ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 912-921 (2021)

www.astesj.com

Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Procrustes Dynamic Time Wrapping Analysis for Automated Surgical Skill Evaluation

Safaa Albasri[1,*], Mihail Popescu[2], Salman Ahmad[3], James Keller[1]

[1]*Electrical Engineering and Computer Science, University of Missouri, Columbia, 65211, USA*

[2]*Health Management and Informatics, University of Missouri, Columbia, 65211, USA*

[3]*Acute Care Surgery, University of Missouri, Columbia, 65211, USA*

ARTICLE INFO

Article history:
Received: 16 September, 2020
Accepted: 21 January, 2021
Online: 12 February, 2021

Keywords:
DTW
Procrustes
k-NN
Accelerometer
Surgery skills evaluation

ABSTRACT

Classic surgical skill evaluation is performed by an expert surgeon examining an apprentice in a hospital operating room. This method suffers from being subjective and expensive. As surgery becomes more complex and specialized, there is an increase need for an automated surgical skill evaluation system that is more objective and determines more exactly the skills (or lack thereof) the apprentice has. The main purpose of our proposed approach is to use an existing skill database with known proficiency levels to evaluate the skills of a given apprentice. The skill of the apprentice will be assessed to be similar to the closest skill example found in the database (case-based reasoning). A key element of the system is the skill distance measure employed, as each skill example is a multidimensional time series (sequence) with widely varying values. In this paper, we discuss a new surgery skill distance measure denoted as Procrustes dynamic time warping (PDTW). PDTW integrates the search for exact alignment between two skill sequences using DTW and Procrustes distance as a measure for the similarity. The Procrustes approach is a shape distance analysis that involves rotation, scaling, and translation. We evaluated our proposed distance on three surgical motion data, a widely used JIGSAWS robot surgery dataset, a wearable sensor dataset, and a Vicon motion system dataset. The results showed that the proposed framework produced a better performance for surgeon skill assessment when PDTW was used compared to other time series distances on all three datasets. Also, some experimental results for the JIGSAWS dataset outperformed existing deep learning-based methods.

1. Introduction

This paper is an extension of work initially presented in the E-Health and Bioengineering Conference (EHB) [1] . Recently, the need for objective surgical skills assessment has captured the interest of practitioners and medical institutions due to the ever-increasing complexity and degree of specialization of the surgical procedure [2]. Traditionally, a senior expert surgeon performs direct observation, scores, assess, and gives feedback to the trainee surgeon (apprentice) with less practice in the hospital operating room. This traditional surgical proficiency evaluation approach is problematic due to its subjectivity, time consumption and cost. Furthermore, it is prone to errors and sometimes insufficient as lacking details related to deficiencies. To address these difficulties, an automated skill assessment procedure is needed for an objective and detailed measure of proficiency levels [3, 4].

As any healthcare domain, surgery is continuously changed by technological advances and medical innovations that alter everyday surgical procedures. The challenge is to assist surgical procedure via the quantifiable data analysis to a better understanding of the surgical operating and to obtain more knowledge about human activities during surgery for advance and further study [5]. A reasonable solution to these challenges is to use technological advances like Robotic Minimally Invasive Surgery (RMIS) that improve overall operating room efficiency [5]. For instance, da Vinci surgical technology provides data-driven that potentially helps optimize and develop training skills

* Corresponding Author: Safaa Albasri, University of Missouri, USA. saaxfc@mail.missouri.edu

for surgeons [6]. This information includes kinematic and video data that conduct a useful resource of quantifiable human motion during surgical operating [7, 8]. Wearable sensing devices that provide detailed motion information for surgical activities are a further example [9]. These recorded data give spacious resources to assess surgical proficiencies by modeling and analyzing descriptive mathematical approaches. The emergence of using machine learning methods with recent robotic surgery systems such as da Vinci and wearable sensing devices via data-driven enable and encourage developers to build and analyze automatic models for evaluating surgeon expertise and may help better coaching potential apprentices [10–12] .

Different earlier works focused on the automated surgical assessment seen good progress. The current techniques for objective surgical evaluation can be divided into three main research areas [10, 13]: 1) surgeon skill assessment, 2) surgical task analysis, and 3) surgemes recognition. These methods considered the surgeon movement using either: 1) kinematic information recorded by a robotic surgical system, 2) video records and 3) wearable sensors data. In this paper, we focused on the surgical skill evaluation based on kinematic and wearable sensors information. One of the initial works used Hidden Markov models (HMM) [14] to evaluate the surgical skills. This approach is structured-based and depends on the number of training samples, tuning parameters and it takes massive pre-processing. This type of model needs complicated preprocessing [3] and leads to low performance with a low number of samples [14]. Another method was proposed by [3] to predict the surgeon skill level (expert and novice) based on movement features of the surgical arms using logistic regression (LR) and support vector machines (SVM) classifiers for suturing surgical task. They extended their work to include eight global movement features (GMF) in [15], they applied LR, SVM, and kNN classifier to distinguish between the previous expertise levels for suturing and knot tying surgical tasks. In [16], a framework based on trajectory shape using DTW and k-nearest neighbor classifier proposed for surgical skill evaluation. This model can also provide online performance feedback through training. More recently, [13] proposed an approach based on symbolic aggregate approximation (SAX) and vector space model (VSM) to identify distinctive patterns of surgical procedure. They used the SAX to obtain the sequence of letters by discretizing the time series first. Then they utilize the VSM to find the discriminative patterns that represent a surgical motion which finally used them to be classified. A variety of holistic analysis features and a weighted features integrated approach proposed by [9] for automated surgical skill evaluation and GRS score prediction. These holistic features include approximate entropy, sequential motion texture, discrete Fourier and discrete cosine transform. They used the nearest neighbor as a classifier and linear support vector regression (SVR) for prediction. The works of literature mentioned above used the kinematic data information obtained from RMIS for surgical skill assessment. However, none of these methods were applied to the wearable sensors data like accelerometer which might give more information about the surgeon's motion during a surgical practice.

Recently, several advanced techniques applied the convolution neural network and deep learning methods for automated surgical skill evaluation. A parallel deep learning framework was proposed by [17] to identify the surgeon skill and task recognition. In their approach, they used a fusion technique between convolution neural networks and gated recurrent networks. Alternative deep convolution neural architecture based on ten layers proposed by [12] for surgical expertise evaluation. Another parallel deep learning approach was proposed in [18] by combining the LSTM recurrent network and CNN to indicate the skill levels. Additionally, recent studies have suggested approaches that use motion from videos [19,20] and wearable sensors to evaluate surgical skills [21,22]. These methods platform various features to perform Objective Structured Assessment of Technical Skills (OSATS) assessments. An approach proposed for surgical skill assessment is based on the acceleration data of both hands performing a basic surgical procedure in dentistry [2]. Also, an entropy-based features technique that utilizes both video and accelerometer data proposed for surgical skill assessment [4]. Despite these techniques which are building the basis and inspire performance results in the surgical skill area, however, some limits and drawbacks occur for the existing methods. some methods need predefined boundaries of the surgemes which done usually by a chief surgeon, i.e., consuming a large time. In other methods, decomposing the motion sequence requires a massive and complicated preprocessing in addition to a deficiency of robustness. Alternatively, the need to developing a new distance measure might have an advantage to a more robust and accurate assessment framework.

In this paper, our contribution to this work can be abridged as follows: 1) we defined a new surgical skill distance combined the best alignments between two multidimensional signals using DTW and measuring the distance between the two aligned sequences using Procrustes analysis 2) we proposed an automated skill classification framework based on using PDTW and kNN technique in the proposed framework to distinguish between the expertise levels focusing on overall performance 3) we investigated the proposed framework on a wearable sensor data for a surgical task. The purpose of this work is to present a technique that handles different kinds of sensor data in addition to the existing public JIGSWAS dataset. Some surgery motion results obtained by a Vicon camera with a 3D marker-based system and wearable device data are examples of the data we use.

2. Methodology

In this section, we illustrate the main components of our proposed framework, which are: motion alignment, Procrustes distance, and skills classifier, as shown in Figure 1. First, DTW is used to align two multidimensional time series performed by surgeons, while the Procrustes distance calculates similarity measure. Lastly, the skill levels of the surgeon are classified by kNN.

2.1. Similarity Measure

To obtain a useful classification, defining a reasonable distance is a crucial element to measure between two surgery tasks. Each

surgery task is represented by a set of features obtained from the traces (time series) of the motion capture sensors. One possible method is the Euclidean distance.

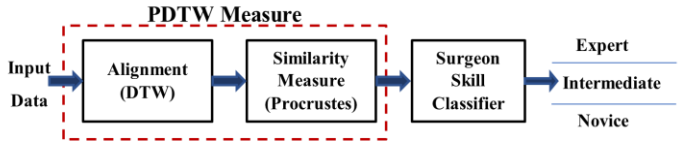

Figure 1: kNN based PDTW evaluation Framework

Euclidian distance is simple and widely used, whereas, it has some limitations and disadvantages. The Euclidean method is very sensitive to outlier and it is suffering from noise, shifting, and requires both signals to have the same length. Thus, we need a measure that can handle sequences with different lengths because the same surgery task might have different lengths even when operated by the same surgeon. A warping distance measure such as the Dynamic Time Warping (DTW), is one solution to do the job. The DTW can process time series with different lengths, it expands or contracts both signals (aligns them) such that their length becomes equal [23].

Let $X_{n\times v} = [X_1, X_2, \dots, X_n]$ and $Y_{m\times v} = [Y_1, Y_2, \dots, Y_m]$ be two sequences having v features and of length n and m respectively. To align X and Y, we form a two-dimensional $(n \times m)$ grid distance. Each point d_{ij} of the grid corresponds to the distance measure (usually Euclidean) between every possible combination of two instances x_i from X and y_j from Y of the same features length (v) as follow [24]:

$$d_{ij}(x_i, y_j) = \sqrt{\sum_{k=1}^{v}(x_{ik} - y_{jk})^2} \quad (1)$$

The next step is to find the warping path through the grid, the path that attempts to minimize the total distance (warping cost) and give the best match between two signals and satisfy boundary conditions, continuity, and monotonicity constraints. It is usually achieved by using a dynamic program to calculate the cumulative distance $\gamma(i,j)$, which is the distance of the current cell (d_{ij}) and the minimum of the cumulative distance of the adjacent cells [24]:

$$\gamma(i,j) = d_{ij} + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (2)$$

Despite the wide use of DTW in many applications and is a more robust distance measure than Euclidean distance, it fails for complex multidimensional signals. Also, when the unevenness occurred in the Y-axis, DTW can produce singularities by warping the X-axis. Inflection points, valleys, and peaks features can cause DTW to fail to align two signals properly [24].

The Procrustes analysis is a standard method in statistical analysis to compare the similarity of shape objects [25, 26]. The Procrustes distance is a shape metric that involves matching two shapes using similarity transformations (rotation, reflection, scaling, translation) to be as close as possible in the least-squares sense [27]. The Procrustes analysis also can estimate the mean shape to examine the shape variability in a dataset [28].

Assume X_1 and X_2 be two configuration matrices of the same $k \times m$ dimension (k points in m dimensions) that can be centered (normalized) using the following equation [28]:

$$(X_i)_c = CX_i \quad , \; i = 1,2 \quad (3)$$

$C = H^T H$ is the centering matrix and H is the Helmert submatrix, let Z_1 and Z_2 be the pre-shapes unit size of X_1 and X_2 respectively, where the original configuration is invariant under the scaling and translation with the pre-shape [28]:

$$Z_i = \frac{(X_i)_H}{\|(X_i)_H\|} = \frac{H(X_i)}{\|H(X_i)\|} \quad , \; i = 1,2 \quad (4)$$

$$(X_i)_H = HX_i \quad , \; i = 1,2 \quad (5)$$

The full Procrustes distance between X_1 and X_2 is achieved by fitting the pre-shape Z_1 and Z_2 as closely as possible as the following [25]:

$$D_P(X_1, X_2) = \inf_{s,a,b,\theta,} \left\| Z_1 - Z_2 \, s \, e^{j\theta} - (Z_2 + jb)1_k \right\| \quad (6)$$

where $\|.\|$ is the Euclidean norm, s is the scale, Θ is the rotation, and $(a + jb)$ is the translation, 1_k is a k-dimensional vector of ones.

This work presents a distance measure PDTW based on a pairwise synchronization between two time series by utilizing a combination of Procrustes distance and DTW to overcome the drawbacks of using DTW alone. First, we use DTW as an alignment approach and then use Procrustes as a distance measure. DTW is used to locate the best matching between two signals, whereas Procrustes is used to minimize the distance.

2.2. Classification

The simplicity of the k-Nearest Neighbors (kNN) method and its reasonable results made it a handy feature classifier. It predicts the new unlabeled query point by using the labels of training data based on their similarity measure. kNN classifier assigns a label for the test point to the majority label of the k-closet neighborhoods [29]. We found k = 3 is a reasonable value and the one we utilize in this paper.

3. Experimental Evaluation

We used three datasets to evaluate the proposed PDTW-kNN model on the public surgical data JIGSAWS [7], and our two data MU-EECS [30], and EM-Cric. The JIGSAWS is a minimally invasive surgical skill assessment working set consist of various fundamental surgical tasks. Each task performed by a surgical surgeon with a different proficiency degree; an expert surgeon who performs the da Vinci Surgical System (dVSS) more than 100 hours of training, a novice surgeon who practice less than 10 hours on dVSS, and an intermediate surgeon (practice on dVSS between 10 and 100 hours). A motion capture based on markers, a Vicon system is used to collect the data from a resident surgeon in the MU-EECS data. The surgeon presented a tracheostomy surgery performed the same procedure six times. The EM-Cric data includes data from four surgeons with different expertise levels who performed the Emergency Cricothyrotomy task. Each surgeon performs the task four times, where the wrist wearable sensors are used to capture both hand motions. More details about the three datasets in the following parts:

3.1. JIGSAWS Data

We evaluate the proposed PDTW-kNN method for surgical proficiency assessment on a public widely used JIGSAWS dataset [7]. Moreover, we use this dataset for direct comparisons with other state-of-the-art approaches for surgical skill evaluation. MIS surgeons performed many types of elementary procedures on Da Vinci robotic systems because it gives confidence, precision, and real-time feedback to improve overall surgical treatment for the patient in the operation room [31].

JIGSAWS dataset consists of kinematic and video data collected from surgical surgeons with various surgical robotic skills performing basic surgical training curricula. All surgeons were right-handed: two expert surgeons (E) with > 100 robotic surgical practice hours, four novice trainee surgeons (N) having < 10 practice hours, and four intermediate surgeons (I) reported between 10 and 100 surgical robotic experience practice hours. The dataset provides two types of data: video and kinematic records for each trail get done by a subject in each task. All the subjects were required to do three fundamental surgical tasks five times repetitively. In this work, we use only kinematic data captured as 76-dimensional time series at 30 Hz from the da Vinci Surgical System (dVSS) using its Application Programming Interface (API). The three elementary surgical tasks are identified as suturing (SU), knot-tying (KT), and needle-passing (NP). Figure 2 presented sample frames of the three surgical tasks achieved by a surgical surgeon and defined them as follows [7]:

- Suturing: the surgeon picks the needle up, first and advances it to the bench-top model toward the incision. Then, the subject stitches up the needle through a dot-marked tissue on one aspect of the incision and extracts it out from the corresponding dot-marked on the other part of the incision. Lastly, the surgeon passes it to the right-hand and repeats the same process till the surgeon gets four times in total.
- Knot Tying: the surgeon makes one tie after selecting one side of a stitch that is tied to an elastic tube connected by its rims to the surface of the bench-top model.
- Needle Passing: the surgeon selects the needle. Then, passes the needle from the right side to the left through 4 tiny metal hoops that are placed over the surface of the bench-top model.

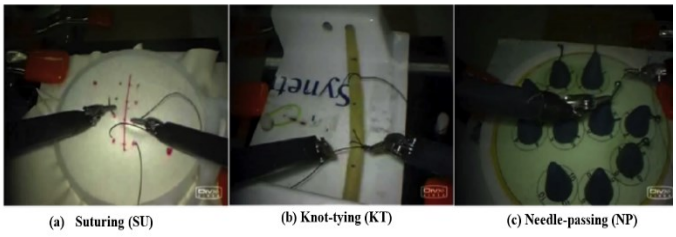

Figure 2: RMIS basic surgery tasks [7].

This dataset consists of a surgical manual annotation for the surgical skill of each trial. An annotating surgeon, with extensive robotic surgical experience, watched the entire trial and appointed a score based on a modified global rating score (GRS). GRS is the measure of the surgical technical skill of the surgeon who performed the trial. GRS presents the total score of six elements illustrated in Table I. Where each component rating scale is between 1 and 5 and the best with a higher total score [7].

Table 1: Elements of Global Rating Score (GRS) [7]

Element	**Rating scale**		
Respect for tissue	Force on tissue	Careful tissue handling	Consistent handling
Suture/needle handling	Poor knot tying	Majority appropriate	Excellent suture
Time and motion	Unnecessary moves	Efficient time/ unnecessary moves	Economy moves/ Max efficiency
Flow of operation	Frequent interrupted	Reasonable progress	Planned operation/ efficient transitions
Overall performance	Very poor	Competent	Superior
Quality of the final product	Very poor	Competent	Superior
Rating score	**1**	**3**	**5**
Min. score =	$\sum = 6$	Max. score =	$\sum = 30$

3.2. MU-EECS Vicon Data

In this dataset, a Vicon system and IR reflective markers were used synchronously to trace and visualize the arms movement of the surgeon while carrying out a surgical procedure. Ten IR reflective markers were placed in different positions on both surgeon's arms as displayed in Figure 3 (a). Also, we can see seven Vicon cameras were located inside the lab to capture the resident surgeon's motions. The MU-EECS includes data presented by a resident surgeon who performed the same tracheostomy surgical procedure six times repeatedly. The earliest three procedures repeat in a consistently appropriate manner, whereas the remaining practices were performed with inaccurately way. This working set was collected through a project at the Center for Eldercare and Rehabilitation Lab in the Dept. of EECS at the University of Missouri Columbia [30].

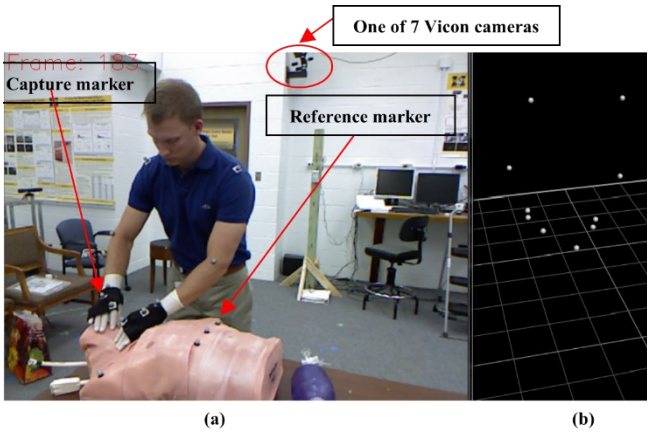

Figure 3: Tracheostomy surgery with Vicon Camera

3.3. EM-Cric Dataset

Emergency Cricothyrotomy (Cric) is a procedure for potentially lifesaving a human being under a high-stress situation, it happens when a person fails to restore enough oxygenation. Cric

is an incision through the skin and cricothyroid, which results in a better patient airway [32]. There are three main steps of the surgical Cric procedure skin incision, incision cricothyroid, and endotracheal tube placement membrane [33].

The EM-Cric dataset includes data from four surgical surgeons (subjects) who performed the Cric procedure with varying expertise levels to study skilled surgical human motion. Two residents reported as Novice (N) surgeon, one intermediate (I) surgeon, and one expert (E) surgeon, respectively. All surgeons are reportedly right-handed except one lefty hand. All surgeons perform the Cric procedure five times on a Trauma Man Surgical Simulator at the Medical Intelligent System Laboratory (MISL) in the Medicine School at the University of Missouri-Columbia. We placed the wristband sensors on both wrists of the surgeon's hands to capture the data, as shown in Figure 4. We use low cost synchronized data transmission MetaMotionR (MMR) sensors introduced by MbientLab. MMR is a 9-axis IMU wearable device that provides continuous monitoring of movement and real-time sensor data [34].

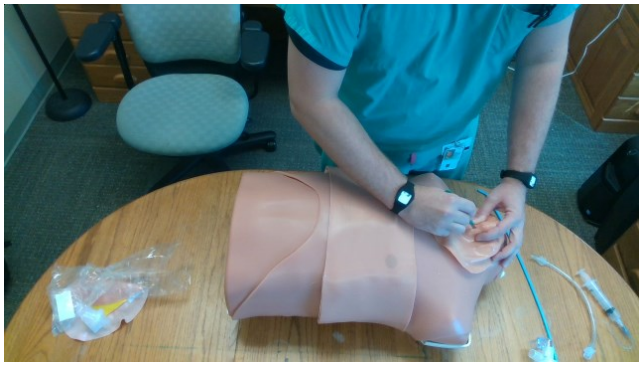

Figure 4: Cric surgical operation on TraumaMan Simulator by a medical surgeon.

The data was conducted for a total of three male right-handed, and one female left-handed participants with different expertise levels were recruited for this study. Two MMR sensors were used for the Cric procedure task, one attached to each wrist of the surgeon's hand. The captured data consists of three-dimensional acceleration with respect to time for each accelerometer, and result in 6-dimensional time series for both sensors. For this study, we use only raw accelerometer data which range was set to ±16 g. The sampling rate of data collection was set to 100Hz.

3.4. Performance Evaluation

We used different cross-validating schemes to evaluate our skill assessment framework on both kinematic and accelerometer data to compare our results with other approaches.

- Leave-One-Trial-Out (LOTO): For each surgical task, training all the trials except one i-th trial reserved for testing ($i = 1,..,N$). N is the total number of trials in a task.
- Leave-One-Supertrial-Out (LOSO): Different from LOTO setup, where we created five folds ($j = 1,2..5$). The j-th fold combines all the j-th trials from all the surgeons for a given surgical task. Then, we repetitively training on four sets and keeping a single set for testing and reporting the average classifying results. The fold j-th is known as supertrial j-th. In this scheme, the robustness of a technique can be assessed by keeping a supertrial out each time [7]. Also, repeating the task in a row can possibly impact the performance of the surgical apprentice in terms of boredom or tiredness, hence keeping the supertrial out perhaps catch that effect on the surgeons.

To evaluate the performance of our proposed technique and to quantitatively compare with other methods, we used the mean accuracy of surgical classification for each output class on the data-driven to validate the performance. The average accuracy, defined in (10), is the percentage of the sum of accurately predicted (TP+TN) over the total number of predictions (TP+TN+FP+FN) [35]:

$$\mathrm{ACC} = \frac{\mathrm{T_P}+\mathrm{T_N}}{\mathrm{TP+TN+FP+FN}} \tag{10}$$

where T_P, T_N, F_P, and F_N represent the number of true positive (predicted correctly belong to the target class), true negative (correctly classified not belong to the target class), false positive (incorrectly predicts to the target class), and false negative (incorrectly predict not belong to the class level) respectively [35].

4. Results and Discussions

In this part, the proposed approach and evaluation metrics described in the preceding sections were evaluated on kinematic and accelerometer data. Also, the results for all the datasets that were explained previously were reported in the following sections, respectively.

4.1. JIGSAWS Dataset

For JIGSAWS data, we perform two sets of experiments for the LOSO validation set up to identify the three expertise levels (E, I, and N) on our proposed approach. For the first assortment, we made use of all the 76-dimensional movement features of the time series. Whilst, in the second set we utilized just the coordinates features (x, y, z) of the two hands.

Figure 5 (a) illustrates the comparison of classification accuracy for surgical expertise levels versus k (the number of neighborhoods) in each task using all kinematic information. For the LOSO scheme, the improvement in accuracy for almost all cases of k of our kNN classifier based PDTW for all surgical tasks. e.g., the mean accuracy for all tasks at k = 3 is 95.7%. Also, kNN-PDTW provide an advantage over the traditional method (DTW) with a reduction in sensitivity to changing the number of neighbors (k) in k-NN.

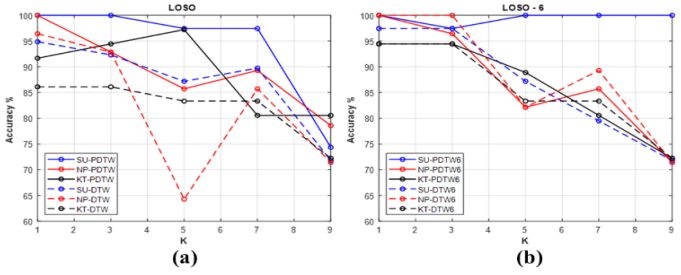

Figure 5: Accuracy of the proposed approach using PDTW and DTW as a function of k.

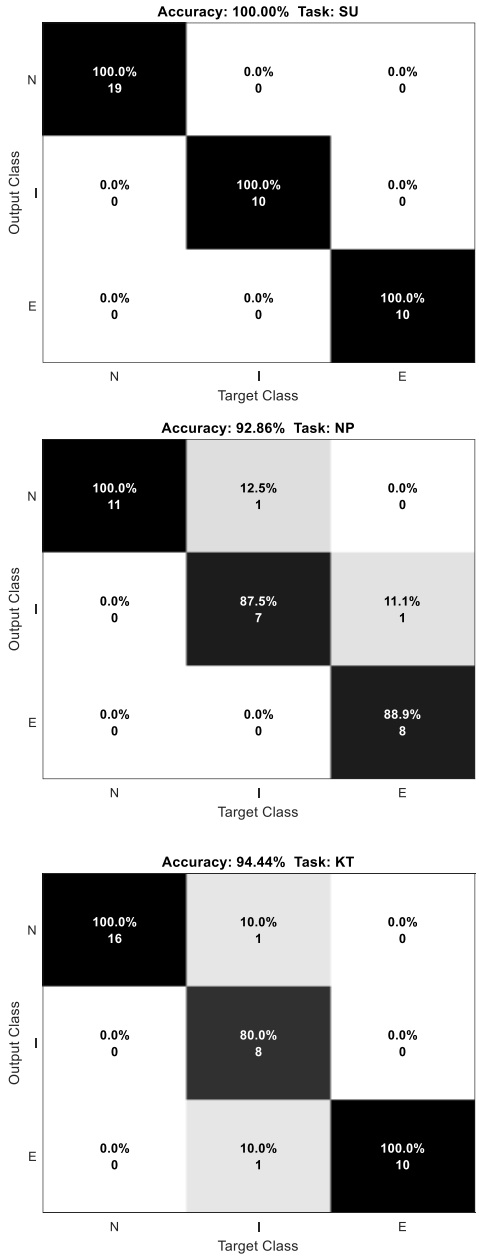

Figure 6: kNN-PDTW Confusion matrix of the three tasks SU, NP, KT for LOSO at k=3.

We also perform another experiment by using only 3D location information of the two hands for the LOSO scheme. Some interesting intuitions results can be seen in Figure 5(b). The accuracy results of the proposed kNN-PDTW6 using the Cartesian coordinates almost achieved the same results as using all the 76-dimensional motion data. This can be explained by the fact that Procrustes analysis works on the similarity of shapes and the motion data are traces in three dimensions space, which encourages us to use the wearable sensors later.

For a further comprehensive comparison, the confusion matrices result for each task is shown in Figure 6 at k=3. For the suturing task, surgeon expertise levels are 100% correctly classified. However, for the other tasks, the misclassifying happened when distinguishing between intermediate level and other levels which in turn reduced the average accuracy to about 94% and 93% for knot tying and needle passing tasks, respectively. We must put into our perspective that each surgeon performs the task in a different style from other surgeons, even within the same expertise level regardless of the hours spent on practice. Because individual surgeons like to improve their proficiencies following their mentor. Thus, small differences between an intermediate surgeon and an expert make the classifier to introduce an error to recognize their skill levels and vice versa. The same case between intermediate and novice surgeons happened.

Another interest intended of our analysis, that we calculate the pairwise PDTW distance inside a group of expert-expert, expert-intermediate, and expert-novice surgeons, separately for each task. Figure 7 illustrates the boxplot of each group distance in each task. From the results, it is clear that the smallest distance is among expert surgeons, and then between expert-intermediate surgeons followed by the expert-novice group for each task. Also, we can see that the differentiating among expert-intermediate surgeons is more complicated in needle-passing than other tasks. one explanation is the needle-passing might be more challenging to learn or more complicated than suturing or knot tying. This might be related to the complication level of the task as can be seen in Figure 6 for the needle-passing task where an expert surgeon classified as intermediate surgeon mistakenly.

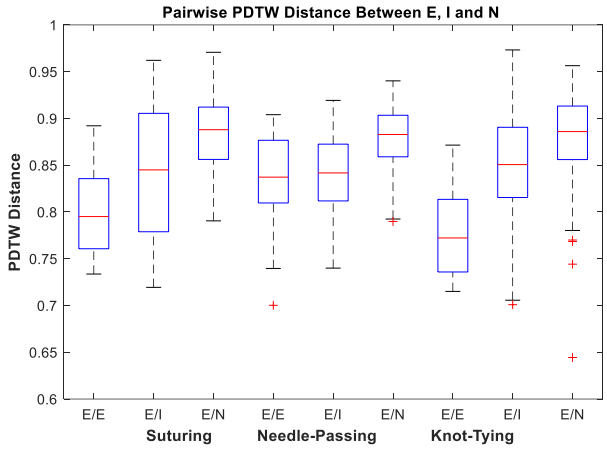

Figure 7: PDTW-distance within E/E, E/I, and E/N surgeons in each task.

Table 1 shows the classification accuracy results of our proposed skill assessment for the JIGSAWS dataset using the kinematic data only. Also, we report the state-of-the-art results for comparative intent under the LOSO validation scheme for each task separately. The results show that the proposed kNN-PDTW properly recognizes the surgeon skill levels and matched the work from CNN [36] for suturing. From Figure 7 we can see that it is straightforward to differentiate between the expertise levels with the help of using PDTW measure. Additionally, the NN classifier learned the dynamic information which already comes from various motion patterns of the surgeons that might benefit this result. In knot-tying, our proposed kNN-PDTW approach outperforms both CNN [36] and Deep Learning [12] approaches in terms of accuracy. Also, our results were near the CNN+LSTM+SENET method [18]. Our results were improved more for suturing and knot-tying tasks than the needle-passing

task, and we did slightly better than [12] in this task. The small distinctions between intermediate surgeons with other surgeons in this task illustrated in Figure 7 might explain the less performance on the needle-passing task. Furthermore, we can notice from Table I that no technique is suitable for the three tasks. In other words, an integration methodology of various approaches is needed for surgical proficiency assessment purposes for these tasks.

Table 2: Skill Assessment Classification Comparative of kNN-PDTW Performance using LOSO for JIGSAWS Data.

Approach		Accuracy		
		SU	NP	KT
Farad [15]	kNN	89.7%	-	82.1%
	LR	89.9%	-	82.3%
	SVM	75.4%	-	75.4%
Wang [12]		93.4%	89.8%	84.9%
Forestier [13]		89.7%	96.3%	61.1%
Fawaz [36]		**100**%	**100**%	92.1%
Anh [18]		98.4%	98.4%	**94.8**%
kNN-PDTW (proposed)		**100**%	92.8%	94.4%

As mentioned previously in section 3.1, the modified global rating score measures the surgical technical skill done by the annotation surgeon for the entire trial provided in the JIGSAWS dataset. Figure 8 presents the boxplot of the surgeons' GRS scores for each task. We can see from this figure, the consistency of the expert surgeons compared to the novice and intermediate surgeons in all tasks. Where the lowest variance the expert surgeons have ultimately implied their steadiness. Another interesting viewpoint from Figure 8, that we can see the scores challenge to differentiate among the surgeon's proficiency in the needle-passing task, which produces the misclassifications. One more thing to be observed in Figure 8, some intermediate subjects score better than expert subjects. This means that these surgeons might be eligible to be in a higher skill level or position.

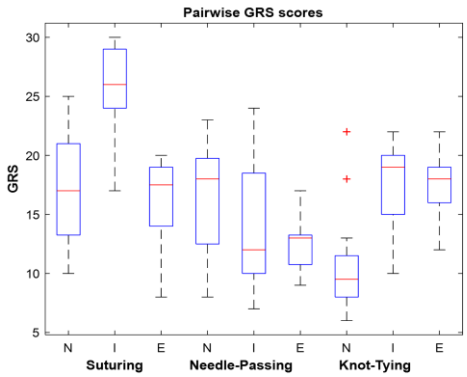

Figure 8: Boxplot of GRS scores for each task.

4.2. MU-EECS dataset

We experiment on the tracheostomy dataset to classify the trial level as either *Good* or *Bad*. In this experiment, we calculate the pairwise PDTW distance among the six trials that operated by a resident surgeon [30]. Figure 9 presents the resulting distance of this experience for the MU-EECS dataset, where the yellow color is the farthest and the closer trials to each other are in darker blue.

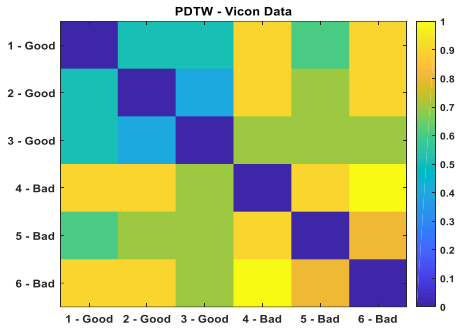

Figure 9: PDTW distance matrix for MU-EECS data.

Overall, the *Good* trials, which are the first three trials in Figure 9, has a similarity less than or equal to 0.5. e.g., about 0.3 is the difference between trials 2 and 3. On the other hand, the pairwise distance between *Bad* procedures, the last three trials, is greater than 0.7 in distance to each other. Also, we can see those *Good* procedures are nearly 0.7 far away from *Bad* trials except among trial-Good 1 and trial-Bad 5 about 0.55 difference.

Another insight from Figure 9, it is straightforward to cluster the trials into *Good* (the upper left corner) and *Bad* (in the lower right corner). That means the PDTW distance helps accurately to identify between the trials in this task where each group looks to cluster together. Finally, the boxplot of the PDTW measure among the Good and Bad trials separately is presented in Figure 10. In this figure and from a statistical viewpoint comparison, the mean and variance of the *Good* procedures ($\mu_{G-G} = 0.12,\ \sigma_{G-G} = 0.08$) is less than the *Bad* procedures ($\mu_{B-B} = 0.21,\ \sigma_{B-B} = 0.16$) which is consistent along with prior results.

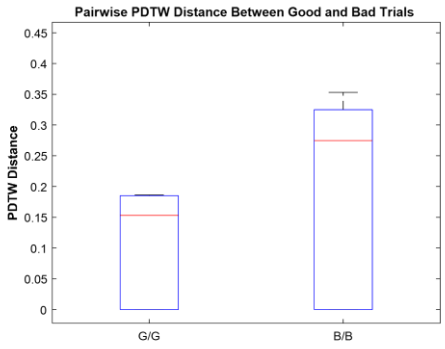

Figure 10: Boxplot of PDTW distance for MU-EECS dataset.

4.3. EM-Cric dataset

For the EM-Cric dataset, we performed two sets of cross-validation schemes, the LOTO for the trial level and the LOSO to identify the surgical proficiency levels (Expert, Intermediate, or Novice) of the subjects. As we mentioned previously in section 3.3, this dataset includes accelerometer data collected from four surgeons (expert, intermediate, and two novices) who performed the same task five times repetitively.

Before evaluating the classification accuracies, we calculate the pairwise distance among all the collected trials. Figure 11 (a) and (b) illustrate pairwise distance matrices comparison between DTW and PDTW measures, respectively. The first five trials represent the expert surgeon procedures, the second five stand for the intermediate surgeon trials, and the remaining ten trials are for the two novice surgeons, all performing the same task. Where the similar performances made by participants are indicated in strong blue squares in this figure. Also, the three separate square blocks in Figure 11 (b) give a visual insight for the possibilities of clustering expertise levels where the task is performed by different surgeons for this data using only the accelerometer data. Also, we can notice from this figure that PDTW distance separates well between expertise levels better than using DTW distance alone. The results in Figure 11 (b) shows that the expert surgeon has a dissimilar pattern to both intermediate and novice surgeons. Moreover, novice surgeons themselves are quite like each other.

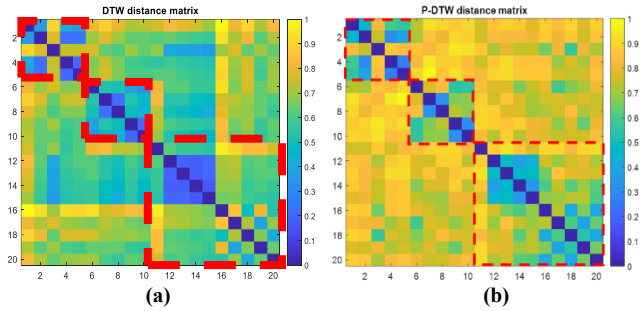

Figure 11: The pairwise distance for each trial on EM-Cric using (a) DTW and (b) P-DTW

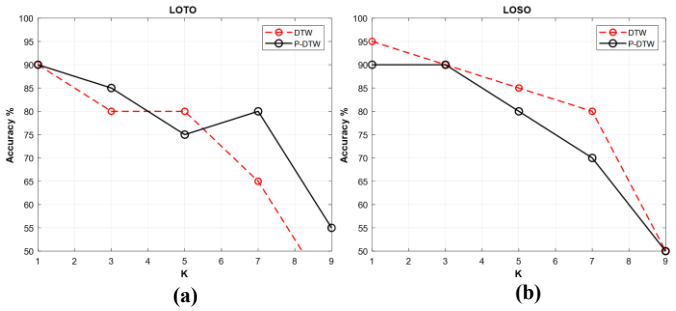

Figure 12: Classification accuracy as a function for k (a) LOTO and (b) LOSO cross-validation for Cric data

First, we performed experiments to compare how DTW and PDTW perform for classifying surgeon levels on Cric data using both LOTO and LOSO configurations. Figure 12 presents comparisons of the classification accuracy results of the proposed model for different values of K (number of neighbors) using LOTO and LOSO cross-validations, respectively. Figure 12 (a) shows that the results of our method based on PDTW performs better compared to using only DTW distance. These results indicate that our approach can identify the surgical skill levels well at trial levels because it utilizes the Procrustes analysis. Secondly, Figure 12 (b) presents the kNN-PDTW performance for the LOSO setup for the Cric dataset. The kNN based DTW approach performs slightly better for the accelerometer data. Whereas our approach results were improved, and the performance was reasonably well and still having a higher classification accuracy of 90% at k = 3.

Figure 13 shows the confusion matrix of our kNN based PDTW for surgeon expertise at k = 3 for Cric data using LOSO configuration. We can see that the intermediate surgeon was classified correctly, whereas both expert and novice surgeons were misclassified in one trial. From Figure 11 (b), we can notice that there is one trial (#3) from the expert surgeon that seems far from other trials with Expert trials and the same for novice surgeons with the trial (#11) in the same figure. The average classification accuracy was 90%.

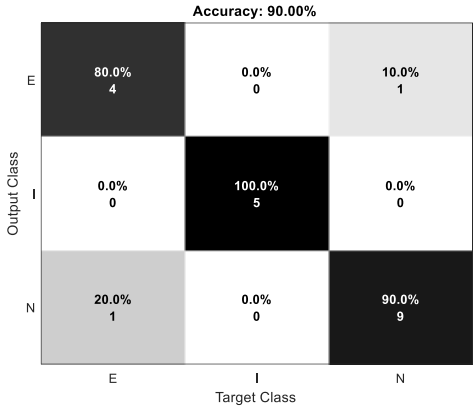

Figure 13: kNN-PDTW Confusion matrix for LOSO at k=3 for Cric data

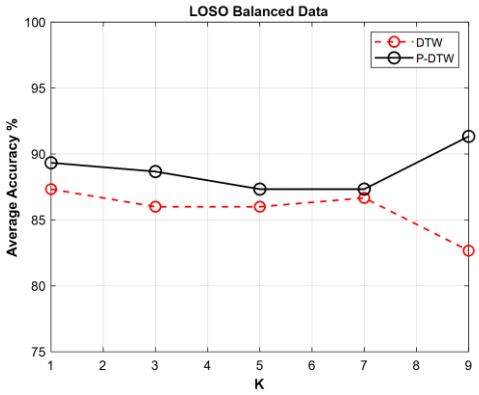

Figure 14: Balanced data classification results for the Cric data

Lastly, for a more thorough comparison, we perform another experiment for Cric data by using balanced data and evaluating using LOSO with a k-fold cross-validating scheme. The balanced data was obtained by having equal trials from each surgeon level. The reason we chose the balanced data experiment because we had two novice surgeons, one expert, and one intermediate surgeon. In this conduct experiment, we pick five trials randomly from a total of ten novice surgeon's trials and put them together with other trials from the expert surgeon and the intermediate surgeon trials. Then repeat the process ten times and report the average classification accuracy. Figure 14 shows the comparison classification accuracy as a function of k between PDTW and DTW based kNN classifier. Furthermore, Figure 15 presents the confusion matrix of kNN-PDTW predictions of the surgical skill classes. We can see from both above figures that the average accuracies of using PDTW much better than using DTW for all

values of k. Also, our approach using balanced data achieved average classification accuracy about 3% higher than using unbalanced data. the balancing data helps classified the novice surgeon's skill correctly with 100%.

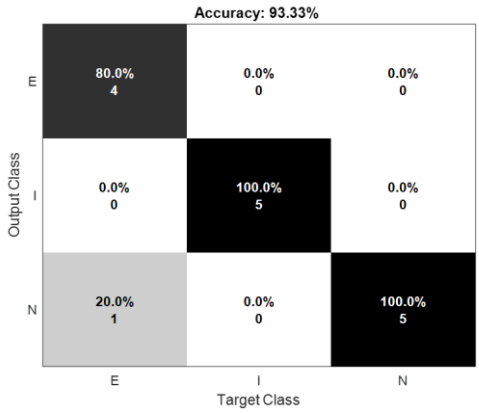

Figure 15: Balanced data confusion matrix for the Cric data

5. Conclusions

In this paper, we define a new surgery skill distance measure PDTW. It incorporates the exploration for best alignment using DTW and the similarity measure using Procrustes distance among two multidimensional time series. We show that the proposed framework based PDTW can enhance the overall performance for surgical proficiency evaluation. We attain an average accuracy of 97% for the JIGSAWS dataset and the results outperform most state-of-the-art methods using kinematic data and are comparable to techniques based on deep schemes.

Also, here we have examined the use of wearable motion sensor devices in proficiency assessment to achieve an entirely objective evaluation. Although our results are encouraging, there are quite a few limitations. The number of subjects is relatively small, not as desired. Furthermore, only one surgical task the subjects were asked to work on and there is no break between the trials which might impact the performing of the trials. Despite the limitations, our results indicate that PDTW distance can be used by classifying techniques to categorize the expertise levels accurately. In the future, we plan to increase the number of participants with a variety of expertise which might have the potential to give more information and robustness to our method. Also, more tasks to be utilized instead of only a given surgical task. Furthermore, consider using another or a combination of classifiers to improve the overall classification accuracy for skill assessment.

References

[1] S. Albasri, M. Popescu, J. Keller, "A novel distance for automated surgical skill evaluation," in 2019 7th E-Health and Bioengineering Conference, EHB 2019, 2019, doi:10.1109/EHB47216.2019.8970029.

[2] G. Arbelaez-Garces, D. Joseph, M. Camargo, N. Tran, L. Morel, "Contribution to the objective assessment of technical skills for surgery students: An accelerometer based approach," International Journal of Industrial Ergonomics, **64**, 79–88, 2018.

[3] M.J. Fard, S. Ameri, R.B. Chinnam, A.K. Pandya, M.D. Klein, R.D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," Lecture Notes in Engineering and Computer Science, **2225**, 433–437, 2016.

[4] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," International Journal of Computer Assisted Radiology and Surgery, **13**(3), 443–455, 2018.

[5] F. Lalys, P. Jannin, "Surgical process modelling: A review," International Journal of Computer Assisted Radiology and Surgery, **9**(3), 495–511, 2014, doi:10.1007/s11548-013-0940-5.

[6] Intuitive Surgical, 2020, doi:https://www.intuitive.com/en-us.

[7] Y. Gao, S.S. Vedula, C.E. Reiley, N. Ahmidi, B. Varadarajan, H.C. Lin, L. Tao, L. Zappella, B. Béjar, D.D. Yuh, C.C.G. Chen, R. Vidal, S. Khudanpur, G.D. Hager, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 1–10, 2014.

[8] C.E. Reiley, H.C. Lin, D.D. Yuh, G.D. Hager, "Review of methods for objective surgical skill evaluation," Surgical Endoscopy, **25**(2), 356–366, 2011.

[9] A. Zia, I. Essa, "Automated surgical skill assessment in RMIS training," ArXiv, **13**(5), 731–739, 2017.

[10] M. Jahanbani Fard, Computational modeling approaches for task analysis in robotic-assisted surgery, Ph. D. Thesis, Wayne State University, 2016.

[11] H.C. Lin, I. Shafran, D. Yuh, G.D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," Computer Aided Surgery, **11**(5), 220–230, 2006, doi:10.1080/10929080600989189.

[12] Z. Wang, A.M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," ArXiv, **13**(12), 1959–1970, 2018.

[13] G. Forestier, F. Petitjean, P. Senin, F. Despinoy, A. Huaulmé, H.I. Fawaz, J. Weber, L. Idoumghar, P.A. Muller, P. Jannin, "Surgical motion analysis using discriminative interpretable patterns," Artificial Intelligence in Medicine, **91**(July), 3–11, 2018, doi:10.1016/j.artmed.2018.08.002.

[14] L. Tao, E. Elhamifar, S. Khudanpur, G.D. Hager, R. Vidal, "Sparse hidden Markov models for surgical gesture classification and skill evaluation," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **7330 LNCS**, 167–177, 2012, doi:10.1007/978-3-642-30618-1_17.

[15] M.J. Fard, S. Ameri, R. Darin Ellis, R.B. Chinnam, A.K. Pandya, M.D. Klein, Automated robot-assisted surgical skill evaluation: Predictive analytics approach, International Journal of Medical Robotics and Computer Assisted Surgery, **14**(1), 2018, doi:10.1002/rcs.1850.

[16] M.J. Fard, S. Ameri, R.D. Ellis, "Skill Assessment and Personalized Training in Robotic-Assisted Surgery," CoRR, 2016.

[17] Z. Wang, A.M. Fey, "SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE: 1793–1796, 2018.

[18] X.A. Nguyen, D. Ljuhar, M. Pacilli, R.M. Nataraja, S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," Computer Methods and Programs in Biomedicine, **177**, 1–8, 2019, doi:10.1016/j.cmpb.2019.05.008.

[19] Y. Sharma, V. Bettadapura, T. Ploetz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, I. Essa, "Video Based Assessment of OSATS Using Sequential Motion Textures," in Fifth Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), Georgia Institute of Technology, 2014.

[20] Y. Sharma, T. Plötz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, I. Essa, "Automated surgical OSATS prediction from videos," in 2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014, IEEE: 461–464, 2014, doi:10.1109/isbi.2014.6867908.

[21] N. Ahmidi, M. Ishii, G. Fichtinger, G.L. Gallia, G.D. Hager, "An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data," in International forum of allergy & rhinology, Wiley Online Library: 507–515, 2012.

[22] A.L. Trejos, R. V. Patel, M.D. Naish, A.C. Lyle, C.M. Schlachta, "A sensorized instrument for skills assessment and training in minimally invasive surgery," in Journal of Medical Devices, Transactions of the ASME, IEEE: 965–970, 2009, doi:10.1115/1.4000421.

[23] F. Gómez-Vela, F. Martínez-Álvarez, C.D. Barranco, N. Díaz-Díaz, D.S. Rodríguez-Baena, J.S. Aguilar-Ruiz, "Pattern recognition in biological time series," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **7023 LNAI**, 164–172, 2011, doi:10.1007/978-3-642-25274-7_17.

[24] E.J. Keogh, M.J. Pazzani, "Derivative dynamic time warping," in Proceedings of the 2001 SIAM international conference on data mining, SIAM: 1–11, 2001.

[25] J.T. Kent, "New directions in shape analysis," The Art of Statistical Science,

115, 1992.
[26] K. V Mardia, P.E. Jupp, Directional statistics, John Wiley & Sons, 2009.
[27] M.B.B. Stegmann, D.D.D. Gomez, "A brief introduction to statistical shape analysis," Informatics and Mathematical …, (March), 1–15, 2002.
[28] I.L. Dryden, K. V. Mardia, Statistical shape analysis, with applications in R: Second edition, John Wiley & Sons, 2016, doi:10.1002/9781119072492.
[29] J.M. Keller, D. Liu, D.B. Fogel, Fundamentals of computational intelligence: neural networks, fuzzy systems, and evolutionary computation, John Wiley & Sons, 2016.
[30] M. Popescu, C.J. Cooper, S. Barnes, "Automated Operative Skill Assessment Using IR Video Motion Analysis," in AMIA, 2014.
[31] B.J. Dlouhy, R.C. Rao, "Surgical skill and complication rates after bariatric surgery," The New England Journal of Medicine, **370**(3), 285, 2014.
[32] M.G. Katos, D. Goldenberg, "Emergency cricothyrotomy," Operative Techniques in Otolaryngology-Head and Neck Surgery, **18**(2), 110–114, 2007.
[33] A. MacIntyre, M.K. Markarian, D. Carrison, J. Coates, D. Kuhls, J.J. Fildes, "Three-step emergency cricothyroidotomy," Military Medicine, **172**(12), 1228–1230, 2007.
[34] INC, MBIENTLAB, 2020, doi:https://mbientlab.com/metamotionr/.
[35] T. Fawcett, "ScienceDirect.com - Pattern Recognition Letters - An introduction to ROC analysis," Pattern Recognition Letters, **27**(8), 861–874, 2006.
[36] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in arXiv, Springer: 214–221, 2018.