overall performance for intermediates, followed by experts and novices. This means the mismatch is between the GRS and the self-proclaimed skill labels. Therefore, the GRS in JIGSAWS does not accurately distinguish the three levels of surgical operation skill.

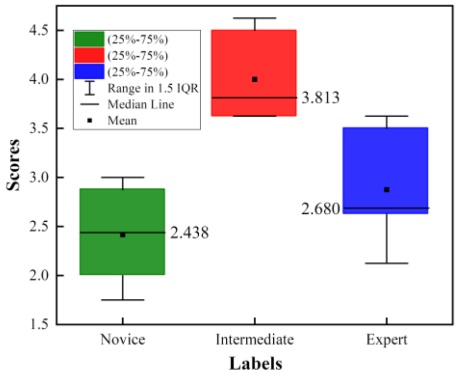

**Figure 12.** The box plot of the GRS of three skills.

The truth can be seen in Figure 11 and Figure 12, in which the labels are inconsistent based on the self-proclaimed and the GRS. Intermediates perform better than experts in most situations in this dataset. Therefore, a significant impact is presented to the model of classifications whether the true label is consistent with the true skill level. Most of the classification errors occurred in the failure to identify self-proclaimed intermediates and experts in this study. The classification accuracy significantly improves (7.24%) when the intermediates are removed. It can be attributed to the fact the self-proclaimed labels based on the practice time of surgical robot operation cannot accurately reflect the true skills. So, the more accurate labels, the better performance of the assessment framework.

### *4.4. Limitations and Future Research*

The development of the RAMIS has promoted great research for objective skill assessment methods [39]. The current work has made some progress, but there still are some limitations to practicing online skill assessment in this new model. First, this study has shown the potential use of the KCF in RAMIS skill assessment, proving that visual solutions may replace kinematics [40]. However, the accuracy of motion tracking cannot reach 100% accuracy in the surgery due to the complex working environment and occlusion problems. Second, supervised deep learning classification accuracy depends mainly on labeled samples. This study focuses on the videos of the JIGSAWS, which lacks strict essential fact labels for skill levels. The self-proclaimed skill is labeled according to the operation time. It isn't easy to judge whether it is true or accurate. Besides, skill labels are annotated according to predefined GRS score thresholds in GRS-based labels, but there is no universally accepted threshold. So, a more precise labeling method and more professional and in-depth surgeon knowledge may improve the skill assessment accuracy [41, 42]. Finally, the interpretability of automatic learning representations is still limited due to the black-box nature of the deep learning model.

This work proposes a new and feasible method rather than finding the best one. More advanced neural networks will be used in this framework in further studies. The endoscopic vision technology will be deeply studied to solve occlusion problems and get depth information effectively. The motion tracking technology in three-dimensional space will be explored to improve further the accuracy of skill assessment based on endoscopic. Also, the deep architecture, parameter setting, and improvement strategy of the deep learning neural network will be optimized in detail to better process the data of the motion time series and further improve the performance of online assessment.

## 5. Conclusions

Efficient and accurate skill assessment in RAMIS is essential to ensure patient safety. This study proposes a novel evaluation framework based on endoscopic visual motion tracking technology and deep learning. The new approach replaces traditional sensors with vision technology, innovatively combining vision and kinematics. By using KCF to track and obtain two-dimensional motion signals based on endoscopic vision, such as the trajectory, velocity, and acceleration of SIT. ResNet is then used for automatic and accurate classification and analysis of surgical skills, and the results are compared with state-of-the-art research in the field. Finally, the reasons for some classification errors are discussed, and the limitations of this study are pointed out.

The contributions of this study are: (1) Provides an efficient and accurate framework for skill assessment in RAMIS, with classification accuracies of 84.80% and 92.04%, which can accurately provide feedback on online assessment results. (2) The classification technology framework based on endoscopic vision and neural network simplifies the access process, and the feedback of the results can be realized within 3 to 5 seconds, thereby improving the efficiency of the assessment of surgical skills. (3) The proposed method can automatically complete the whole process of surgical skill assessment without using additional tools other than the endoscope, so it was more valuable for application.

In conclusion, the aim was to propose a method for the assessment of surgical skills that combines vision and kinematics. The new method can effectively consider the advantages of vision and kinematics in the assessment of surgical skills, achieving a higher level of two-dimensional visual assessment. It can be easily integrated and applied to the system in RAMIS. Real-time and accurate feedback can be obtained during personalized surgery, improving surgeon training efficiency and ensuring surgical quality and safety.

**Author Contributions:** Conceptualization, K.L. and S.W.; methodology, M.P.; software, S.W.; validation, M.P., K.L. and S.W.; formal analysis, J.L.; investigation, S.W.; resources, X.Y.; data curation, S.W.; writing—original draft preparation, S.W.; writing—review and editing, J.L.; visualization, X.Y.; supervision, K.L.; project administration, M.P.; funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Guangxi Science and Technology Base and Talent Special Project (Grant No. 2021AC19324), Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology（Grant No.20-065-40S006）.

**Data Availability Statement:** The experiments' source code and required data sets can be obtained upon request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lane, T., A short history of robotic surgery. *Ann R Coll Surg Engl* **2018,** 100, (6_sup), 5-7.
2. Nagy, T. a. D.; Haidegger, T. a., A DVRK-based Framework for Surgical. *Acta Polytechnica Hungarica* **2019,** 16, 68-71.
3. Aggarwal, R.; Mytton, O. T.; Derbrew, M.; Hananel, D.; Heydenburg, M.; Issenberg, B.; MacAulay, C.; Mancini, M. E.; Morimoto, T.; Soper, N.; Ziv, A.; Reznick, R., Training and simulation for patient safety. *Qual Saf Health Care* **2010,** 19 Suppl 2, i34-43.
4. Birkmeyer, J. D.; Finks, J. F.; O'Reilly, A.; Oerline, M.; Carlin, A. M.; Nunn, A. R.; Dimick, J.; Banerjee, M.; Birkmeyer, N. J.; Michigan Bariatric Surgery, C., Surgical skill and complication rates after bariatric surgery. *N Engl J Med* **2013,** 369, (15), 1434-42.
5. Darzi, A.; Mackay, S., Assessment of surgical competence. *Qual Health Care* **2001,** 10 Suppl 2, (Suppl 2), ii64-ii69.
6. Farcas, M. A.; Trudeau, M. O.; Nasr, A.; Gerstle, J. T.; Carrillo, B.; Azzie, G., Analysis of motion in laparoscopy: the deconstruction of an intra-corporeal suturing task. *Surg Endosc* **2017,** 31, (8), 3130-3139.
7. Shanmugan, S.; Leblanc, F.; Senagore, A. J.; Ellis, C. N.; Stein, S. L.; Khan, S.; Delaney, C. P.; Champagne, B. J., Virtual reality simulator training for laparoscopic colectomy: what metrics have construct validity? *Dis Colon Rectum* **2014,** 57, (2), 210-4.
8. Ebina, K.; Abe, T.; Higuchi, M.; Furumido, J.; Iwahara, N.; Kon, M.; Hotta, K.; Komizunai, S.; Kurashima, Y.; Kikuchi, H.; Matsumoto, R.; Osawa, T.; Murai, S.; Tsujita, T.; Sase, K.; Chen, X.; Konno, A.; Shinohara, N., Motion analysis for better understanding of psychomotor skills in laparoscopy: objective assessment-based simulation training using animal organs. *Surg Endosc* **2021,** 35, (8), 4399-4416.
9. Jiang, J.; Xing, Y.; Wang, S.; Liang, K., Evaluation of robotic surgery skills using dynamic time warping. *Computer Methods and Programs in Biomedicine* **2017,** 152, 71-83.
10. Oquendo, Y. A.; Riddle, E. W.; Hiller, D.; Blinman, T. A.; Kuchenbecker, K. J., Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surg Endosc* **2018,** 32, (4), 1840-1857.
11. Sbernini, L.; Quitadamo, L. R.; Riillo, F.; Lorenzo, N. D.; Gaspari, A. L.; Saggio, G., Sensory-Glove-Based Open Surgery Skill Evaluation. *IEEE Transactions on Human-Machine Systems* **2018,** 48, (2), 213-218.

12. Beulens, A. J. W.; Namba, H. F.; Brinkman, W. M.; Meijer, R. P.; Koldewijn, E. L.; Hendrikx, A. J. M.; van Basten, J. P.; van Merrienboer, J. J. G.; Van der Poel, H. G.; Bangma, C.; Wagner, C., Analysis of the video motion tracking system "Kinovea" to assess surgical movements during robot-assisted radical prostatectomy. *Int J Med Robot* **2020,** 16, (2), e2090.
13. Ganni, S.; Botden, S.; Chmarra, M.; Goossens, R. H. M.; Jakimowicz, J. J., A software-based tool for video motion tracking in the surgical skills assessment landscape. *Surg Endosc* **2018,** 32, (6), 2994-2999.
14. Rivas-Blanco, I.; P'erez-del-Pulgar, C. J.; Mariani, A.; Quaglia, C.; Tortora, G.; Menciassi, A.; Muñoz, V. F., A surgical dataset from the da Vinci Research Kit for task automation and recognition. *ArXiv* **2021,** abs/2102.03643.
15. Gao, Y.; Vedula, S. S.; Reiley, C. E.; Ahmidi, N.; Varadarajan, B.; Lin, H. C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D. D.; Chen, C. C. G.; Vidal, R.; Khudanpur, S.; Hager, G. In *JHU-ISI Gesture and Skill Assessment Working Set ( JIGSAWS ): A Surgical Activity Dataset for Human Motion Modeling*, 2014; 2014.
16. Kitaguchi, D.; Takeshita, N.; Matsuzaki, H.; Takano, H.; Owada, Y.; Enomoto, T.; Oda, T.; Miura, H.; Yamanashi, T.; Watanabe, M.; Sato, D.; Sugomori, Y.; Hara, S.; Ito, M., Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical Endoscopy* **2019,** 34, (11), 4924-4931.
17. Anh, N. X.; Nataraja, R. M.; Chauhan, S., Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. *Comput Methods Programs Biomed* **2020,** 187, 105234.
18. Zhang, J.; Nie, Y.; Lyu, Y.; Yang, X.; Chang, J.; Zhang, J. J., SD-Net: joint surgical gesture recognition and skill assessment. *Int J Comput Assist Radiol Surg* **2021,** 16, (10), 1675-1682.
19. Nguyen, X. A.; Ljuhar, D.; Pacilli, M.; Nataraja, R. M.; Chauhan, S., Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer Methods and Programs in Biomedicine* **2019,** 177, 1-8.
20. Wang, Z.; Majewicz Fey, A., Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg* **2018,** 13, (12), 1959-1970.
21. Yanik, E.; Intes, X.; Kruger, U.; Yan, P.; Diller, D.; Voorst, B.; Makled, B.; Norfleet, J.; De, S., Deep neural networks for the assessment of surgical skills: A systematic review. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* **2021**, 154851292110345.
22. Lee, D.; Yu, H. W.; Kwon, H.; Kong, H.-J.; Lee, K. E.; Kim, H. C., Evaluation of Surgical Skills during Robotic Surgery by Deep Learning-Based Multiple Surgical Instrument Tracking in Training and Actual Operations. *Journal of Clinical Medicine* **2020,** 9, (6).
23. Funke, I.; Mees, S. T.; Weitz, J.; Speidel, S., Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Assist Radiol Surg* **2019,** 14, (7), 1217-1225.
24. Ming, Y.; Cheng, Y.; Jing, Y.; Liangzhe, L.; Pengcheng, Y.; Guang, Z.; Feng, C., Surgical skills assessment from robot assisted surgery video data. In *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 2021; pp 392-396.
25. Lajkó, G.; Nagyné Elek, R.; Haidegger, T., Endoscopic Image-Based Skill Assessment in Robot-Assisted Minimally Invasive Surgery. *Sensors* **2021,** 21, (16), 5412.
26. Henriques, J. F.; Caseiro, R.; Martins, P.; Batista, J., High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans Pattern Anal Mach Intell* **2015,** 37, (3), 583-96.
27. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016; pp 770-778.
28. Zheng, K.; Zhang, Z.; Qiu, C., A Fast Adaptive Multi-Scale Kernel Correlation Filter Tracker for Rigid Object. *Sensors (Basel)* **2022,** 22, (20).

29. Rifkin, R.; Yeo, G.; Poggio, T., Regularized Least-Squares Classification. *Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and Systems Sciences* **2003,** 190.
30. Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; Yang, M.-H. In *Fast Visual Tracking via Dense Spatio-temporal Context Learning*, Computer Vision – ECCV 2014, Cham, 2014//, 2014; Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Eds. Springer International Publishing: Cham, 2014; pp 127-141.
31. Lefor, A. K.; Harada, K.; Dosis, A.; Mitsuishi, M., 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* **2020,** 15, (12), 2017-2025.
32. da Vinci Surgical System, Intuitive Surgical, Inc. https://www.davincisurgery.com/. (accessed on 25 October 2022)
33. Martin, J. A.; Regehr, G.; Reznick, R.; Macrae, H.; Murnaghan, J.; Hutchison, C.; Brown, M., Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery* **2005,** 84, (2), 273-278.
34. Azari, D. P.; Frasier, L. L.; Quamme, S. R. P.; Greenberg, C. C.; Pugh, C. M.; Greenberg, J. A.; Radwin, R. G., Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Ann Surg* **2019,** 269, (3), 574-581.
35. Frasier, L. L.; Azari, D. P.; Ma, Y.; Pavuluri Quamme, S. R.; Radwin, R. G.; Pugh, C. M.; Yen, T. Y.; Chen, C. H.; Greenberg, C. C., A marker-less technique for measuring kinematics in the operating room. *Surgery* **2016,** 160, (5), 1400-1413.
36. Liang, K.; Xing, Y.; Li, J.; Wang, S.; Li, A.; Li, J., Motion control skill assessment based on kinematic analysis of robotic end-effector movements. *Int J Med Robot* **2018,** 14, (1).
37. Ahmidi, N.; Tao, L.; Sefati, S.; Gao, Y.; Lea, C.; Haro, B. B.; Zappella, L.; Khudanpur, S.; Vidal, R.; Hager, G. D., A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. *IEEE Trans Biomed Eng* **2017,** 64, (9), 2025-2041.
38. Kumar, R.; Jog, A.; Malpani, A.; Vagvolgyi, B.; Yuh, D.; Nguyen, H.; Hager, G.; Chen, C. C., Assessing system operation skills in robotic surgery trainees. *Int J Med Robot* **2012,** 8, (1), 118-24.
39. Vedula, S. S.; Ishii, M.; Hager, G. D., Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. *Annu Rev Biomed Eng* **2017,** 19, 301-325.
40. Hasan, M. K.; Calvet, L.; Rabbani, N.; Bartoli, A., Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Med Image Anal* **2021,** 70, 101994.
41. Dockter, R. L.; Lendvay, T. S.; Sweet, R. M.; Kowalewski, T. M., The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. *Int J Comput Assist Radiol Surg* **2017,** 12, (7), 1151-1159.
42. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A., Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017; pp 843-852.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.