ADVANCES IN AUTOMATED SURGERY SKILLS EVALUATION

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

SAFAA ALBASRI

Dissertation Supervisors,

Dr. Mihail Popescu and Dr. James Keller

July 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ADVANCES IN AUTOMATED SURGERY SKILLS EVALUATION

presented by Safaa Albasri,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Professor James M. Keller

_____

Professor Marjorie Skubic

_____

Professor Mihail Popescu

_____

Associate Professor Salman Ahmad

DEDICATION

This dissertation is proudly dedicated ……

To my Beloved Parents,

To my Sister and Brothers,

To my all Friends,

Thanks for your Endless Love, Sacrifices, Prayers, Support and Guidance.

ACKNOWLEDGEMENTS

First, praise to Allah for giving me the strength to complete this work at the University of Missouri-Columbia.

I would like to express my deepest gratitude to my advisor, Dr. Mihail Popescu, for his valuable guidance, encouragement, constructive suggestions, and his assistance throughout the final preparation of this work. I am truly grateful especially throughout writing this dissertation. I really appreciate him for granting me especially during the year of the pandemic. I would also like to thank my co-advisor Dr. James Keller for all his assistance, useful discussions and especially during the difficult stages of doing this study. It was a great honor for me to work with my both supervisors; the work would not have been finished without their support.

Also, many thanks go to the rest of my committee members: Dr. Marjorie Skubic and Dr. Salman Ahmad for their valuable feedback on the dissertation.

My sincere thanks go to my dear friend and lab mate Hasanain Al-Sadr for his assistance, his insightful comments, suggestions, and for a cherished time spent together in the lab and daily social life.

Special thanks go to my friends Mohanned A-Gharawi, Hayder Yousif, Saif Altai, and Zakariya Oraibi; without their enormous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

My gratitude extends to the Higher Committee of Education Development in Iraq (HCED) for the funding opportunity to undertake my studies at the Department of Electrical Engineering, University of Missouri.

In the loving memory of my father, who has passed away in 2004, I just wished to make him see me when I graduate and make him feel proud of me. I wish he were here to share my happiness while writing these words.

Distinctive gratitude goes to my MOM, who inspired and supported me during this journey. She had sacrificed and walked next to me throughout the entire way. I would not be able to finish this work without her encouragement and support until the end. You have been constantly in my heart and will always be!

Finally, I want to thank my entire family, my sister, my brothers for their support, prayers, and words of love during the difficult times that I had.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ADVANCES IN AUTOMATED SURGERY SKILLS EVALUATION

SAFAA ALBASRI

Dr. Mihail Popescu, Dissertation Supervisor

Dr. James Keller, Dissertation Co-advisor

ABSTRACT

Training a surgeon to be skilled and competent to perform a given surgical procedure, is an important step in providing a high quality of care and reducing the risk of complications. Traditional surgical training is carried out by expert surgeons who observe and assess the trainees directly during a given procedure. However, these traditional training methods are time-consuming, subjective, costly, and do not offer an overall surgical expertise evaluation criterion. The solution for these subjective evaluation methods is a sensor-based methodology able to objectively assess the surgeon's skill level.

The development and advances in sensor technologies enable capturing and studying the information obtained from complex surgery procedures. If the surgical activities that occur during a procedure are captured using a set of sensors, then the skill evaluation methodology can be defined as a motion and time series analysis problem. This work aims at developing machine learning approaches for automated surgical skill assessment based

on hand motion analysis. Specifically, this work presents several contributions to the field of objective surgical techniques using multi-dimensional time series, such as 1) introduce a new distance measure for the surgical activities based on the alignment of two multi-dimensional time series, 2) develop an automated classification framework to identify the surgeon proficiency level using wrist worn sensors, 3) develop a classification technique to identify elementary surgical tasks: suturing, needle passing, and knot tying , 4) introduce a new surgemes mean feature reduction technique which help improve the machine learning algorithms, 5) develop a framework for surgical gesture classification by employing the mean feature reduction method, 6) design an unsupervised method to identify the surgemes in a given procedure.

CHAPTER 1 **Introduction**

1 . 1 Statement of the Problem

The skills of surgeons receive increasing attention from healthcare organizations due to the possible increase of financial and legal expenses. Factors such as teaching, training, and practice, that vary from one surgeon to another, significantly influence the surgeons' skill levels. Existing evidence shows that more surgical training and objective assessment can improve patient care outcomes and provide faster recovery [1]. Several factors determine the surgeon proficiency levels, including cognition capabilities, decision-making and dexterity skills. Some of these factors are learned in Medical School, whereas others depend on the apprentice training [2, 3].

"See one, do one, teach one," the traditional surgical training approach by William Halsted, consists in a relationship between the senior surgeon and the apprentice [4]. This master-trainee surgical pattern involves subjective assessment of surgery skills, which can produce bias and subjectivity in the surgical evaluation [5]. Also, it requires the senior surgeon to directly attend the surgical procedure performed by the apprentice as an observer, which produces extra expenses for the health system [6]. To address this problem, various methodologies for automated evaluation of surgical skills have been proposed recently. The need for objective surgical skill assessment methods in surgical training was mentioned by multiple authors in the medical literature [7]. These methods are objective, more accurate, affordable, and provide analytic information about surgeon surgical skills within various surgery environments [5].

One methodology for surgical training and assessment, is based on virtual reality simulator systems or benchtop camera-based versions (http://medvisionsim.com). Such systems help surgeons to practice and develop skills in a safe and secure environment before trying them on living patients [1, 8]. These training methods that offer quantifiable and statistical measurements for modeling and evaluating the surgeon skills, can benefit medical curriculum and hospital surgeon training [1]. The challenge then becomes to evaluate the surgical process using the quantitative data from these advances to understand the surgeon movements during a given surgical task. Robotic Minimally Invasive Surgery (RMIS) is one solution to overcome this challenge to enhance the effectiveness at the operating room (OR) [9]. Da Vinci robotic surgical system is an example to offering a data-driven that has the potential to evolve the surgeon skills [10]. Kinematic and video data represent valuable information sources about the surgeon's motion during a surgical task from da Vinci systems and enable them to analyze their performance using data-driven methods [1, 11]. Recently, wearable sensor devices another instance of the recoded data that added further information used for surgical events regarding surgeon movements and unknown patterns during surgery by statistical and machine learning models [12, 13].

The automated surgical skill methods are recently gaining more attention and seeing decent development in surgery curricula [5]. Employing the machine learning approaches to data obtained from robotic surgery systems and wearable sensor devices inspired the researchers to develop and investigate automated models to evaluate and assess surgeon professional knowledge. Also, it might help improve prospective coaching apprentices [12, 14, 15]. Current advances in machine learning and computer vision techniques help understand and analyze surgeon motions during a surgery task. Also, these methods

provide the chief surgeon a source of information to evaluate the trainee's skills and feedback to the trainees about their performance while doing a particular procedure [12].

In this work, we elucidate some of the obstacles and challenges in automated surgical skill assessment by evolving novel approaches to address these obstructions. The escalating needs for objective surgery skill evaluation motivated us to develop techniques based on machine learning. This work aims to assess a surgeon's skill based on a multi-dimensional time series to tackle the problems of surgical aptitude levels, surgical tasks classification, and surgemes recognitions to be achieved.

1 . 2 Background

Automated approaches to assess surgeon dexterity in surgery are recently rising in medical education fields. It is not straightforward to build a computerized model that duplicates the expert surgeon evaluation for trainee surgeons. However, there are continuing works taking place toward an objective surgical assessment.

We found many excellent reviews and research studies for automated surgical assessment and its importance across numerous surgery environments. We provide a survey about earlier and current key works accomplished on objective surgical methods and available datasets. Most of the works breakdown into one of three surgical categories: 1) Surgical Skills, 2) Surgical Tasks, and 3) Surgemes segmentation. The following gives more description about each group.

1. 2. 1 Surgical Skill Assessment

Traditional surgical skill assessment is based on competency checklists measure, where a senior faculty surgeon usually observes a surgical trainee complete a given surgical task [6, 16]. However, this approach is susceptible to the assessor and might be biased. Alternative methods suggest using surgical movement data (e.g., kinematic and video) that is recorded by surgical robotic or surgical simulator systems as for examples, and then using the statistical information (e.g., speed, time, distance, number of hand movements) to evaluate the surgeon overall performing [6]. But these global statistical methods [17-20] provide an average performance for a given measurement, and it is unable to provide feedback where the trainee needs to practice more [6].

Different earlier works focused on the automated surgical assessment have seen good progress. These methods considered the surgeon movement using either: 1) kinematic information recorded by a robotic surgical system, 2) video records, or 3) wearable sensor data.

Pioneer attempts to lead the way using data-driven from minimally invasive surgery for skill assessment by utilizing global statistics measures. Early work [19] measures the magnitude of force and torque in addition to the time of completion through laparoscopic surgery to evaluate the skill level between novice and expert surgeons. An alternative approach used the mean values of the number of movements and time to task completion to measure the talent among four groups of surgeons [21]. Another method includes time, distance, speed, phase, and curvature to measure the difference between novice and expert surgeons [18]. Although these predefined statistical features methods are easy to implement and produce substantial results, they are particularly time-consuming, and they

give the overall metrics without local feedback information about the surgical training task [6, 22].

Alternatively, one early approach [23] used the Hidden Markov Models (HMM) based on surgeon 3D motion traces to identify the surgical expertise between two levels, expert and novice. Another method [24] suggested using the sparse hidden Markov (S-HMM) to classify the surgical skill and gestures. For expertise classification, they learn an S-HMM model on kinematic data, then assigning a test trial to one model (Expert, intermediate, or novice) with the highest likelihood.

These approaches are structured-based and depend on the number of training samples, tuning parameters, and it takes massive pre-processing. This model needs complicated preprocessing [25] and leads to low performance with a low number of samples [24]. Another method was proposed by [25] 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 a suturing surgical task. They extended their work to include eight global movement features (GMF) in [26]; they applied LR, SVM, and kNN classifiers to distinguish between the previous expertise levels for suturing and knot tying surgical tasks. In [27], 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, [28] 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 [13] 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 [29] 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 [15] for surgical expertise evaluation. Another parallel deep learning approach was proposed in [22] by combining the LSTM recurrent network and CNN to indicate the skill levels. Additionally, recent studies have suggested approaches that use motion from videos [30, 31] and wearable sensors to evaluate surgical skills [32, 33]. 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 [34]. Also, an entropy-based features technique that utilizes both video and accelerometer data proposed for surgical skill assessment [35]. 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 have been done usually by a chief surgeon, i.e., consuming a considerable 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 develop a new distance measure might have an advantage to a more robust and accurate assessment framework.

1. 2. 2 Surgical Task Recognition

A surgical task is described as a sequence of movements to accomplish a medical procedure [12]. Preliminary surgical task classification approaches focus on identifying high-level tasks (HLT) in the operating room (OR) environment using Hidden Markov Models for surgical workflow analysis [36]. A current technique [37] applied dynamic time warping (DTW) and kNN classifier to distinguish between three training procedures: suturing, needle passing, and knot tying. Another work [38] investigated the use of derivative DTW (DDTW) as a similarity measure with the kNN classifier for the three fundamental surgical tasks. A recent deep-learning approach [29] was designed to join a CNN and gated recurrent unite (GRU) architecture for recognizing surgeon skill levels and surgical tasks simultaneously. The beforehand approaches used the kinematic data for analysis. In contrast, [39] applied the long short term memory (LSTM) and convolution neural network (CNN) on video information to classify the surgical tasks and the surgical gestures (surgemes). Even these approaches could find underlying surgical tasks; however, they are limited and have some common flaws such as time-consuming, heavyweight

computation load and architecture, high pre-processing data, and significant human interaction.

1. 2. 3 Surgemes Segmentation and Recognition

During a given surgical task, a surgical activity component is characterized as a surgical gesture or in medical language well-defined as surgemes [12]. Recently, studying surgemes has raised attention for many developers in the medical training area after using the advances of robotic surgical simulators, which are based on sensors that enable the capture of the surgeon motions [28]. Numerous methods for surgemes segmentation have been proposed, which can be divided into supervised and unsupervised methods depending on the technique used to identify a given surgical task's components. Additionally, the existing surgical activities segmentation approaches categorize their type of information sources into four groups: kinematic, video, wearable sensors, or a combination of video and one of the other data.

1. 2. 3. A Surgemes Supervised Approaches

The task, in this case, is to locate the surgemes to the right categories that it belongs to in a given surgical undertaking, where the boundaries of each surgeme are predefined. Several pioneered studies attempted to decompose a surgical task into a set of surgical activities. One early work [24] used the sparse HMM with learning the dictionaries of each surgeme to classify and represent a new action. A variations technique of HMMs and conditional random fields based on temporal methodologies for joints segmentation and features-based strategies (e.g., linear dynamical system (LDS) and a bag of spatial-temporal features (BOF)) to classify surgemes are proposed by [6]. In [40], they proposed a method of two training stages, learning a shared dictionary among all the possible

surgemes jointly in conjunction with the multi-class linear support vector machine (SVM) for segmentation and classification of the surgical gestures. Another end-to-end model fusion, the conditional random field (CRF), and discriminative sparse were presented by [41] for surgical activities segmentation and recognition. In [37] and [27], a framework based on DTW along with k-nearest neighbor is used for identifying surgemes. Recently, an attempt by [28] applied symbolic aggregate approximation (SAX) and vector space model (VSM) for discovering an interpretable pattern in surgical actions.

The mentioned approaches above used the kinematic data in their works except for [6], where the authors used both the kinematic information and video data. Several techniques proposed categorizing the surgical task into a set of subtasks or surgemes using video information and showing that applying these data can improve the surgical recognition analysis performance [42, 43]. For instance, [44] applied three techniques for surgemes classification: linear dynamic systems (LDS), a bag of features (BOF), and a fusion method for LDS and BOF using multiple kernel learning (MKL). Recent works apply the deep learning techniques after their widespread use in machine learning and computer vision applications for surgemes segmentation and classification. An LSTM approach was proposed by [45] to classify the kinematic information into labeled surgical gestures. In [46] developed a spatiotemporal CNN (ST-CNN) for fine-grained segmentation and recognition of surgemes. Another deep-learning approach based on the encoder-decoder temporal convolution network is suggested by [46, 47] to segment the surgical activities using video data. Recently, [48] deployed using a fusion between temporal convolution kernels and bidirectional LSTM for labeling the video sequences of the suturing task into surgical subtasks. Another branch of a deep-learning method for joint

surgemes segmentation and classification using deep reinforcement learning (RL) was proposed by [49].

1. 2. 3. B Surgemes Unsupervised Approaches

Most of the techniques are supervised classification and based on predefined or pre-segmented surgemes data in the previous categories. These surgical gestures are annotated manually by chief surgeons, consuming more time and being susceptible to human mistakes by missing parts (surgemes) or inconsistently criteria applied throughout a surgical task [12, 50]. Several works intended to identify surgical activities from unsupervised viewpoints without prior knowledge [51-53]. In [51] proposed a framework for segmentation and recognition surgemes from kinematic data. They first applied unsupervised segmentation by finding a relevant selection of dexemes (a numerical representation of subgestures to perform a surgemes). Secondly, learning features from dexemes to associate them to corresponding surgemes (composed of a set of dexemes) [51]. Another approach introduced by [54] is known as soft boundary unsupervised surgemes segmentation. The temporal sequence of surgemes segment and merge based on some criteria and then smooth the boundaries between parts. A recent deep-learning approach was proposed by [55], based on a deep convolution network using both kinematic and video data for surgical gestures segmentation.

1 . 3 Key Contributions

This work puts the spotlight on novel methods that depend on a surgery similarity measure. The key roles of our work present several contributions accomplished in the region of surgical skill assessment and surgical task analysis, which can be summarized as follows:

- Proposed a new distance measure, namely PDTW, for time series analysis mainly for surgical skill assessment classification (Chapter 2) and surgical task recognition (Chapter 3). The novel surgery skill distance PDTW comprises two main elements: Dynamic Time Warping (DTW) and Procrustes analysis (PA). The DTW method aligns two multi-dimensional sequences with various lengths by stretching both signals to become equal. The Procrustes measure, which includes reflection, scaling, and translation, can then be used as a distance measure between two aligned time series.
- We also developed a framework to automatically classify the surgical skill assessment into three expertise levels: expert, intermediate, and novice. The new distance is applied in the proposed model to align various length time series using DTW, then calculate their distance using Procrustes analysis. The k-nearest neighbor (k-NN) is used as a classifier in the framework to indicate the surgeon proficiency level. Our proposed approach is employed on multi-dimensional raw data and applied for various data kinds, e.g., kinematic and Vicon system data (Chapter 2 and 4) and wearable sensors data (Chapter 4). The performance of the proposed framework is validated using two sets of experiments on various types of information. Our results demonstrate the high performance of the PDTW over the conventional distances in classifying the surgeon skill levels on different tasks. Also, we conduct another scenario using only the right and left hands' cartesian coordinates instead of using the complete information. The experimentation illustrates that it reduces time consumption. Therefore, it can be applied for surgeon proficiency evaluation using low-cost worn devices instead of using pricey tools.

- We developed a surgical task recognition framework to classify three essential robotic minimally invasive surgery (RMIS) tasks: suturing, needle passing, and knot tying. We used the multi-channels kinematic data representing the surgeon's motion traces during a given surgical task in this work. The main components of the proposed model are the PDTW distance and the Fuzzy k-NN. A key benefit in addition to our distance measure PDTW is utilizing the Fuzzy k-nearest neighbor as a classifier by assigning a membership to get a confidence classification measure. Two different experiments were used to evaluate the proposed technique's performance over experiments comprising state-of-the-art methods on a widely utilized surgery dataset (Chapter 3).
- We built a classification approach to identify the surgical gestures based on raw kinematic data. In addition, we provide the mean feature reduction of surgical movements, which is efficient, reliable, and minimizes the complexity of the proposed approach while maintaining its functionality. Our proposed method utilizes k-NN and SVM algorithms to classify surgemes. Also, we employ two distance measures, the DTW distance with the k-NN technique and the Euclidean distance with the SVM algorithm. The best performance was achieved by combining the mean feature reduction method and the SVM algorithm. For both LOSO and LOUO cross-validation setups, our suggested methodology outperforms the existing method based on raw kinematic data.
- We created a technique for identifying surgical gestures by clustering surgical activities learned directly from raw kinematic data using unsupervised approaches based on predefined surgemes. First, we build the prototypes by clustering the

surgemes of the expert surgeon from all his\her trials. Then, we map the other surgeons surgemes to the closest representative. Finally, we report the clustering accuracy using the rand index technique. Our unsupervised surgemes clustering approach uses four approaches based on Hierarchical and FCM algorithms. We validated our methods using a real dataset by analyzing raw kinematics data from a suturing task with varying degrees of expertise levels. Additionally, we discuss the benefits of using mean feature technique of the time series data prior to clustering in terms of computational time savings and reducing the system complexity.

The research conducted in this dissertation has materialized in the following publications:

1. Albasri, Safaa, Mihail Popescu, and James Keller. "A Novel Distance for Automated Surgical Skill Evaluation." In 2019 E-Health and Bioengineering Conference (EHB), pp. 1-6. IEEE, 2019.
2. Albasri, Safaa, Mihail Popescu, and James Keller. "Surgery Task Classification Using Procrustes Analysis." In 2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1-6. IEEE, 2019.
3. S. Albasri, M. Popescu, S. Ahmad, J. Keller "Procrustes Dynamic Time Wrapping Analysis for Automated Surgical Skill Evaluation", *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 1, pp. 912-921 (2021).

CHAPTER 2 A Novel Distance for Automated Surgical Skill Evaluation

The assessment of surgical skills of surgeons with different levels of expertise has traditionally been performed by direct observation by a senior expert surgeon inside the surgical room. However, this method suffers from being subjective, costly, and lacking any quantitative indication of the appropriate skills level. Instead, an automatic evaluation would provide an objective and quantitative measure of skill levels [25]. To provide a rich source of quantitative motion information, robot-assisted minimally invasive surgery (RMIS), such as the Da Vinci surgical system (synchronized kinematic data and video), can be used [1]. These recorded data provide great opportunities to evaluate and train surgeons by creating descriptive mathematical models. Applying machine learning techniques to the Da Vinci data has motivated researchers to develop automated computational models and approaches to measure surgical proficiency, assess surgeon skills, and coach prospective trainees [14, 15]. Several methods have been proposed in the past to address this problem using either predefined surgical gestures (surgemes) or overall motion trajectories. For surgemes, statistical approaches such as Hidden Markov Models (HMM) provide an objective tool to evaluate laparoscopic surgical skills between expert and resident surgeons [56]. Despite the ability of this method to find the structure of RMIS tasks, it suffers from requiring an enormous number of training samples and parameter tuning, which leads to low performance [24] and need for complicated preprocessing [25]. An automated personalized gesture training model based on DTW was proposed [25] to

evaluate surgeon skills and provide online performance feedback during training. A texture-, frequency- and entropy-based feature set for automated surgical skill assessment framework was proposed in [13]. Also, in their study they proposed a weighted feature fusion technique for skill scoring using four holistic features on kinematic data: sequential motion texture (SMT), discrete Fourier transform (DFT), discrete cosine transform (DCT) and approximate entropy (ApEn) [13].

In contrast, most researchers have been focusing on motion features because of their simplicity in extraction and implementation. These features are named global movement features (GMF). It includes operation time, path length, depth perception, speed, and motion smoothness. They have been used to distinguish the relation between surgical tools movement patterns of an expert and novices during a surgery task [25]. Additionally, two features: turning angle, and tortuosity were added to the GMF for more accurate and robust in automatically evaluating surgeons [26].

Most of the prior works used temporal features or surgemes which lack efficiency and need domain-specific knowledge. More recently, deep conventional learning [15] and convolutional neural networks (CNN) [57] have been proposed for surgeon skills assessment using multivariate time series kinematic data in three independent tasks. Although these methods build the foundation for automated surgeon skills evaluation and obtained promising results, they have a few drawbacks such as complexity, complicated preprocessing, lack of robustness and are time consuming due to the need of parameter estimation. Instead, defining an adequate proficiency distance measurement may lead to a more robust and accurate evaluation framework.

In this chapter, we introduce a novel motion-based distance for surgical skill assessment using kinematic data for surgical training that consists of two main components: Dynamic Time Warping (DTW) and Procrustes analysis (PA). The DTW method aligns two time series with different lengths by contracting/dilating both signals such that their lengths become equal. The Procrustes analysis, that includes reflection, scaling, and translation, can then be used as a distance measure between two aligned sequences. Finally, a k-nearest neighbor (kNN) method is employed to classify a new sample surgery using the PDTW similarity to other data samples. In this paper, we use the surgery samples to classify surgeons in three classes: novice, intermediate and expert. While data obtained from RMIS systems is easiest to process, our method can handle any type of sensor data that can capture surgery motion. As an example, we show some results obtained from surgery motion captured with a 3D marker-based system (Vicon, www.vicon.com).

2 . 1 Methodology

The main components of our proposed framework are motion alignment, Procrustes distance, and classifier as shown in Figure 2.1. DTW is used to align two multidimensional time series performed by surgeons, while similarity measure is calculated by the Procrustes measure. Finally, the kNN is used to classify the skill levels of the surgeon.

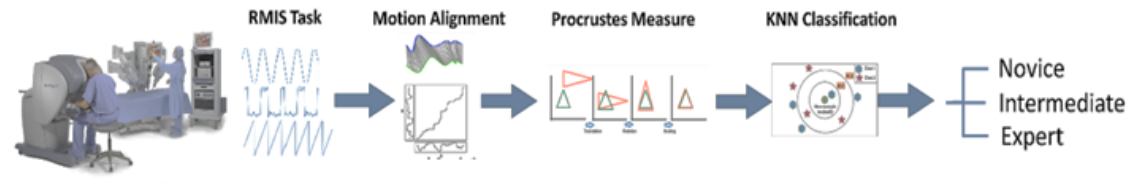


Figure 2.1 PDTW Skill Assessment for RMIS Framework.

2. 1. 1 Similarity Measure

To obtain a good classification, a crucial element is to define a good distance measure between two surgery tasks. Each task is represented by a set of features extracted from the traces (time series) of the motion capture sensors. Euclidean distance is one possible method. Assume two time series $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$of the same length, their Euclidian distance is [58]:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2-1)$$

Euclidian distance is simple and widely used; however, it has some limitations and drawbacks. The Euclidean method suffers from being very sensitive to outliers, translation, and requires both signals to be of the same length. For that reason, we need a measure that can process signals with different lengths as the same surgery task might have different lengths even when performed by the same surgeon. One solution is to use a warping distance measure such as the Dynamic Time Warping (DTW). DTW can process signals with different lengths, it stretches or contracts both signals (aligns them) such that their length becomes equal [58].

Let $\mathrm{X}_{n\times v} = [X_1, X_2, \dots, X_n]$ and $\mathrm{Y}_{m\times v} = [Y_1, Y_2, \dots, Y_m]$ 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 follows [59]:

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

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 [59]:

$$\Gamma(I,j) = d_{ij} + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(I,j-1)\} \quad (2-3)$$

Although DTW is widely used in many applications and is a more robust distance measure than Euclidean distance, it fails for complex multidimensional signals. It can also produce singularities by warping the X-axis where unevenness in the Y-axis is encountered. Features like valleys, peaks, and inflection points can cause DTW to fail to properly align two signals [59].

A common method for measuring the similarity of two shapes in directional statistics is the Procrustes shape analysis [60, 61]. The Procrustes analysis consists of best matching two shapes using similarity transformations (rotation, reflection, scaling, translation) to be as close as possible in the least squares sense. To examine the shape variability in a dataset, the Procrustes analysis can also estimate the mean shape [62].

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 [62]:

$$(X_i)_c = CX_I \quad , \quad i = 1,2 \qquad (2-4)$$

Where C, is the centering matrix and calculated in Eq. 2-5:

$$C = H^T H = I_k - \frac{1}{k} 1_k 1_k^T \qquad (2-5)$$

H is the Helmert submatrix, that the j^{th} row consists of h_j repeated j times followed by $-jh_j$ and then $k-j-1$ zeros:

$$\left(H_j, h_j, \dots, h_j, -jh_j, 0,0, \dots, 0\right) \quad h_j = -\{j(j+1)\}^{-\frac{1}{2}} \qquad (3-6)$$

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 [62]:

$$Z_i = \frac{(X_i)_H}{\|(X_i)_H\|} = \frac{H(X_i)}{\|H(X_i)\|} \quad , \quad i = 1,2 \qquad (2-7)$$

$$(X_i)_H = HX_I \quad , \quad i = 1,2 \qquad (2-8)$$

The full Procrustes distance between X_1 and X_2 is obtained by matching the pre-shape Z_1 and Z_2 as closely as possible as shown below [62]:

$$D_P(X_1, X_2) = \inf_{\Gamma,\beta} \|Z_2 - \beta Z_1 \Gamma\| \quad , \quad i = 1,2 \qquad (2-9)$$

where $\|.\|$ is the Euclidean norm, Γ is the rotation matrix (known as a special orthogonal matrix) with $m \times m$ dimensions and $\Gamma\Gamma^T = \Gamma^T\Gamma = I_m$ and $\det(\Gamma) = +1$. Scale parameter $\beta > 0$.

To overcome the limitations of using DTW alone, we use DTW as an alignment approach first and then apply Procrustes as a distance measure. This work introduces a distance measure PDTW based on a pairwise synchronization between two signals by using a combination of Procrustes and DTW. DTW is used to find the best matching between

two sequences, while Procrustes is used to minimize the distance from the estimated average to each realization.

2. 1. 2 Classification

The simplicity, and reasonable results, have made the k-Nearest Neighbors (kNN) algorithm a very successful feature classifier. The kNN uses the label of the training data to predict the class of new unlabeled test point x based on their similarity measure. The majority label of the k- closest neighborhoods assign the label of the query point [63]. In this paper, we found k = 3 is a reasonable value and the one we use.

2 . 2 Experimental Results

2. 2. 1 The JIGSAWS Dataset

The Da Vinci robotic surgical system provides an ideal environment for surgeons to perform fundamental training tasks. We use the JIGSAWS [11] dataset set acquired with the Da Vinci surgical system to implement our model for surgical skill assessment. Eight right-handed surgeons with different skill levels, expert (E), intermediate (I) and novice (N) performed three elementary surgical tasks, suturing (SU), knot tying (KT) and needle passing (NP) as shown in Figure 2.2. Each task was performed five times, each one named a trial. Due to some corruption of the data, the actual total number of trials is 39 for SU, 36 for KT, and 28 for NP.

The JIGSAWS dataset includes data from two experts, two intermediate, and four novices of robotic surgical experience. The kinematic data of the Da Vinci robot was captured at 30 Hz and it consists of 76 motion variables for all four manipulators (left and right master tool manipulators (MTM), left and right patient side manipulators (PSM)).

Each manipulator has 19 variables, 3 Cartesian positions, 9 rotation matrices, 3 linear velocities, 3 angular velocities, and a gripper angle. The three surgical skill training tasks that the surgeons performed are shown in Figure 2.2 and defined as [11]:

(a) Knot-tying (KT) (b) Needle-passing (NP) (c) Suturing (SU)

Figure 2.2 Three Da Vinci Surgical tasks [11].

Suturing (SU): the trainer proceeds the needle to the incision after picking it up first. The needle passes through the tissue from one side of the incision at the dot marked to the other side at the corresponding dot marked. Then, the surgeon extracts the needle out of the tissue and repeats the procedure three more times.

Knot-Tying (KT): the trainer picks up one end of a suture tied to a flexible tube attached at its ends to the surface of the bench-top model and ties a single loop knot.

Needle-Passing (NP): The trainer passes the needle after picking it up first, from right to left through four small metal hoops which are attached at a small height over the surface of the bench-top model.

Also, the dataset contains manual annotation for each trial based on a modified global rating score (GRS) by an annotating surgeon. GRS describes the sum of six elements: respect for tissue, suture/needle handling, time and motion, the flow of operation, overall performance, and quality of the final product. Each element score is between 1 and 5, so

the total is arranged between a minimum score of 6 through a maximum score of 30, where the highest score is the best [11].

2. 2. 2 MU-EECS Vicon Dataset

Six repetitions of the same surgical procedure (tracheostomy) were performed by a resident surgeon at the Center for Eldercare and Rehabilitation Technology laboratory in the department of Electrical Engineering and Computer Science at the University of Missouri, Columbia, MO [64]. The lab has a 7 camera Vicon system. Ten IR reflective markers to simultaneously track and visualize the motion involved in the surgical procedures were placed on both hands, wrists, elbows and shoulders of the resident surgeon as shown in Figure 2.3 (a). The first 3 repetitions were performed in conformity with the established procedure while the last 3 were conducted in an erroneous fashion.

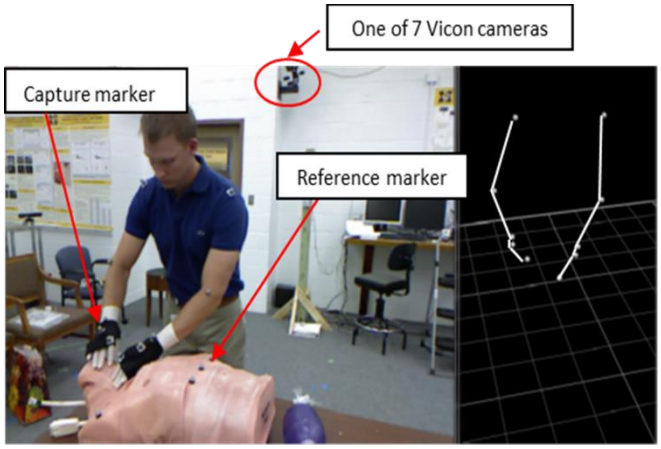


Figure 2.3 Surgical procedure captured with Vicon IR markers.

2. 2. 3 Performance Evaluation

To evaluate our skill assessment method, we used two cross validation settings for each task.

2. 2. 3. A Leave-One-Trial-Out (LOTO)

In each task, one trial is kept out for testing while the remaining trials of that surgical task are used for training. The procedure is repeated for each trial of each task.

2. 2. 3. B Leave-One-Supertrial-Out (LOSO):

In this cross validation, we created five folds similar to the number of repeated trials, each fold includes data from one of the five supertrials of the eight surgeons. The set of the ith trial from all surgeons for a given surgical task is called a supertrial i (i =1,2…5). Each time, one set was left out for testing and the union of the remaining supertrial sets was used for training [11]. Leaving out the supertrial ith might be able to capture the effect of tiredness or boredom on the surgeons, as repeating the task consecutively might produce some reduction in performance. This cross-validation procedure was performed for comparison purposes (see Table 2.I).

2. 2. 3. C Leave-One-User-Out (LOUO):

The LOUO system resulted in the creation of eight folds, and each fold contains all the trials of one surgeon in each task. The LOUO cross-validation is used to assess the robustness of the procedure when a surgeon has not been observed previously in training.

The performance of the proposed framework is compared using the overall accuracy of the classification results. The ratio of the total number of correct predictions over the total number of predictions defined as accuracy [65].

$$\text{Accuracy} = \frac{T_P + T_N}{P + N} \qquad (2-10)$$

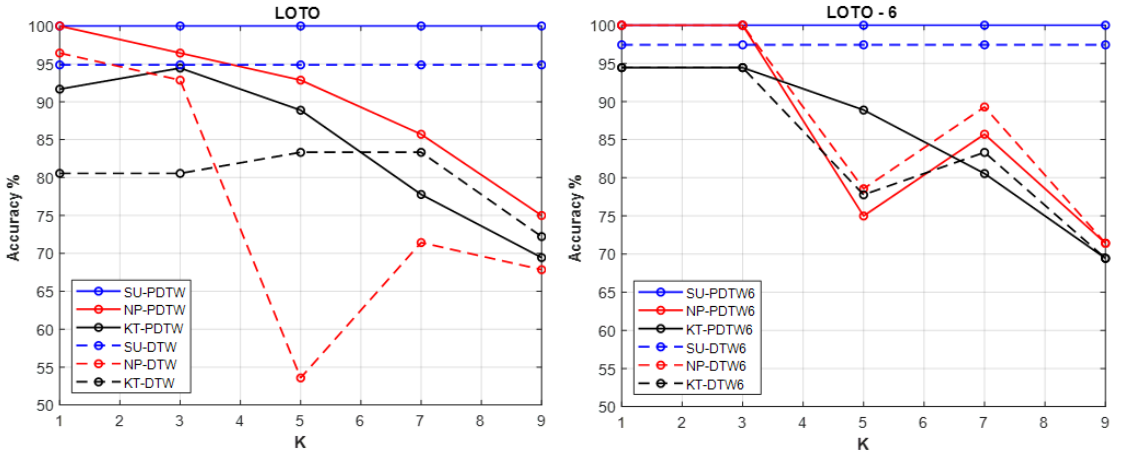
True positives (TP): the number of correctly classified as belonging to the target class.

True negatives (TN): the number of correctly classified that does not belong to the target class. where P: Positive predictions and N: negatives predictions.

2. 2. 4 Classification Results for the JIGSAW data

In this section, we report the experimental results of our proposed framework on the JIGSAW dataset. The three surgeon skill levels (expert, intermediate, and novice) are assessed on three different RMIS tasks: suturing, needle passing, and knot tying. To detect the expertise levels, we performed two sets of experiments. In the first set, LOSO and LOTO, we used all the kinematic data (76 motion variables). In the second set, LOSO-6, and LOTO-6, we used only two-hand Cartesian coordinates data (6 motion variables).

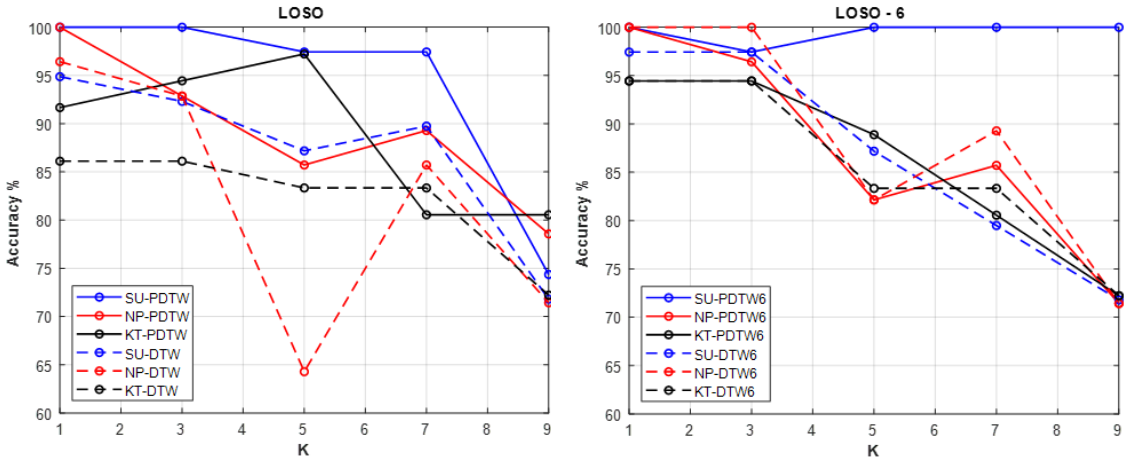
In Figure 2.4 and 2.5, we show the comparison of the kNN skills level classification accuracy as a function of k validating techniques, respectively. It can be seen that our PDTW based kNN classifier (continuous lines) produces better results than the DTW based one for most values of k. For kNN-PDTW using LOTO cross validation shown in Figure 2.4(a), skills detection in the SU task is better than in the other two tasks: it is 100% for all values of k. In the meantime, the best classification accuracy in KT is about 94.5% at k = 3 and it drops to about 70% when k = 9. At NP surgeon skills detection, the performance decreases gradually from 100% at k = 1 to 75% at k = 9. However, kNN -DTW has lower overall accuracy and is more sensitive to varying values of k as compared to PDTW.



(a) (b)

Figure 2.4 Accuracy of kNN-PDTW and DTW as function of k for LOTO validation using (a) all the 76 features (b) movements features (X,Y,Z)..

Figure 2.5(a) shows the results in all tasks using the LOSO validation technique for different values of k. From Figure 2.4(a) and 5(a), the average accuracy of recognizing surgeon expertise in all tasks at k = 3 is about 97% and 95.7% for LOTO and LOSO of all the kinematic information.



(a) (b)

Figure 2.5 Accuracy of kNN-PDTW and DTW as a function of k for LOSO validation using validation (a) all the 76 features (b) movements features (X,Y,Z).

The confusion matrices of our kNN-PDTW skill levels classification are summarized at k =3 in Figure 2.6 for the three tasks for LOSO validation, where the diagonal corresponds to the correct predictions. Surgeon expertise levels are 100% correctly classified in the suturing task, while in the knot-tying task and in the needle-passing task with about 94% and 93%, respectively. We can also observe from Figure 2.4 (a)/(b) and Figure 2.5 (a)/(b) that for two-hands Cartesian data, the accuracy using kNN-PDTW6 achieves almost the same performance as using all the kinematic information. In general, we observe that the accuracy of using two-handed Cartesian data alone is somewhat similar to using all the kinematic data. This is not surprising if we remember that Procrustes acts on shapes, which is exactly what the Cartesian data is. With other distance measures, we expect the result to be different, that is, using all 76 variables might lead to better results than using 6.

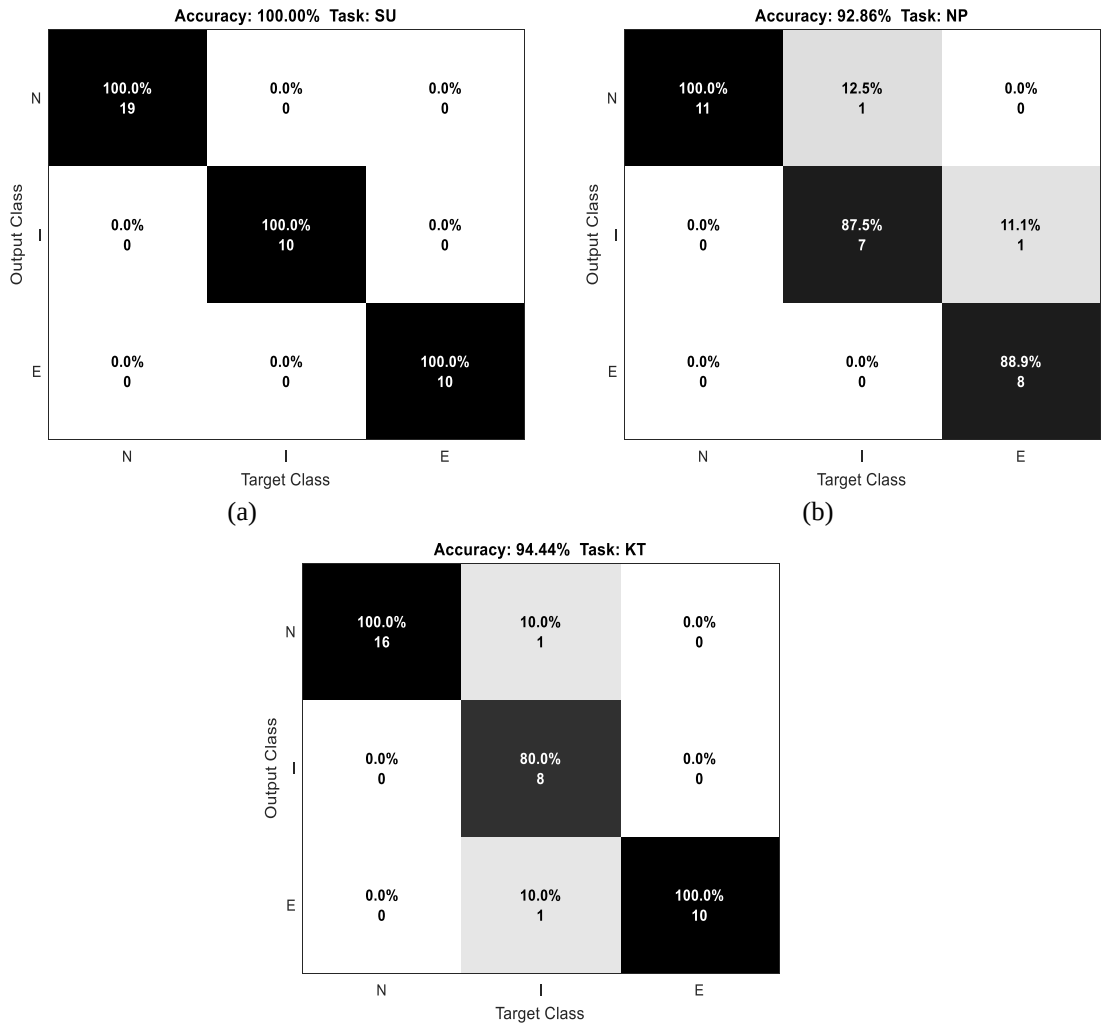
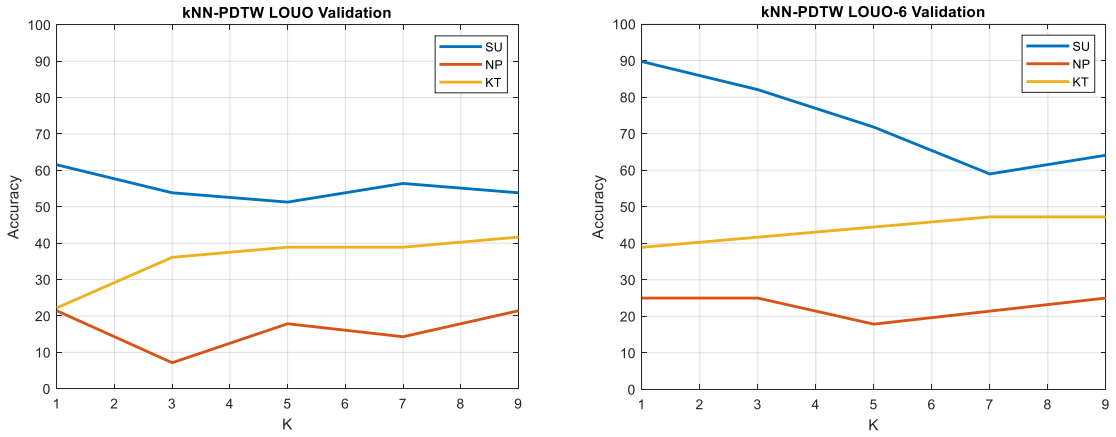


Figure 2.6 Confusion matrices results of the kNN-PDTW classification using LOSO Validation at k = 3 for (a) SU, (b) NP, and (c) KT.

Figure 2.7 shows the results for the LOUO cross-validation of the PDTW-kNN approach using all the features in (a) and utilizing only the motion traces of the left and right hands for each surgeon. The best results were obtained for the suturing task, which had an average accuracy of approximately 60%, whereas the NP and KT tasks had low performance. We can explain this because the LOUO validation is more challenging where the surgeon has not been seen before in training. Also, some trials of the intermediate surgeons were misclassified as novice or expert skill levels which makes it difficult for the

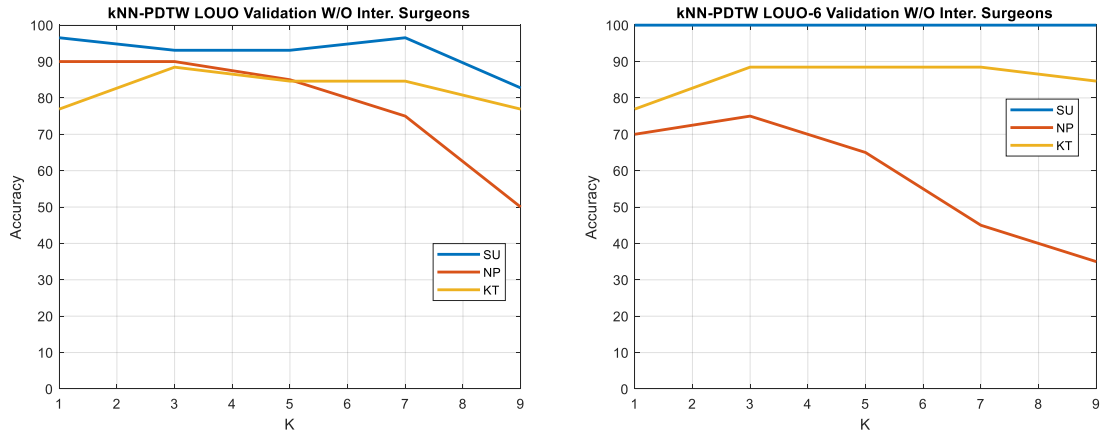
algorithm to uncover a pattern to distinguish between proficiency levels. Furthermore, as the number of samples in the dataset decreases, the results and the classification method do not draw a solid conclusion.



(a) (b)

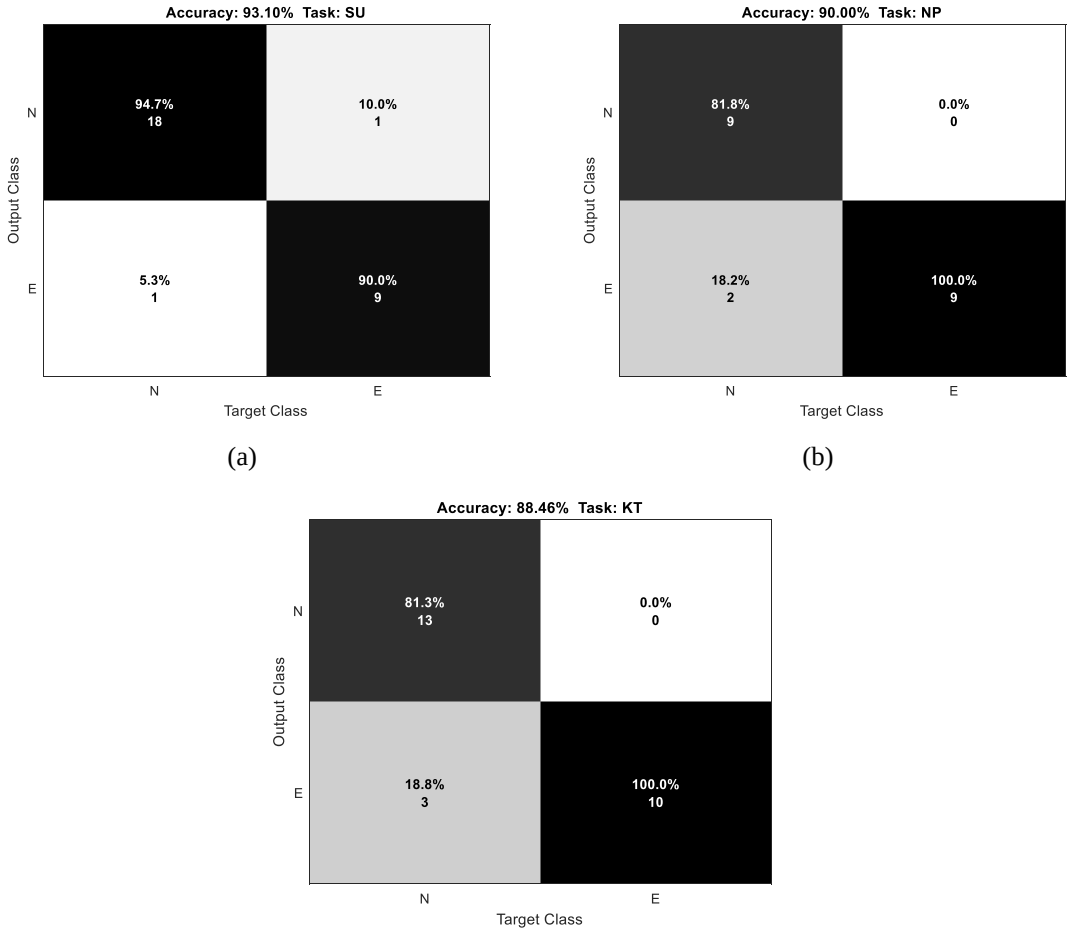
Figure 2.7: Accuracy of kNN-PDTW as a function of k for LOUO validation using (a) all the 76 features (b) movements features (X, Y, Z).

We run another experiment of the k-NN based on PDTW distance for the LOUO scheme to classify between two classes of surgeon skill levels, i.e., only novice and expert levels. We can observe from Figure 2.8 that the average accuracy increased for the three tasks using all the kinematic features. The best overall accuracy achieved was at k = 3 with an average accuracy of about 93%, 90%, and 88% for SU, NP, and KT, respectively. The confusion matrices at k = 3 for kNN-PDTW are shown in Figure 2.9 for the LOUO validation setup. It can be noted that the expert skill level was correctly classified in NP and KT tasks, while only one trial from both novice and expert was misclassified alternatively. Table 2.I demonstrates the comparison with the state-of-the-art method based on two classes (N and E skill level) classification results for the SU and KT tasks, where we outperform their method based on SVM for LOUO cross-validation.



(a) (b)

Figure 2.8: Accuracy of kNN-PDTW for two classes (E and N) as a function of k for LOUO validation using (a) all the 76 features (b) movements features (X, Y, Z).



(a) (b)

(c)

Figure 2.9: Confusion matrices results of the kNN-PDTW classification for two classes (E, N) using LOUO Validation at k = 3 for (a) SU, (b) NP, and (c) KT.

Table 2.I: Comparison of the classification accuracy (%) with the state-of-the-art method for LOUO setup (best accuracy highlighted in bold)

Algorithm	SU Task (%)			KT task (%)		
	Novice	Expert	Overall	Novice	Expert	Overall
SVM based on S+C (spatial motion and curvature features) [26]	74.4	81.2	79.8	75.3	80.5	77.9
Our method kNN-PDTW	94.7	90	93.1	81.3	100	88.5

The boxplot of the pairwise PDTW distance between expert surgeons (E/E), expert-intermediate surgeons (E/I), and expert-novice surgeons (E/N) computed for all three tasks in JIGSAWS dataset is displayed in Figure 2.10. It shows that the distance between expert surgeons (E/E) is the smallest for each task, followed by the distance between experts and intermediate (E/I) and experts to novice surgeons (E/N). However, discriminating between intermediate and expert surgeons is more challenging in NP compared to other tasks. This might be related to the difficulty level of the task: needle passing might be more complicated than knot tying or suturing, hence difficult to learn. It can also be seen in Figure 2.6, where one expert was misclassified as intermediate for the NP task.

Table 2.II shows a comparison of our skill evaluation results to existing similar methods for the LOSO validation scheme. We see that for suturing, our method matched the result from [57] , 100%, obtained using a convolutional neural net (CNN). However, for knot tying, we surpassed the same CNN [57], and we were close to the results from [22] based also on CNNs. In needle passing both above CNNs did better than our method. We

note that no method shown in table I is best for all tasks, which means that a skill evaluation framework might need to contain multiple methods joined by some fusion methodology.

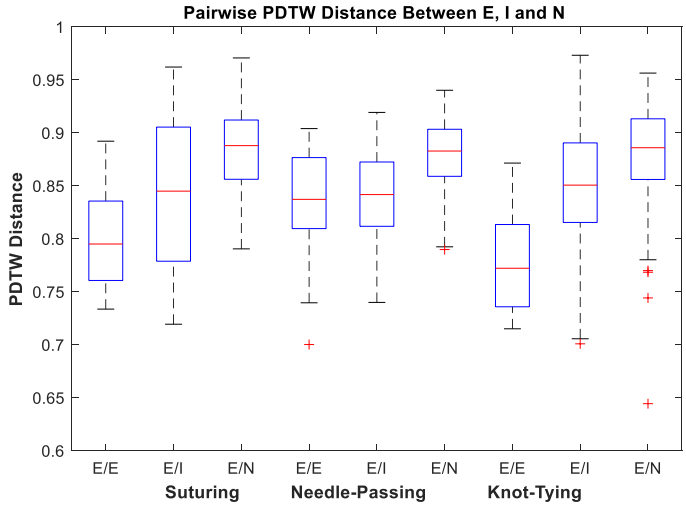


Figure 2.10 Boxplot of PDTW distance between group of E-E, E-I, and E-N surgeons for all tasks.

Table 2.II JIGSAWS Dataset Classification Comparison of expertise levels using LOSO with state-of-the-art methods.

Method		**Accuracy (%)**		
		SU	***NP***	***KT***
GMF [26]	*k*NN	89.7	-	82.1
	LR	89.9	-	82.3
	SVM	75.4	-	75.4
Deep learning CNN [15]		93.4	89.8	84.9
SAX-VSM [28]		89.7	96.3	61.1
CNN [57]		**100**	**100**	92.1
CNN-LSTM+SENET+Restart [22]		98.4	98.4	**94.8**
***k*NN-PDTW (proposed)**		**100**	**92.8**	**94.4**

Figure 2.11 illustrates the surgeon's global rating score (GRS) for each trial and for all the tasks. The GRS score describes the measure of the entire trial for surgical technical skills. This figure shows a more consistent pattern for the expert surgeons in comparison to the intermediate and the novice surgeons, where the higher score implies better skills in performing the task. Also, we can see from this figure that the experts have the lowest variance in all the tasks, which indicates their consistency.

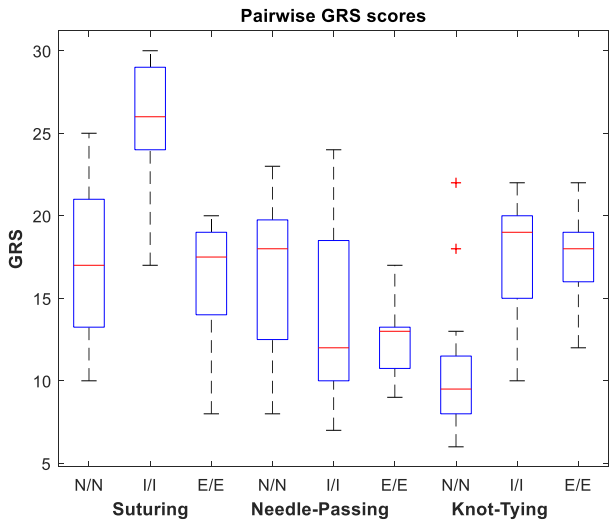


Figure 2.11 Pairwise Boxplot of Skill-level GRS for: a) Suturing, b) Needle-passing, and c) Knot-tying.

Furthermore, this figure gives an intuition of the challenge of discriminating between surgeon skill levels in needle-passing, which causes the misclassifications. Another thing to be noticed is that some intermediate surgeons (residents) have higher scores than the expert surgeons (attending), which indicates that they might be qualified to be an attending physician.

2. 2. 5 Results on the MU-EECS data

In addition to the JIGSAWS dataset, we used a PDTW analysis approach to compute the distance between six trials of a tracheostomy procedure performed by a resident surgeon [64]. The resulting distance matrix is shown in Figure 2.12, where "0" represents the closest and "1" as the farthest procedure (trial).

We can see that the distance between the first 3 "good" trials are less than or equal to 0.5 (e.g., the similarity between procedures 2 and 3 is about 0.3). At the same time, none of the "bad" trials is less than or equal to 0.7 similar to each other. The "bad" trials are about 0.7 from any of the "good" ones. An exception is the similarity between trial 1 and 5 is about 0.55.

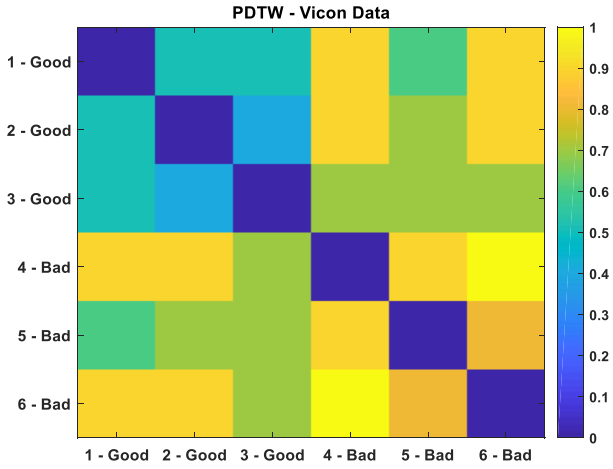


Figure 2.12 Pairwise distance matrix using PDTW for the Vicon dataset.

However, overall, we can see that the "good" trials seem to cluster together in the left-upper corner. Furthermore, the boxplot of the pairwise PDTW distances between Good trials and Bad trials computed respectively is shown in Figure 2.13. It's clearly seen from this figure that the mean and variance of the Good trials is smaller ($\mu_{G-G} = 0.12,\ \sigma_{G-G} = 0.08$), as compared to the Bad ones ($\mu_{B-B} = 0.21,\ \sigma_{B-B} = 0.16$).

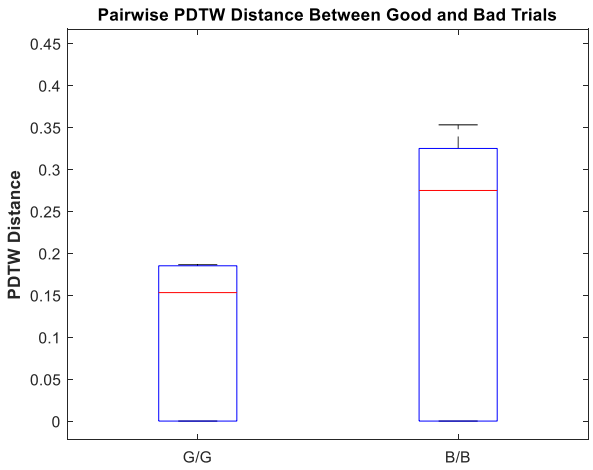


Figure 2.13 Pairwise Boxplot of PDTW distance between Good and Bad trials separately for MU-EECS Vicon dataset.

2 . 3 Conclusions

This work introduces a distance measure based on a pairwise synchronization between two signals by using a combination of Procrustes and DTW. DTW finds the best matching between two sequences, while Procrustes is used to minimize the distance from the estimated average to each realization. The performance of our model achieves an average accuracy of 97% on the JIGSAWS dataset, which is comparable to state-of-the-art results based on CNNs without the heavy training required by deep models. In addition, the study highlights the possibility of using only the Cartesian coordinates of the left and right hands for computing the PDTW distance measure for surgical skill assessment.

CHAPTER 3 **Surgery Task Classification Using Procrustes Analysis**

Surgery is one of the crucial factors in global health care because a massive number of surgical operations are performed throughout the world annually in all fields [66]. This has encouraged the rising interest in computer-assisted surgery (CAS) tools. Furthermore, systems with sensing devices that capture the surgeon's motions are getting equipped in the operating rooms [67]. Thus, improving the safety and the efficiency of surgical patient care is the main goal for studying surgical activities because the common reasons for post-surgical complications, i.e., re-operation and re-admission, are the surgical technical errors [11, 67].

Robot-assisted minimally invasive surgery (RMIS), such as da Vinci surgical system (dVSS), is an excellent source to obtain the motion information related to training surgeons (for both video and kinematic data) [1]. RMIS has the potential to improve trainee skills and patient healing results by shortening the stay at the hospital and providing faster time recovery as well [68]. These recorded data motivate the researcher to develop computational models for better surgery procedures understanding and to analyze surgical tasks automatically [14].

To recognize and identify the surgical tasks (knot-tying, needle-passing, suturing), multiple approaches have previously been proposed for RMIS data. These methods are mainly based on two types of data: kinematic or video. The work on da Vinci robotic public dataset JIGSAWS [11] can be categorized into three types: surgeon skill assessment,

surgical gestures analysis, and surgical task recognition. In this chapter, we focus on the third part by analyzing the huge kinematic information of the surgical activities captured from dVSS using its application programming interface (API) [11]. A sequence of events to achieve a surgical objective is defined here as a task which it is often called in the literature [9].

Several existing works address this problem; two approaches utilized the dynamic time warping [37] and derivative dynamic time warping [38] along with k-nearest neighbor to recognize the three surgical tasks. A multi-output deep learning architecture proposed [29] to characterize the surgical operation activity by combining CNN and gated recurrent unit network. All these approaches are based on multiple channels of kinematic sequences. For task classification based on video data, a multi-model design has been proposed based on CNN for visual features and long short-term memory (LSTM) network as motion cues for classifying gestures and surgical tasks in robot-assisted surgery (RAS) [39]. Even these approaches could find underlying surgical tasks, however, they are limited and have some common flaws such as time consuming, heavy weight computation load and architecture, high pre-processing data, and significant human interaction.

In this chapter, we develop and investigate a new framework to classify different surgical tasks. We apply our framework to an existing dataset, JIGSAWS, that contains three tasks (classes): knot-tying, needle-passing, and suturing. Our framework is based on using dynamic time warping (DTW), Procrustes analysis, and Fuzzy k-nearest neighbor. First, DTW finds the best matching between two multi-channels surgery sensor data by compressing and stretching the time series. The distance between the aligned time series is measured by the Procrustes distance. Finally, the fuzzy k-nearest neighbor is used to

classify a new kinematic surgical task data performed by subject as one the known tasks from our database.

3 . 1 Methodology

The goal of this work is to classify the RMIS tasks (knot tying, needle passing, and suturing in our case) using raw motion kinematics da Vinci data. The pipeline of our framework (see Figure 3.1) is composed of three main parts: sequences alignment, distance measure, and task classification. In this section, we present details on the methods and the proposed tasks classification framework. For alignment, DTW is used to align two multidimensional sensor signals performed by surgeons, while Procrustes is used to calculate the similarity measure. Finally, the fuzzy kNN is used to classify the surgical tasks.

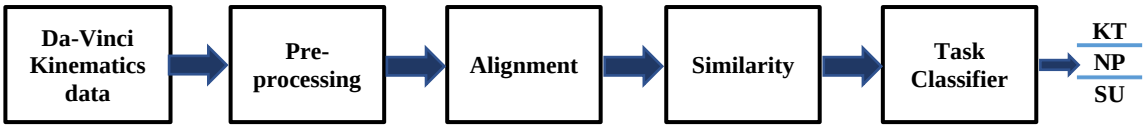


Figure 3.1 Fuzzy k-Nearest Neighbor PDTW RMIS Task classification Pipeline.

3. 1. 1 Pre-Processing Data

Each individual channel (of the total of 76) of the raw da Vinci motion sensor data was normalized to a zero mean and unit variance. Then, the noisy data smoothed with a gaussian-weighted moving average filter [69] using a window size of 15 samples. The sliding window was carefully chosen to reduce the noise and to keep the important details of the pattern as in the following equation [32]:

$$\tilde{X}(j) = \frac{1}{W}\sum_{r=0}^{W-1} X(j+r) = \frac{1}{W}\sum_{r=0}^{W-1} \frac{x(j+r)-\mu}{\sigma} \qquad (3-1)$$

Where W is the window size at each point j after being normalized with the mean μ and the variance σ of the time series X.

3. 1. 2 Task Similarity Measure

A good similarity measure between two sequences is one of the most important steps to get better classification results. Euclidean distance is one of the common and simple methods, but it requires both sequences to have the same length. Other drawbacks of this measurement are its sensitivity to outliers, noise, and shifting. For instance, if two sequences are similar in shapes but have a different offset, this leads to false or mismatched similarity measurements [70].

Dynamic time warping (DTW) is one solution for finding the optimal matching between two timeseries by stretching and compressing to align them [71]. It was first used in speech recognition applications and introduced by [72, 73]. DTW can re-align two sequences by matching the coordinates inside both sequences. Figure 3.2 shows the alignment for two signals, where one signal from a novice surgeon and the other from an expert surgeon; it seems that they have the same shape but are shifted slightly. Each point from sequence A is globally compared by DTW to any arbitrary point of the second sequence B, it results in a sequence pattern that represents the best coupling between the two sequences A and B where Euclidean distance is unable to match [71]. The warping cost, which minimizes the total distance between two timeseries, provides monotonicity constraints, boundary conditions, and continuity. The cost can be recursively calculated by [59]:

$$D(a_i, b_j) = \delta(a_i, b_j) + min\begin{Bmatrix} D(a_{i-1}, b_{j-1}) \\ D(a_i, b_{j-1}) \\ D(a_{i-1}, b_j) \end{Bmatrix} \qquad (3-2)$$

where $A_{M\times v} = [a_1, a_2, \dots, a_M]$ and $B_{N\times v} = (b_1, b_2, \dots, b_N)$ are two time series with N and M instances, respectively, $\delta(a_i, b_j)$ is the Euclidean distance that stretches the i*th* sample of A ($a_i \in R^M$) and the j*th* sample of B ($b_j \in R^N$) onto a common set such that the overall distance of the time warping path is smallest which is [73]:

$$D_{DTW}(A, B) = D(a_M, b_N) \qquad (3-3)$$

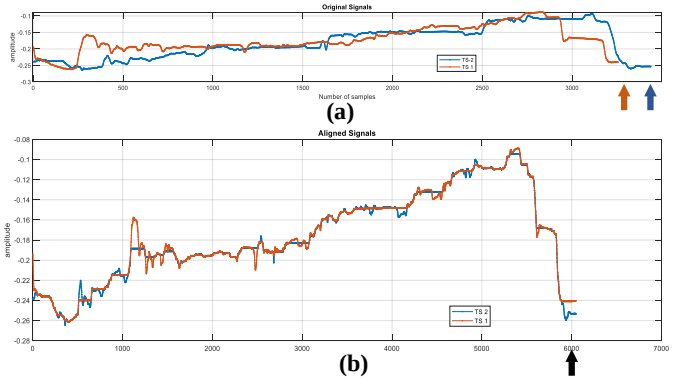


Figure 3.2 Dynamic time warping alignment.

In many domains, DTW has widely been used for its robustness as a distance measure. However, one limitation is that it can produce unreasonable results when the algorithm may try to explain Y-axis variability by bending the X-axis. Thus, a single point on one time series can change onto a wide part of another time series. Also, the algorithm may fail to align two signals when they differ because of features such as peaks, valleys, inflection points, ... etc. For example, DTW may fail to detect natural matches between two sequences

merely because a feature in one sequence is somewhat higher or lower than the equivalent feature in the other sequence [59].

In directional statistics, Procrustes distance analysis is used to compare two shapes [60, 61]. In geometry, pre-shape refers to the geometric information that remains after location and scale effects have been removed. Full Procrustes distance between A and B is defined as the Euclidean distance between the full Procrustes that fit of pre-shapes Z_1 and Z_2 as [74]:

$$D_p(X,Y) = \inf_{s,a,b,\theta,} \left\| Z_1 - Z_2\, s\, e^{j\theta} - (Z_2 + jb)1_k \right\| \qquad (3-4)$$

Where s is the scale, Θ is the rotation, and $(a + jb)$ is the translation, and the centered pre-shaped are given below. 1_k is a k-dimensional vector of ones, and I_k is a kxk identity matrix [75]. Figure 3.3 shows three steps of the procrustes measure to minimize the distance between two different shapes by translating, rotating, and scaling to the center shape of the desired one

$$Z_i = \frac{C(X_i)}{\|C(X_i)\|} \quad and \quad C = I_k - \frac{1}{k} 1_k 1_k^T \ , \qquad i = 1,2 \quad (3-5)$$

The fusion of Procrustes and DTW measures overcome the limits of using DTW alone as mentioned before. Procrustes-DTW (or PDTW) utilizes DTW to align and find the best fitting between two sequences, then uses the procrustes to calculate the similarity measure.

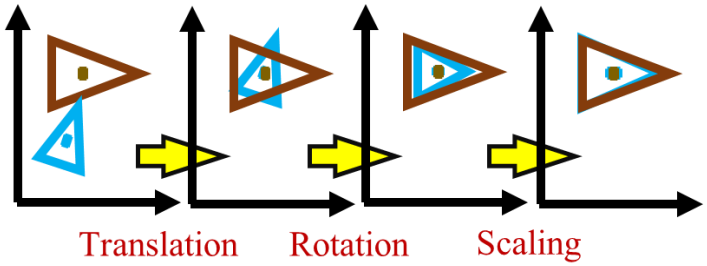


Figure 3.3 Procrustes measure steps.

3. 1. 3 Fuzzy k-Nearest Neighbor

The crisp k-NN algorithm is widely used because of its simplicity and because it needs only one parameter k (the number of the closest training neighbors to a test point). The k-NN assigns an unknown classification pattern to the class of the majority of its known k-NNs according to a distance or similarity measure [63, 76]. Figure 3.4 illustrates the difference between using fuzzy kNN and crisp kNN for k = 7. For example, the class label goes to class B for crisp kNN by the majority vote, while it assigns to class A using the FkNN case because of the higher fuzzy memberships.

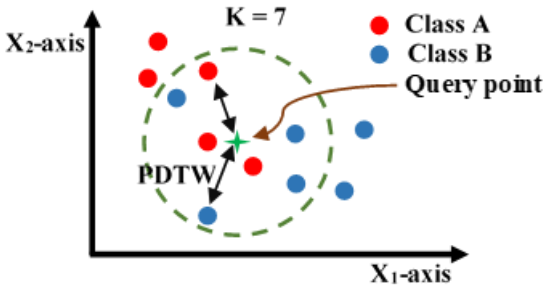


Figure 3.4 kNN versus Fuzzy kNN.

One drawback of the crisp k-NN classification is in the assignment of the query point where the k-NN training samples are handled as being equally important. Another problem of k-NN is that there is no indication of how much a pattern belongs to a given class. Fuzzy

k-NN (FKNN) algorithm overcomes these problems, which allocates a fuzzy membership to each training input neighbor instead of using a hard class membership [76]. let $x_1, x_2 \dots x_k$ be the k nearest neighbor of the query point x_q. The assigned membership value u_i to x_q in a class i is computed by [63]:

$$u_i = \frac{\sum_{j=1}^{K} u_{ij} \left(1/dist\left(x_q, x_j\right)^{\frac{1}{m-1}}\right)}{\sum_{j=1}^{K} \left(1/dist\left(x_q, x_j\right)^{\frac{1}{m-1}}\right)} \qquad (3-6)$$

$$u_{ij} = \begin{cases} 0.51 + 0.49\left(\frac{n_i}{n}\right) & i = j \\ 0.49\left(\frac{n_i}{n}\right) & i \neq j \end{cases} \qquad (3-7)$$

Where $m > 1$ is the fuzzifier, n_i is the number of n closest neighbor of x_j labeled *i*, and u_{ij} be the pattern's fuzzy membership in class i of the jth actual training labeled k-NN [63, 76]. In this work, we picked m = 2 and PDTW as the distance measure in (7) for reasonable results, where $dist$ stands for the PDTW distance. The pseudo code of our proposed FKNN using PDTW is shown in Figure 3.5.

Algorithm 1: FKNN-PDTW

Input: k (number of closest neighbors).
1. X (query multi-dimensions time series).
Y (set of N multi-dimensions training data)
2. **Output**: class label for X.

Compute PDTW

3. **for** i = 1,2, … N **Do**
4. Align X and Y_i by DTW using (2) and (3).
5. Find the distance by Procrustes measurement using (4).
6. **End**

Calculate k-NN

7. Sort the training set Y ascendingly based on their distance to X.
8. Y_j ($1 \leq j \leq k$) is the set of k nearest neighbors.

Find Fuzzy Classifier

9. **for** i = 1,2, … C (number of classes) **Do**
Assign membership u_i to X in all classes using:
10 $$u_i = \frac{\sum_{j=1}^{k} u_{ij} / \{PDTW(X, Y_j)\}^{\frac{1}{m-1}}}{\sum_{j=1}^{k} 1/PDTW(X, Y_j)^{\frac{1}{m-1}}}$$
11. **End**
12. **END**

Figure 3.5 Fuzzy KNN algorithm using PDTW.

3 . 2 Experimental results

3. 2. 1 JIGSAWS Dataset

We test our proposed model FKNN-PDTW on a public robotic surgical dataset presented in [11]. The JIGSAWS dataset includes kinematic and video data captured by application programming interface (API) using da Vinci surgical system (dVSS) at 30 Hz sampling frequency. Eight right-handed surgeons with different expertise performed three fundamental surgical tasks knot tying (KT), needle passing (NP), and suturing (SU) as shown in Figure 3.6 and defined below:

1) Knot-Tying (KT): The trainer ties a single loop knot after picking up one end of a suture tied to a flexible tube attached at its ends to the surface of the bench-top model.

2) Needle-Passing (NP): The trainer picks the needle up and passes it through four small metal hoops, from right to lift, which are attached at a small height over the surface of the bench-top model.

3) Suturing (SU): The trainer first picks up the needle and proceeds it to the incision. Second, the trainer passes the needle at the dot marked from one side to the corresponding dot marked side through the tissue. Finally, the subject extracts the needle out of the tissue then repeats the procedure three more times.

The surgeons performed each task above for about five repetitions, each one known as a trial. KT consists of 36 trials, NP has 28 trials, and 39 trials for SU because some data were corrupted. In this work, we analyze our model using kinematic data that consists of 76 sensor motion dimensional. Multi-channels kinematic data includes 19 variables for the four manipulators of dVSS ends (two master sides and two patient sides for each hand). The 19 variables for each end have 3 Cartesian positions, 9 rotation matrices, 3 linear velocities, 3 angular velocities, and a gripper angle [11].

(a) Knot-tying (KT) (b) Needle-passing (NP) (c) Suturing (SU)

Figure 3.6 Three Surgical tasks used in our experiments [3].

3. 2. 2 Performance Evaluation

Two validation schemes were conducted to evaluate our proposed framework for the three surgical tasks classification on each testing set and to make a comparison with other states of art methods as suggested in [11]. We also report the performance metrics regarding the sensitivity or recall, specificity, precision, and f1-score for each output class results.

Leave one supertrial out (LOSO) and leave one user out (LOUO) settings were used for this dataset. In LOSO validation, five folds were created; each has the ith supertrial (i = 1,2…5) from all the surgeons for the three tasks. One folder was kept for testing while the union of the remaining supertrial sets used for training each time. This LOSO evaluated the robustness of the algorithm when the ith trial from all the surgeons left out, while the robustness of the method when a surgeon had not been seen previously in training is evaluated in LOUO cross-validation. Eight folds were created in the LOUO scheme, each one consisting of all the trials that belong to three tasks of the surgeon as shown in Figure 3.7 [11].

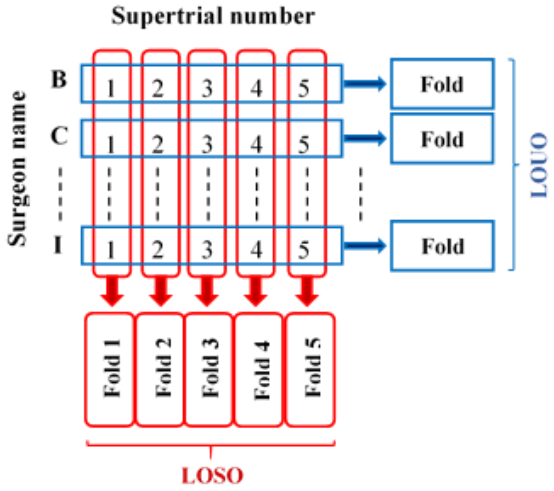


Figure 3.7 LOSO and LOUO validation techniques.

To evaluate our classification results, we compared each class output to the ground truth of the provided data by [11]. The evaluation metrices can be defined as follows [77]:

1) The accuracy is defined as the number of tasks that correctly classified as positive divided by the total number of tasks:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3-8)$$

Where TP is the number of tasks correctly classified as belonging to the target class. TN is the number of tasks that are correctly classified not belonging to the instances task. If a class is reported incorrectly to the task, it is FP, and it is FN, if it is not classified as an instance task [77].

2) Precision: the ratio of the number of tasks that are correctly classified as positive over the number of tasks that are labeled by the classifier positive.

$$Presision = \frac{TP}{TP + FP} \qquad (3-9)$$

3) Sensitivity (Recall): The number of tasks that correctly classified as positive divided by the number of tasks that labeled positive in the ground truth.

$$Sensitivity = \frac{TP}{TP + FN} \qquad (3-10)$$

4) f1-score: The harmonic-mean of the precision and the recall:

$$f1\_score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \qquad (3-11)$$

3. 2. 3 Classification Results

The experimental results of identifying the three surgical tasks: knot-tying, needle passing, and suturing using our proposed classification framework on the JIGSAWS dataset are reported in this section. First, we used our PDTW distance to measure the similarity between traces that represent the surgeon motion, then we applied the Fuzzy kNN classifier to recognize between different RMIS surgical tasks. For results evaluation on the testing set, we performed two cross-validation setups: LOSO, and LOUO techniques which are widely used and suggested by the JIGSAWS dataset. We implemented our algorithm using only kinematics data of the JIGSAWS dataset which was collected from the surgeons as described before.

Figure 3.8 shows the variation of the classification accuracy as a function of k (the number of neighbors), which in turn, it tests the robustness of our model based on Fuzzy kNN for each surgical task using LOSO and LOUO validating schemes. As we can see from this figure, we need at least three neighbors for LOSO and five for LOUO validation techniques to ensure model robustness. The accuracy of our proposed framework is better performed in the LOSO setting than the LOUO technique which the surgeon has not seen before.

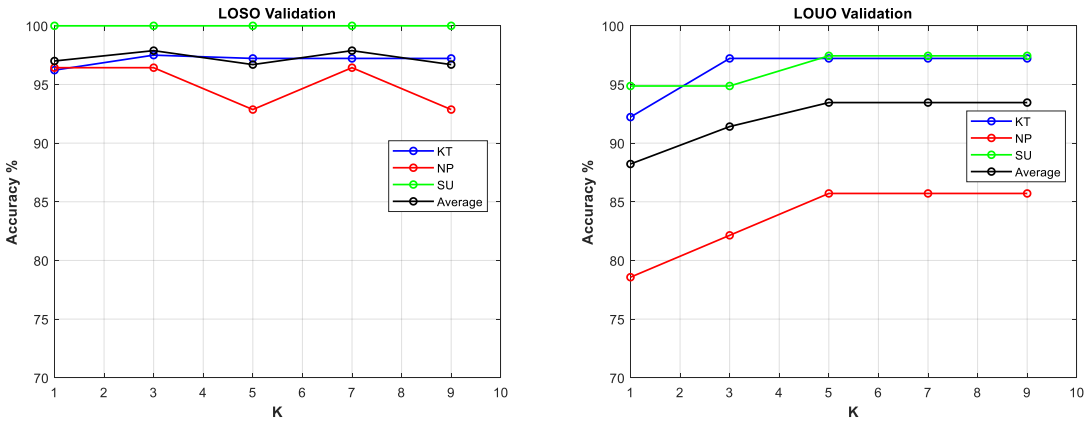


Figure 3.8 Accuracy comparison as function of k for LOSO and LOUO validation technique.

For LOUO cross-validation, the higher classification accuracy was achieved starting at k = 5. We picked this value to reduce the computational complexity and time consumption. Our Fuzzy kNN model provides more consistency and robustness for all three robotic surgical tasks. KT and SU tasks achieved better results as compared to NP tasks. Which indicates the challenge of recognizing NP surgical tasks compared to the other two tasks performed by study subjects.

Figure 3.9 illustrates the visual intuition of the confusion matrices of Fuzzy kNN-PDTW prediction results for three classes (KT, NP, and SU). The performance matrices of these surgical tasks are evaluated on the testing set for both 5-fold LOSO at k = 3 and 8-fold LOUO at k = 5 cross-validation schemes. Suturing task is 100% correctly classified in LOSO, and one class misclassified for both knot-tying and needle-passing. Also, we can observe that KT keeps the same accuracy in both validation settings, while the accuracy degrades to 97.4% and 85.7% for SU and NP when switching from LOSO to LOUO validation.

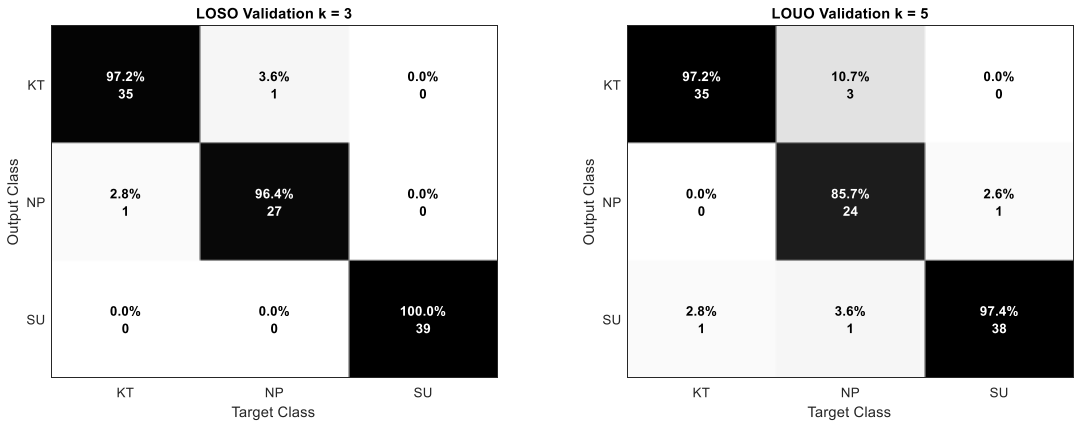


Figure 3.9 Confusion metrices for LOSO and LOUO validation techniques using Fuzzy kNN-PDTW.

To further assess, we compared our model results of the overall surgical task performance to the state-of-art classification methods, as shown in Table 3.I. All benchmarks are evaluated using the same LOSO and LOUO cross-validation sets applied to raw kinematics data suggested by JIGSAWS dataset. For the LOSO validation scheme, our proposed method overall results are close to the results of [29] which are based on deep task learning networks. Also, our approach provides accuracy improvement of at least 1.5% compared to other existing methods. We outperformed the other techniques with the accuracy improvements ranging from 4.5% to 10.1% for LOUO validation.

Table 3.I Performance Comparison Using Overall Accuracy (%) for LOSO and LOUO Schemes.

Method	**LOSO**	**LOUO**
Farad, 2017 [21]	92.4	84.1
Farad, 2017 [21]	95.5	89.7
Wang, 2018 [22]	96.6	-
Wang, 2018 [23]	**100**	-
FKNN-PDTW	**98.1**	**94.2**

Table 3. II illustrates the performance measurements regarding the sensitivity, specificity, precision, and f1-score. These measurements give an indication that our algorithm can yield better results for each class for LOSO and LOUO schemes. Where the sensitivity metric concerns of identifying the positive tasks by a classifier. The precision focuses on the agreement between positive task labels by a classifier with positive labels given by the data. The specificity measures the effectiveness of the classifier to recognize the negative instances. Finally, f1-score emphases on the relationship between the positive task labels classified by a classifier, and positive tasks given by the data.

Table 3.II Performance Results in Terms of Sensitivity, Specificity, Precision, Recall, and f1-Score for Each Class for LOSO and LOUO Schemes.

	at **k = 3**				
LOSO	**Task**	**Sensitivity**	**Specificity**	**Precision**	***f*1-Score**
	KT	97.2	98.5	97.2	97.2
	NP	96.4	98.7	96.4	96.4
	SU	100	100	100	100
	at **k = 5**				
LOUO	**Task**	**Sensitivity**	**Specificity**	**Precision**	***f*1-Score**
	KT	97.2	95.5	92.1	94.6
	NP	85.7	98.7	96.0	90.6
	SU	97.4	96.9	95.0	96.2

Furthermore, we repeated our experiment using two-hands cartesian data (6 variables) instead of using all the kinematics information (76 variables) with LOSO and LOUO cross-validating techniques. Different values of nearest neighbors (k) were tested to identify the best k to be used in training our multi-class classifier. We found that k = 5 is a reasonable value to obtain the highest overall accuracy for surgical task classification. The performance of Fuzzy kNN for both hands data almost achieved similar or better results

than using all the kinematic data, around 99%, for both validations schemes. This indicates the potential of using two-hands information which leads to the possibility of using low-cost wrist wearable techniques instead of using expensive tools.

3 . 3 Conclusions

This study introduces a surgical multi-tasks classification framework in robotic assisted minimally invasive surgery based on a new similarity measure that uses the combination of dynamic time warping and Procrustes analysis. DTW aligns two timeseries, and Procrustes is used to obtain the similarity distance measure between two motion paths of the da Vinci kinematic data. Using a Fuzzy k-nearest neighbor as a classifier is a key advantage of our proposed framework in improving the performance results by assigning a fuzzy membership based on our PDTW distance to obtain a confidence classification measure. In addition, the performance of our approach surpasses published results for LOUO validation by at least 4.5% accuracy improvement. Also, it is comparable with state-of-the-art methods that are based on deep architectures for LOSO validation technique. Furthermore, utilizing 3D cartesian motion trajectories of the right and left hands reduces the time consumption and emphasizes the potential toward using low-cost wrist wearable devices instead of using expensive tools for surgical tasks recognition and surgeon skill assessment.

CHAPTER 4 **Procrustes Dynamic Time Wrapping Analysis for Automated Surgical Skill Evaluation**

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 [34]. 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 [25, 35].

As in 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 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 [9]. A reasonable solution to these challenges is to use technological advances like Robotic Minimally Invasive Surgery (RMIS) that improve overall operating room efficiency [9]. For instance, da Vinci surgical technology provides data-driven that potentially helps optimize and develop training skills for surgeons [10]. This information includes kinematic and video data that conduct a useful resource of quantifiable human motion during surgical operating [1, 11]. Wearable sensing devices

that provide detailed motion information for surgical activities are a further example [13]. 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 [12, 14, 15].

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 [12, 28]: 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 chapter, we focused on the surgical skill evaluation based on kinematic and wearable sensors information. One of the initial works used Hidden Markov models (HMM) [24] 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 [25] and leads to low performance with a low number of samples [24]. Another method was proposed by [25] 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 tasks. They extended their work to include eight global movement features (GMF) in [26], They applied LR, SVM, and kNN classifier to distinguish between the previous expertise levels for suturing and knot tying surgical tasks. In [27], 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, [28] 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 [13] 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 [29] 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 [15] for surgical expertise evaluation. Another parallel deep learning approach was proposed in [22] by combining the LSTM recurrent network and CNN to indicate the skill levels. Additionally, recent studies have suggested approaches that use motion from

videos [30, 31] and wearable sensors to evaluate surgical skills [32, 33]. 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 [34]. Also, an entropy-based features technique that utilizes both video and accelerometer data proposed for surgical skill assessment [35]. 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 chapter, 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.

4 . 1 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 4.1. 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.

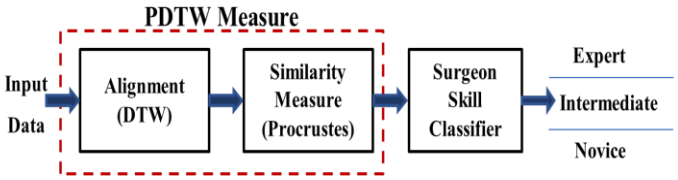


Figure 4.1: kNN based PDTW evaluation Framework.

4. 1. 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.

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 [58].

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 follows [59]:

$$\mathrm{d_{ij}}(\mathrm{x_i}, \mathrm{y_j}) = \sqrt{\sum_{\mathrm{k=1}}^{\mathrm{v}} (\mathrm{x_{ik}} - \mathrm{y_{jk}})^2} \quad (4-1)$$

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 [59]:

$$\gamma(\mathrm{i},\mathrm{j}) = \mathrm{d_{ij}} + \min\{\gamma(\mathrm{i}-1,\mathrm{j}-1), \gamma(\mathrm{i}-1,\mathrm{j}), \gamma(\mathrm{i},\mathrm{j}-1)\} \quad (4-2)$$

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 [59].

The Procrustes analysis is a standard method in statistical analysis to compare the similarity of shape objects [60, 61]. 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 [78]. The Procrustes analysis also can estimate the mean shape to examine the shape variability in a dataset [62].

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 [62]:

$$(\mathrm{X_i})_\mathrm{c} = \mathrm{CX_i} \quad , \quad \mathrm{i} = 1{,}2 \quad (4-3)$$

$C = \mathrm{H^T H}$ 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 [62]:

$$\mathrm{Z_i} = \frac{(\mathrm{X_i})_\mathrm{H}}{\|(\mathrm{X_i})_\mathrm{H}\|} = \frac{\mathrm{H(X_i)}}{\|\mathrm{H(X_i)}\|} \quad , \quad \mathrm{i} = 1{,}2 \quad (4-4)$$

$$(\mathrm{X_i})_\mathrm{H} = \mathrm{HX_i} \quad , \quad \mathrm{i} = 1{,}2 \quad (4-5)$$

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 [60]:

$$\mathrm{D_P(X_1, X_2)} = \inf_{s,a,b,\theta,} \left\| Z_1 - Z_2 \; s \; e^{j\theta} - (Z_2 + jb)1_k \right\| \quad (4-6)$$

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.

4. 1. 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 [63]. We found k = 3 is a reasonable value and the one we utilize in this chapter.

4 . 2 Experimental Evaluation

We used three datasets to evaluate the proposed PDTW-kNN model on the public surgical data JIGSAWS [11], and our two data MU-EECS [64], 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:

4. 2. 1 JIGSAWS Data

We evaluate the proposed PDTW-kNN method for surgical proficiency assessment on a public widely used JIGSAWS dataset [11]. 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 [79].

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 trial 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 4.2 presented sample frames of the three surgical tasks achieved by a surgical surgeon and defined them as follows [11]:

- 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.

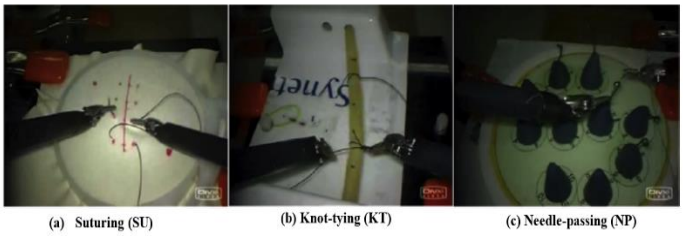


Figure 4.2: RMIS basic surgery tasks [7].

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 4.I. Where each component rating scale is between 1 and 5 and the best with a higher total score [11].

Table 4.4.I. Elements of Global Rating Score (GRS) [11].

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 =	$\sum = 30$

4. 2. 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 4.3 (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 [64].

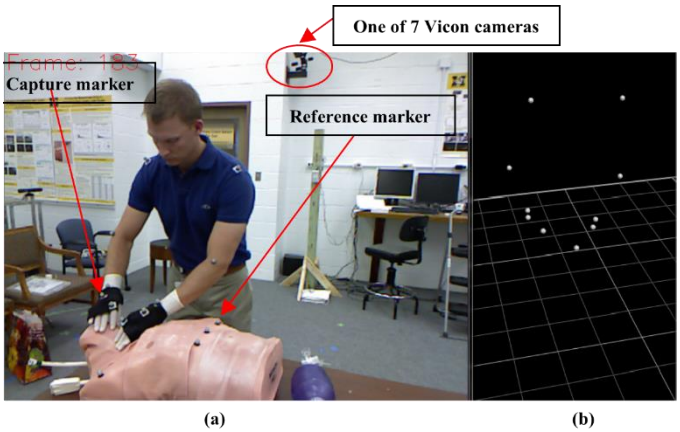


Figure 4.3: Tracheostomy surgery with Vicon Camera [64].

4. 2. 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 [80]. There are three main steps of the surgical Cric procedure skin incision, incision cricothyroid, and endotracheal tube placement membrane [81].

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 4.4. 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 [82].

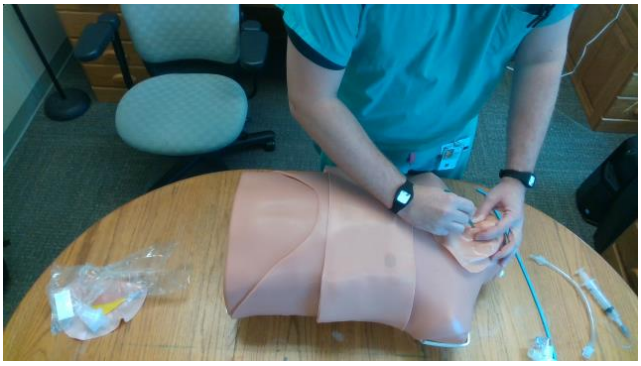


Figure 4.4: Cric surgical operation on TraumaMan Simulator by a medical surgeon.

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 accelerations 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.

4 . 3 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 [11]. 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) [65]:

$$\mathrm{ACC} = \frac{\mathrm{T_P} + \mathrm{T_N}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (4-7)$$

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 [65].

4 . 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. 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 4.5 (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.

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 4.5(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.

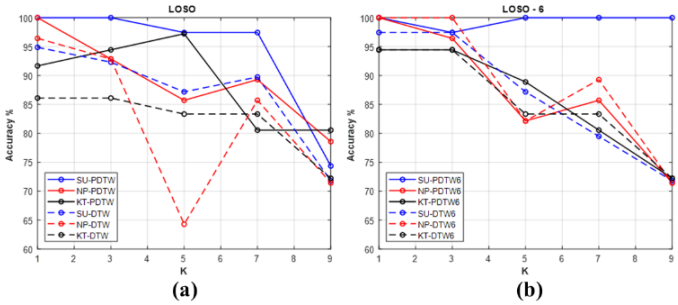


Figure 4.5: Accuracy of the proposed approach using PDTW and DTW as a function of k.

For further comprehensive comparison, the confusion matrices result for each task is shown in Figure 4.6 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 introduce an error to recognize their skill levels and vice versa. The same case between intermediate and novice surgeons happened.

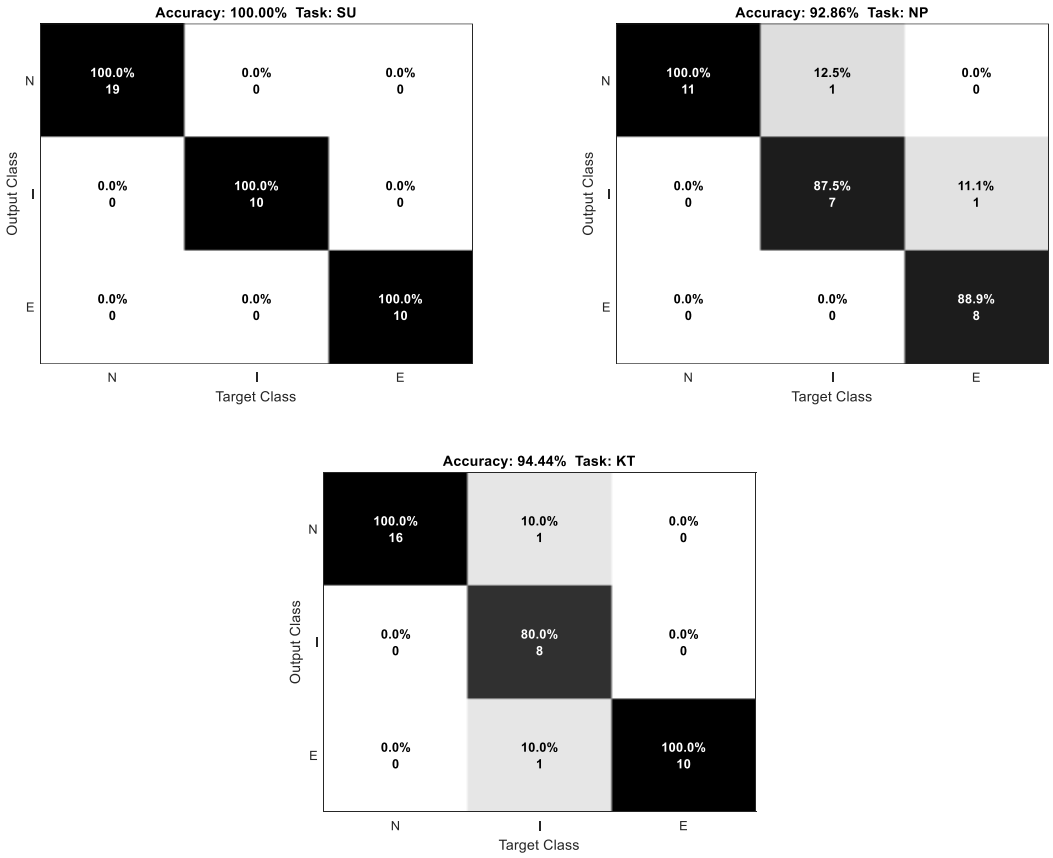


Figure 4.6: kNN-PDTW Confusion matrix of the three tasks SU, NP, KT for LOSO at k=3.

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 4.7 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 4.6 for the needle-passing task where an expert surgeon classified as an intermediate surgeon mistakenly.

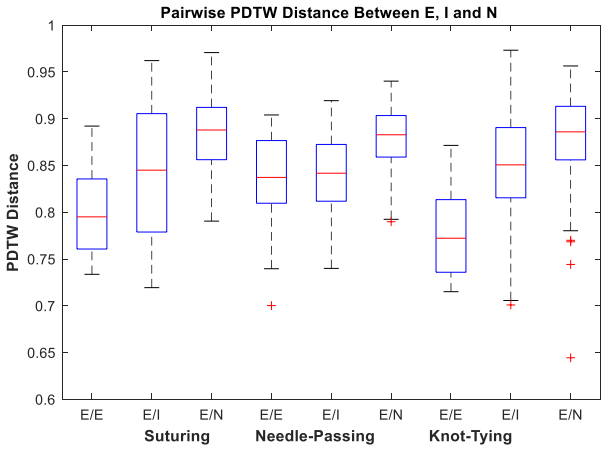


Figure 4.7: PDTW-distance within E/E, E/I, and E/N surgeons in each task.

Table 4.I 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 [57] for suturing. From Figure 4.7 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 [57] and Deep Learning [15] approaches in terms of accuracy. Also, our results were near the CNN+LSTM+SENET method [22]. Our results were improved more for suturing and knot-tying tasks than the needle-passing task, and we did slightly better than [15] in this task. The small distinctions between intermediate surgeons with other surgeons in this task illustrated in Figure 4.7 might explain the less performance on the needle-passing task. Furthermore, we can notice from Table 4.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 4.4.II: Skill Assessment Classification Comparative of kNN-PDTW Performance using LOSO for JIGSAWS Data.

Approach		***Accuracy***		
		SU	***NP***	***KT***
Farad [26]	kNN	89.7%	-	82.1%
	LR	89.9%	-	82.3%
	SVM	75.4%	-	75.4%
Wang [15]		93.4%	89.8%	84.9%
Forestier [28]		89.7%	96.3%	61.1%
Fawaz [57]		**100**%	**100%**	92.1%
Anh [22]		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 4.8 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 4.8, 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 4.8, 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.

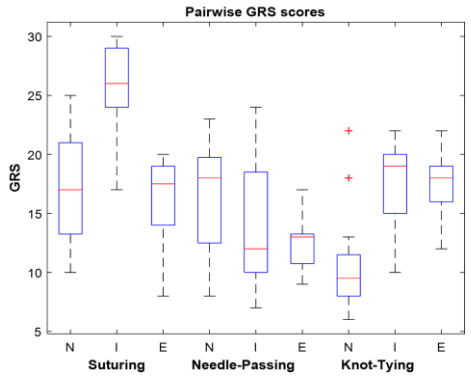


Figure 4.8: Boxplot of GRS scores for each task.

4. 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 operated by a resident surgeon [64]. Figure 4.9 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.

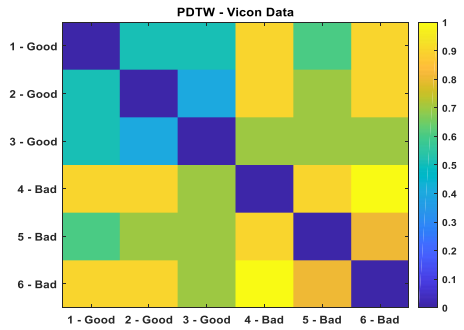


Figure 4.9: PDTW distance matrix for MU-EECS data.

Overall, the *Good* trials, which are the first three trials in Figure 4.9, 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 4.9, 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 4.10. 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.

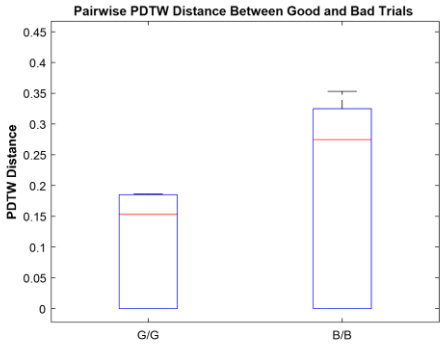


Figure 4.10: Boxplot of PDTW distance for MU-EECS dataset.

4. 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 4.11 (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 Figure11 (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 4.11 (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.

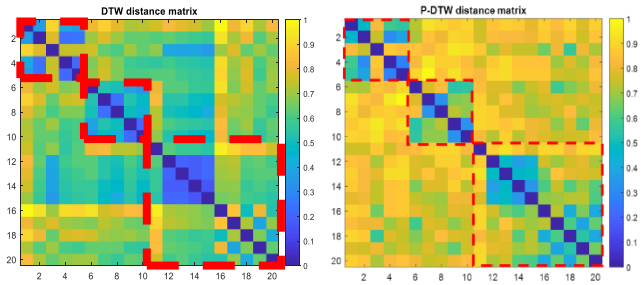
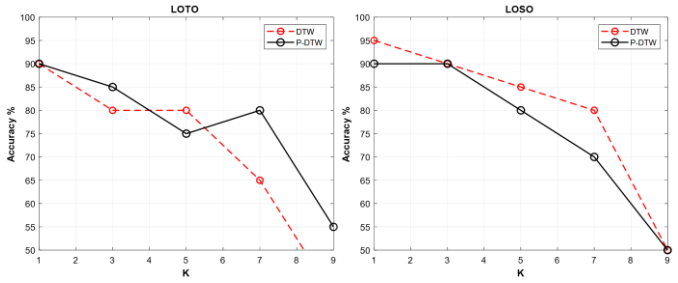


Figure 4.11: The pairwise distance for each trial on EM-Cric using (a) DTW and (b) P-DTW.

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 4.12 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 4.12 (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 4.12 (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.



(a) (b)

Figure 4.12: Classification accuracy as a function for k (a) LOTO and (b) LOSO cross-validation for Cric data.

Figure 4.13 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 4.11 (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%.

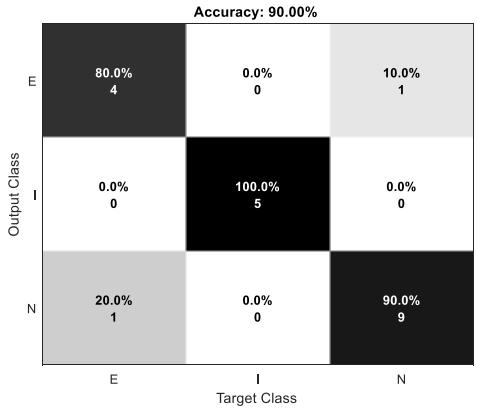


Figure 4.13 kNN-PDTW Confusion matrix for LOSO at k=3 for Cric data.

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 4.14 shows the comparison classification accuracy as a function of k between PDTW and DTW based kNN classifier. Furthermore, Figure 4.15 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%.

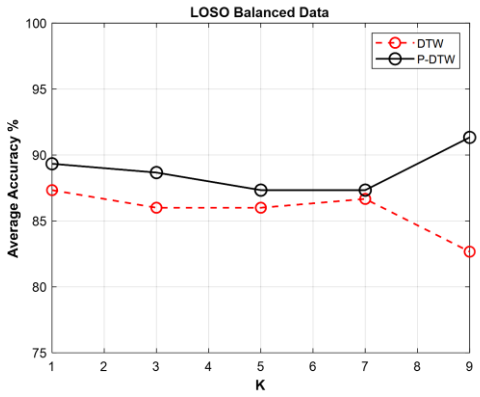


Figure 4.14: Balanced data classification results for the Cric data.

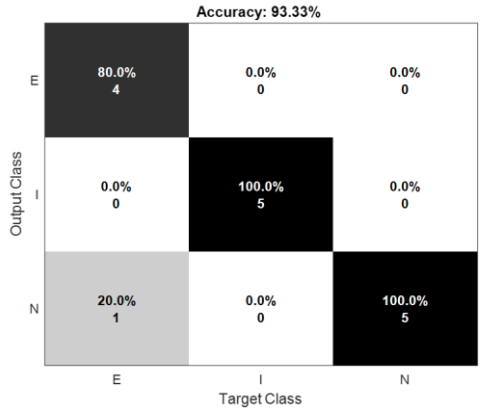


Figure 4.15: Balanced data confusion matrix for the Cric data.

4 . 5 Conclusions

In this chapter, 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.

CHAPTER 5 **Surgemes Classifications using Mean Feature Reduction**

5 . 1 Introduction

A little more than a century ago, Dr. William Halsted established the first surgical resident training program in the United States. His training methodology was quite straightforward: "see one, do one, teach one." [44] Expert feedback is the most frequent method for evaluating surgeon technical performance, either during surgery or afterward via video review. However, an attending physician may not always be available in person, and post-operative video assessment can be time consuming and subjective. With experienced surgeons limited free time, this approach is clearly not scalable. Thus, it is important to create objective, automated, and time-efficient approaches for evaluating surgeon technical skill [83]. Recent technological advancements have altered the way some procedures are conducted. This has created an opportunity to review Halsted's paradigm to identify more effective methods of teaching surgeons [44]. These difficulties have prompted and motivated us to develop techniques for automatically assessing RMIS skills and classifying gestures.

In this chapter, we developed a framework for surgical gesture classification based on raw kinematic data. We employ the k-nearest neighbor and support vector machine algorithm as a classifier in our proposed method. Also, we present the mean feature reduction technique to represent the surgical segments in more feasible way and allow us to utilize various distance measures.

5 . 2 Methodology

In this section, we will demonstrate our developed framework for recognizing surgemes that operate directly on raw kinematic data using a mean-feature approach. To classify the predefined surgemes in each surgical task, we used two distinct distance measures: dynamic time warping and Euclidean distance. We also describe the classification methods that we utilized and will define the performance metrics that will be used to compare our results to those of other proposed methods. Figure 5.1 provides a summary of the proposed surgical gestures classification framework for robotic assisted minimally invasive surgery data. In summary, our approach begins with predefined surgical gestures. Then, determine the similarity between different surgemes. In this step, either using DTW to measure the pairwise distance among the multivariate time series or Euclidean distance (ED) with the mean feature reduction representation, we will explain it later in this chapter. In the first case, we used a k-NN classifier; in the second, we used a support vector machine (SVM).

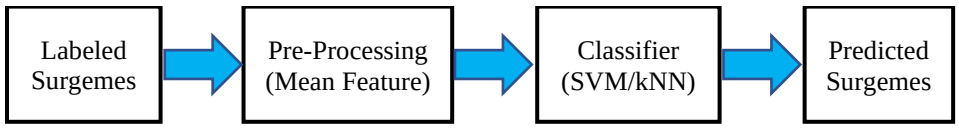


Figure 5.1: Flow diagram for surgemes classification.

5. 2. 1 Support Vector Machine

The support vector machine (SVM) is an effective and simple algorithm which is widely adopted for classification, pattern recognition and regression. The SVM was first developed by Vladimir Vapnik and Alexey Chervonenkis [84] in 1963. In the 1990s, SVM

was improved and suggested by Vapnik [85] and later on used by Byun and Agarwal [86, 87] based on a statistical learning methodology. The concept of SVM method unlike the traditional approaches, such as neural network (NN), is to classify the dataset into two groups or more by using a linearly or nonlinearly hyperplane (a separable line or curve) and to increase the margin between separating data as illustrated in Figure 5.2.

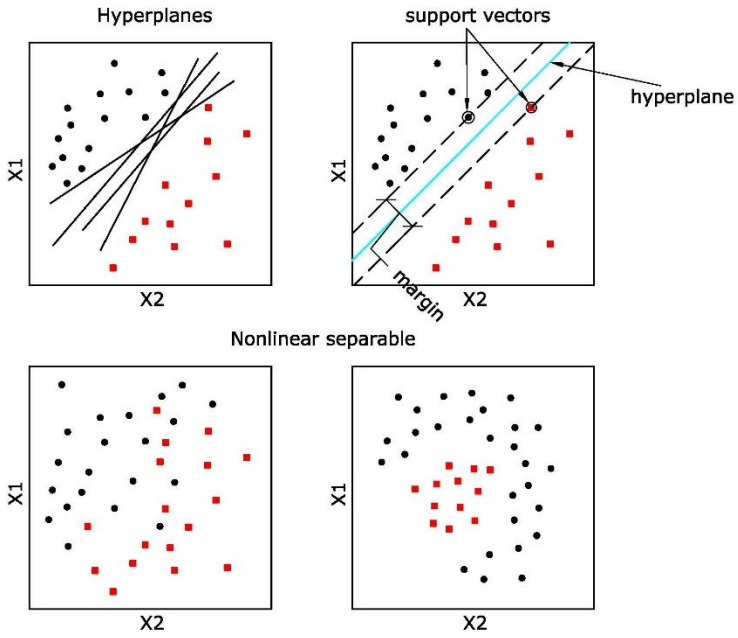


Figure 5.2: Define the hyperplanes in a dataset. H_1 and H_{-1} are the positive and negative support vectors, respectively.

The best hyperplane that has the biggest margin between two categories in one dataset, which represents distances between the hyperplane and the closest points in the classified categories in the dataset. In 1990's, Vapnik developed SVM as a Soft Margin Classifier or support vector machine in case there are some misclassifications of datasets as stated in Figure 5.2. Furthermore, the SVM had been improved by utilizing Kernel techniques by maximizing the features space boundaries to employ non-linearity between classes [88]. Kernel functions are algorithms that quantify resemblances between observations. There

are several types of Kernels that are adopted to classify non-linear datasets, such as polynomial, radial basis and linear Kernels.

For a 2D linear case, the following steps are below summarized to depict the procedure that is employed to solve a problem in the support vector machine (SVM).

- A linear classifier has a form of:

$$f(x) = W^T x_i + b \qquad (5-1)$$

where W is the weight vector, x is the input vector, b is the bias and $i = 1, 2, \dots, N$.

H_1: $W^T x + b = +1$, H_{-1}: $W^T x + b = -1$, H_0: $W^T x + b = 0$ …… (5-2)

- d+ and d- are the smallest distances from the hyperplane to the nearest positive and negative points, respectively. Thus, the margin is d=d++d- for a given weight vector W and bias b.
- Maximizing the distance d that leads to increasing the margin to obtain an optimal hyperplane.
- The distance from a point (x0, y0) to a line:

$ax + by + c = 0$ is $\frac{ax_0 + by_0 + c}{\sqrt{a^2 + b^2}}$ …….. (5-3)

Consequently, the distance between H_1 and H_0 is:

$d^+ = \frac{W^T x + b}{\|W\|} = \frac{1}{\|W\|}$ and then the margin = $\frac{2}{\|W\|}$ …….. (5-4)

1. From the previous equation, $\|W\|$ needs to be minimized to maximize the margin under the status that there are no points between H^1 and H^{-1} lines.

max $\frac{2}{\|W\|}$ is subjected to:

$$W^T x + b \begin{matrix} \geq 1 \; if \; y_i = +1 \\ \leq -1 \; if \; y_i = -1 \end{matrix} \; for \; i = 1, 2, \dots, N. \; \dots\dots\dots\dots\dots\dots\dots\dots \; (5-5)$$

and equivalently min $\|W\|^2$ is subjected to:

$$y_i(W^T x_i + b) \geq 1 \; for \; i = 1, 2, \ldots, N. \ldots\ldots\ldots\ldots\ldots\ldots \text{(5-6)}$$

2. The obtained optimization problem is a quadratic function that can be solved by using the LaGrange multiplier method.

5. 2. 2 K nearest Neighbor

k-Nearest Neighbor (KNN) is considered one of the supreme machine learning classifiers. Its applications are multi-sided and could be used in finance, healthcare, political science, image processing and many other purposes. Anyhow, it has two main features: non-parametric and lazy learning algorithms. The formation of the model is decided from the dataset itself. It means no assumptions are required for data distribution. It is more beneficial since most datasets do not obey mathematical models. Moreover, it does not need to train data to generate models, for lazy algorithms, since all data are trained during the testing phase. In other words, this leads to the training phase being faster and testing phase being slower. In general, KNN needs more time to train and test the whole dataset which means more memory.

Basic steps could be summarized into three points as followed:

1- Calculate distance between a new point (an example) and the other points of the other classes (such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance).
2- Find and decide which are the closest neighbors.
3- Voting for labels depending on the neighbors that are close to a new point.

In general, KNN works perfectly with a lower number of features. This means the high great number of features requires more datasets. In addition, another disadvantage with

KNN is the number of dimensions, which is called the curse of dimensionality. This issue (the growth of dimensions) might cause an overfitting.

Several solutions might be dealt with to avoid such a problem like this. Performing the principal component analysis before subjecting the machine learning technique or using another method which calls a feature selection approach.

There is no optimal value for k-nearest neighbor approach since each dataset has a unique features or own requirements. Researches have demonstrated that a small value of neighbor gives low bias but high variance while a high value of neighbor displays a lower variance but higher bias [89].

The k-Nearest Neighbor is one of the simple algorithms that has been used in machine learning. the concept of KNN is by classifying new points based on other points of the dataset that are more like them. KNN is an algorithm which is treated as both non-parametric and lazy learning. the feature of non-parametric means that there are no assumptions that would be made. In other words, the full model is composed depending on the given datasets in lieu of assuming its structure. For lazy learning, moreover, there is no popularization. In other words, training datasets are little in the training process while all the training datasets are employed second time in testing phase by using KNN method. Figure 5.3 below shows the concept or the idea how the KNN works when attempting to assort a new point in a dataset based on another given datasets.

It would be figured that the process will be started to its nearest points and assort depending on which is closet and more like. There are several ways and methods that calculate the distances between the new points and the points of given datasets such as Euclidean, which is a mathematical method. Anyhow, after KNN calculated the distances

between each new point and tested datasets. The method detects the probability of similarity between the new points and tested datasets and then assorts them according to the highest score of probability.

Despite that the KNN is simple to use and has no assumptions on the datasets, it still has some cons. Its accuracy mainly depends on the quality of the data. Moreover, an optimal k number (the value of nearest neighbor) must be determined to obtain greater accuracy.

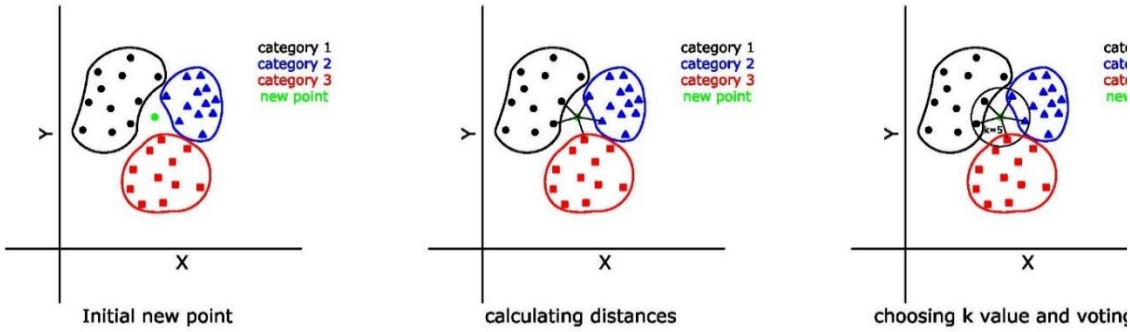


Figure 5.3: The concept of assorting new points depending on given datasets.

5. 2. 3 Similarity Measure

In time series classification, where data are labelled according to their similarity, the distance measure is critical. The Euclidean distance and dynamic time warping (DTW) are the most frequently used similarity measurements in this domain due to their effectiveness and efficiency in determining the similarity between objects. Euclidian distance is a straightforward, fast, and parameter-free calculation. On the other hand, it is susceptible to noise, a time axis shift, and requires one-to-one matching (both signals have similar time durations) [90].

To overcome these constraints, DTW proposes a one-to-many matching between time axes without considering local and global shifting issues in time series data. By resolving this time scale problem (local shift), it is possible to match time series data that have a similar pattern but have a different time axis [91].

The DTW distance can be implemented using dynamic programming, in which the accumulated distance formula recursively determines the optimal warp path between two segments $X_i = [x_1 \dots, x_M]$ and $Y_i = [y_1 \dots, y_N]$:

$$\mathrm{D}(\mathrm{X_i}, \mathrm{Y_j}) = \delta(\mathrm{x_i}, \mathrm{y_j}) + \min\{\mathrm{D}(\mathrm{x_{i-1}}, \mathrm{y_{j-1}}), \mathrm{D}(\mathrm{x_i}, \mathrm{y_{j-1}}), \mathrm{D}(\mathrm{x_{i-1}}, \mathrm{y_j})\} \quad (5-7)$$

In which $\delta(x_i, y_j)$ is the Euclidean distance between the two aligned warp path segments [92]. It can be used to classify among time series data using methods like SVM or kNN which have been validated as effective in the field of time series analysis.

To determine how well our proposed method performs for classifying surgemes, we compare the predicted labels with the ground truth labels using various performance metrics, i.e., the average of accuracy, f-measure, and recall.

5 . 3 Experimental Evaluation

5. 3. 1 Surgemes Mean Feature Reduction Representation

Instead of using the entire segment time frames for each variable, we compute the mean for every single of the 76 variables (features) and then normalize it to re-represent the surgemes or segments of each trial for each of the surgical tasks.

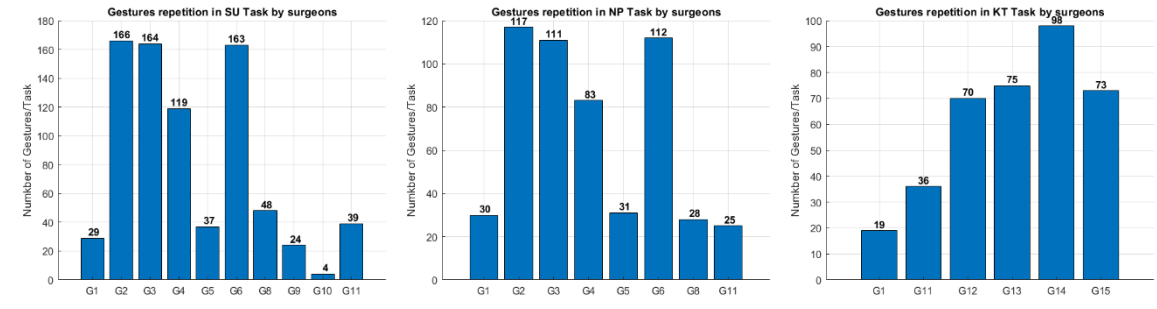
Let $\mathbf{X} = [\boldsymbol{x_{ij}}]_{\boldsymbol{NxP}}$ be a gesture of length *N* and have *P* features. Then the averaging of surgical gestures will be as the following:

$$\widehat{X} = [\hat{x}_{ij}]_{1xP} = \text{Mean}(X) = \frac{1}{N}\sum_{i=1}^{N} x_{ip} \quad , \forall\, p \in P \qquad (5-8)$$

The computation is made easier by mean feature reduction, which also eliminates the bias caused by time. More importantly, by using this approach, we can avoid using DTW or something similar because we can now measure the similarity between surgemes with the same dimension rather than using a different multidimension distance measuring method.

5. 3. 2 Surgical Dataset

The proposed approach is tested on a widely used public robotic surgery dataset, JIGSAWS [11]. This dataset contains kinematic and video data from eight surgeons of varying expertise: expert, intermediate, and novice. Each surgeon repeated the three basic surgical tasks (suturing, needle passing, and knot tying) five times. Raw kinematic data from the da Vinci robotic surgery system was used with different gestures frame lengths. Each of the four manipulators has 76 variables, consisting of 3 cartesian positions, 9 rotation matrices, 3 linear velocities, 3 angular velocities, and 1 gripper angle. Figure 5.4 shows the histogram plot of the numeric gestures grouped in each task.



(a) Suturing task (b) Needle passing task (c) Knot-tying task

Figure 5.4: The frequencies of each gesture in every surgical task.

For each task, the JIGSAWS provided manually annotated ground truth segments (surgemes). In kinematic data attributed to each trial, each predefined surgeme specifies the start and the end frames. Table 5.I lists 15 elements of the common vocabulary of potential surgemes for all three surgical tasks. Some surgemes appear to perform multiple tasks, but they differ in the context of the surgical gestures. Suturing, needle passing, and knot tying are three surgical tasks that include 10, 9, and 6 of the 15 surgemes [11].

Table 5.I: Surgemes Vocabulary for all the surgical tasks [3].

Surgeme index	Gesture description	Suturing	Needle Passing	Knot Tying
G1	Reaching for needle with right hand	☑	☑	☑
G2	Positioning needle	☑	☑	
G3	Pushing needle through tissue	☑	☑	
G4	Transferring needle from left to right	☑	☑	
G5	Moving to center with needle in grip	☑	☑	
G6	Pulling suture with left hand	☑	☑	
G8	Orienting needle	☑	☑	
G9	Using right hand to help tighten suture	☑		
G10	Loosening more suture	☑		
G11	Dropping suture at end and moving to end points	☑	☑	☑
G12	Reaching for needle with left hand			☑
G13	Making C loop around right hand			☑
G14	Reaching for suture with right hand			☑
G15	Pulling suture with both hands			☑

To compare the accuracy results with state-of-the-art methods for gestures classification, we employ three distinct validation schemes: k-fold, leave one supertrial out

(LOSO), and leave one user out (LOUO). In k-fold cross-validation, data is divided into k equal parts and run k times with a different holdout set each time and report the average accuracy of the k times results. For LOSO setup, trials i-th from all the surgeons hold out for testing and training on the rest four trials for each task. In the LOUO validation, we held entire surgemes that belong to one surgeon trials out for testing and training on the remainder [11].

5 . 4 Results and Discussions

5. 4. 1 Surgemes Classification Results by DTW

In each task, we experimented with assessing recognition of surgical surgemes. We used the DTW distance in conjunction with the kNN classifier to distinguish between surgemes using all 76 kinematic variables of the gestures. The results of the LOSO validation setup are shown in Figure 5.5. The model correctly identified the SU and KT tasks with a percent accuracy of 87.3% and 77%, respectively, while correctly identifying the NP task with a percent accuracy of 67.6%. We might have noticed that some surgemes, i.e., 3,4, 6 of SU, NP, and 11, 12, 13 of KT tasks, have the highest accuracy. However, gestures 5, 8, and 10 have the minimum accuracy. We can explain it because the dataset has few numbers of these gestures as shown in Figure 5.5 and that needle passing is more difficult than other tasks. For instance, gesture 8 is not involved in suturing or needle passing. So, if the surgeon cannot complete the gesture, such as inserting the needle into the tissue, they might perform a to reorient the needle and repeat the gesture. As a result, the model is having difficulty accurately recognizing it [12].

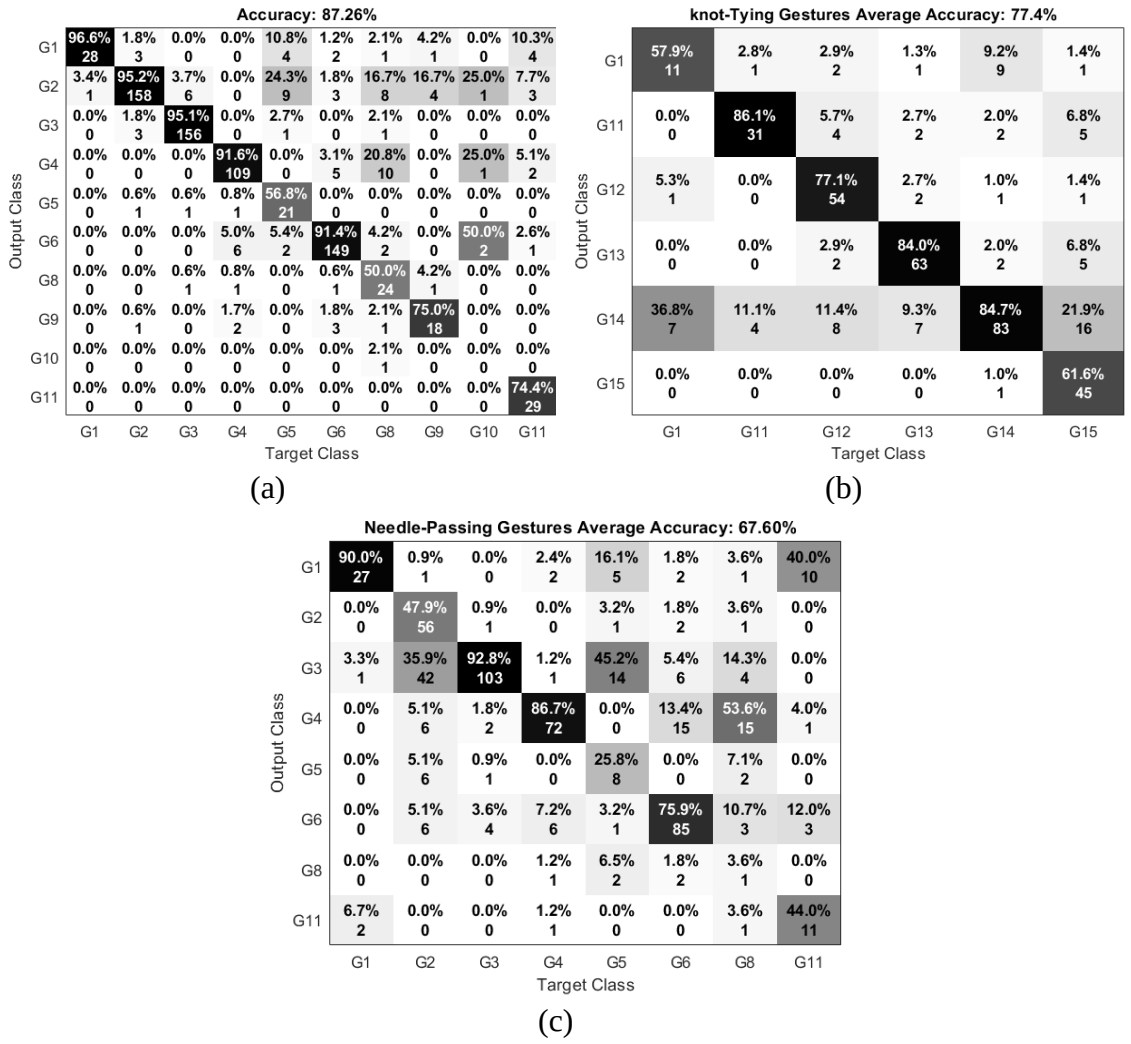


Figure 5.5: Confusion matrix in LOSO validation using kNN-DTW for a) SU b) KT and c) NP tasks.

5 . 5 Surgemes Results by Mean Feature Reduction

As mentioned previously, we convert raw surgemes with various observation lengths into gestures with similar dimensions by applying the mean feature reduction. This section shows the results for gestures classification by using SVM with quadratic kernel function to compute the classifiers for modeling the different surgemes.

Table 5.II shows the 10-fold cross-validation supervised metrics that split the dataset into ten groups and randomly select one group to be used for testing and training the other

nine groups of the dataset. Finally, provide an estimate of the average assessment performance of the ten folding. The confusion matrix for the classification results is depicted in Figure 5.6. Also, the percentage of the number of correctly and incorrectly surgemes classified for true (rows) and predicted (columns) classes is displayed in the exact figure as well. As expected, the results achieved high scores in k-fold because the data was rearranged and partitioned randomly. Consequently, all the observations are used for both training and validation.

Table 5.II: 10-fold results of the average accuracy for SVM Classifier.

Task	Accuracy	f-measure	Specificity	Sensitivity	Precision
SU	98.4	98.4	99.8	98.4	98.4
NP	95.2	94.3	99.3	95.8	93.2
KT	97.6	97.3	99.5	98.2	97.6

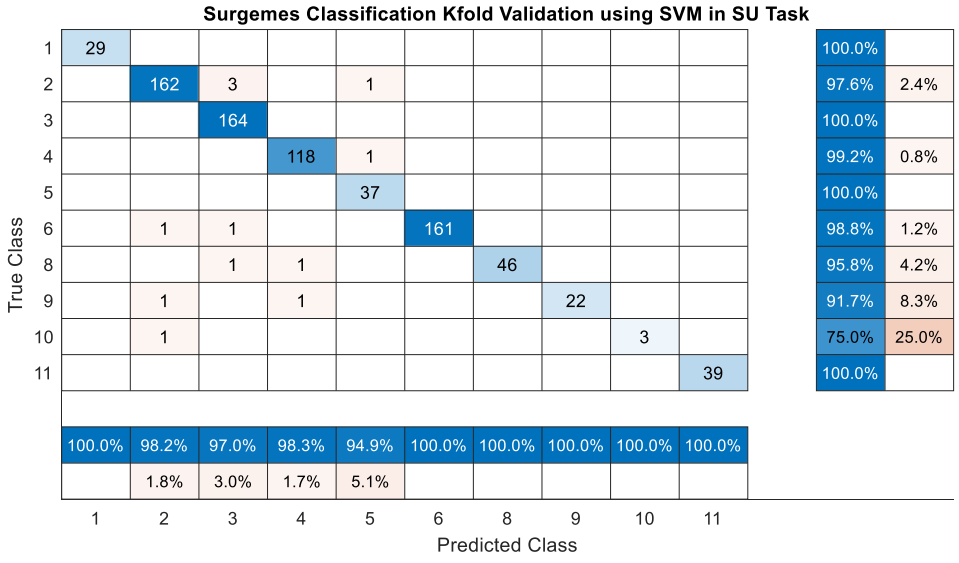


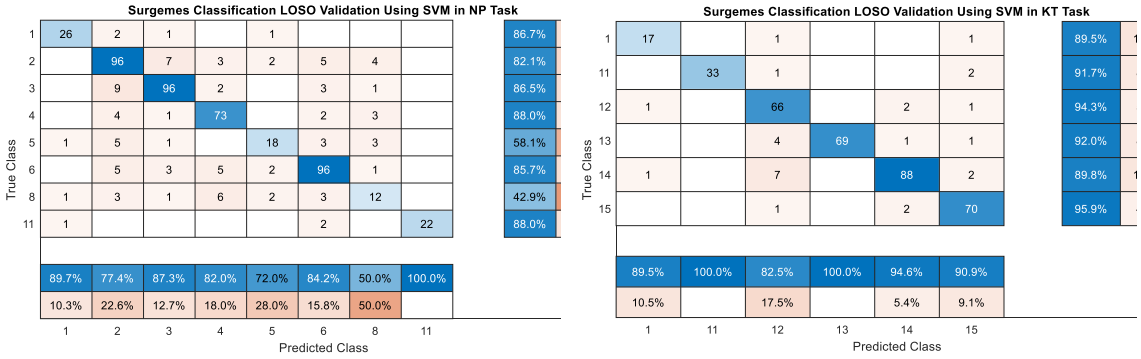
Figure 5.6: Confusion matrix for 10-fold validation for SU task.

We experimented with the LOSO cross-validation to examine the robustness of the model when leaving the i-th trial out from all surgeons for testing and training on the remainder gestures. We used the ground truth annotation data for comparison purposes with the predicted labels by our approach. Table 5.III summarizes the average accuracy results of the SVM method using the mean feature reduction technique for the surgemes for every surgical task. From this table, we can observe that the approach can recognize the surgemes with an average recognition rate of 94.8%, 81.8%, and 92.5% for SU, NP, and KT tasks.

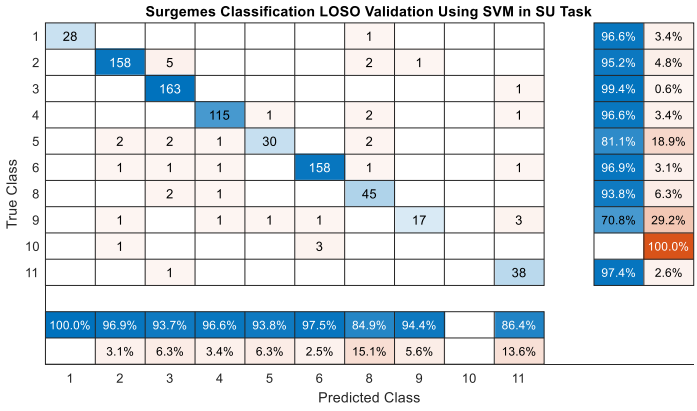
Table 5.III: The average accuracy results of the SVM method using mean feature reduction.

Task	Accuracy	f-measure	Specificity	Sensitivity	Precision
SU	**94.8**	94.8	99.4	94.8	94.8
NP	**81.8**	81.8	97.4	81.8	81.8
KT	**92.5**	92.5	98.5	92.5	92.5

Figure 5.7 shows the confusion matrix comparing the classification results with the ground truth for each surgeme in each task. Interestingly, the overall results of the SVM model of the SU task achieved high accuracy and turned out to be highly discriminating between the gestures. However, we can notice that surgeme ten is misclassified in the SU task. We can explain that this gesture is achieved only four times by three novice surgeons in the SU task. Three of these gestures were performed in the same trial number along with different surgeons. In contrast, one surgeon worked this gesture in other trial numbers. Thus, while experimenting with the LOSO scheme, three of these motions are kept separate from the other for testing or training purposes, resulting in misclassification.



(a) NP (b) KT



(c) SU

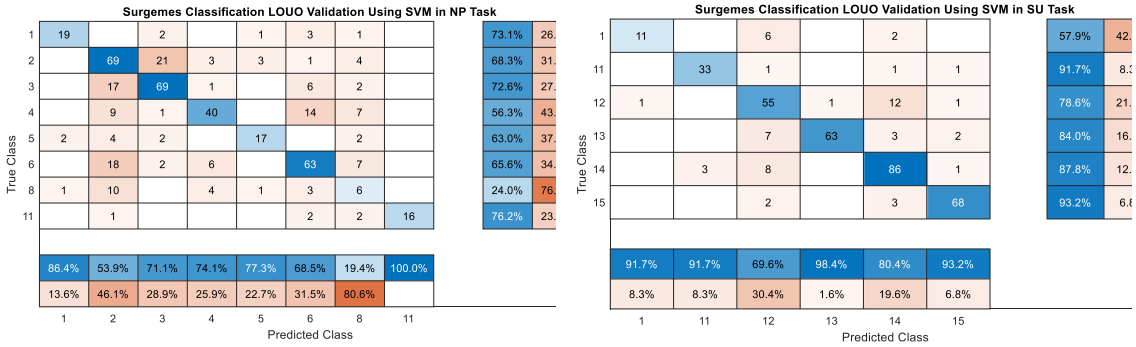
Figure 5.7: Confusion matrix for SVM model using LOSO validation.

Another experiment explores our proposed method for the surgemes when the entire gestures that belong to a surgeon are not seen in the training. We evaluate our approach based on mean feature reduction in the LOUO cross-validation scheme. We can notice from Table 5.IV and Figure 5.8 that the accuracy results using this validation setup decreased by approximately 7.7, 17, and 7 percent of accuracy when switching from LOSO to this scheme for SU, NP, and KT tasks.

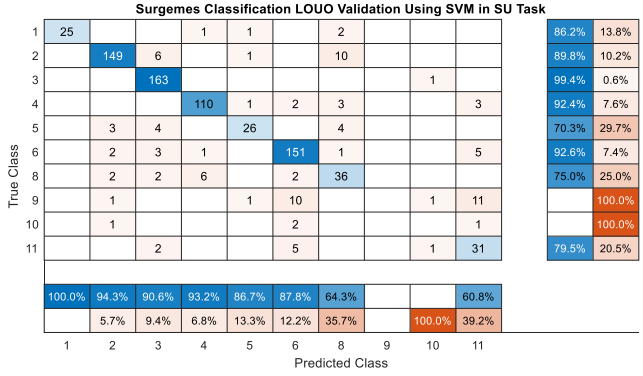
Table 5.IV: Average accuracy of the SVM classification method for LOUO validation in all the surgical tasks.

Task	Accuracy	f-measure	Specificity	Sensitivity	Precision
SU	87.1	87.1	98.6	87.1	87.1
NP	64.7	64.7	95	64.7	64.7
KT	85.2	85.2	97	85.2	85.2

This disparity indicates that this cross-validation is very difficult to achieve high accuracy. Additionally, the considerable reduction in accuracy for the Needle passing task accuracy can be clarified due to the limited available surgemes for this specific task in the dataset compared to the other surgical study. Furthermore, this variation in tasks implies that needle passing, and suturing are performed differently depending on the surgeons.



(a) NP (b) KT



(c) SU

Figure 5.8: Confusion matrix for SVM model using LOUO validation.

We compare the result of our approach using the mean feature reduction technique with the state-of-the-art methods in average accuracy for LOSO in Table 5.V and LOUO in table 5.VI. As we can see that our method outperforms the other methods using kinematic data only for LOSO and LOUO cross-validations in all the tasks. An exception for the needle passing task is that the linear dynamic system (LDS) based on multiple kernel learning proposed by [13] achieved better accuracy. The earlier methods rely on the video information to recognize the gestures or a combination of both kinematic and video data to provide higher accuracy results.

Table 5.V: Comparison with state-of-the-art methods using **LOSO** Validation.

	Kinematic				Video		Hybrid	
Task	Our Proposed	SHMM [24]	LDS [44]	BOF	LDS	BOF+LDS	BOF+LDS (Kin)	BOF+LDS (all)
SU	94.8	79.4	87.3	90.7	87.2	91.8	93.5	94
NP	81.8	76.4	78.8	74.1	69	77.8	85.3	86
KT	92.5	86.8	85.1	88.4	87.3	90.8	93.8	92.8

Table 5.VI: Comparison with state-of-the-art methods using **LOUO** Validation.

	Kinematic				Video		Hybrid	
Task	Our Proposed	SHMM [24]	LDS [44]	BOF [44]	LDS [44]	BOF+LDS [44]	BOF+LDS (Kin) [44]	BOF+LDS (all) [44]
SU	87.1	60.9	74.6	78	74.2	81.2	86.3	86.6
NP	64.7	45.3	67.3	65.5	85.8	66.9	80.1	80.2
KT	85.2	72	78.9	84.9	77.4	86.7	90.1	90.4

It is worth mentioning that our approach is more precise and fast in comparison to other existing methods. One of the key benefits is the mean feature reduction of the gestures, which reduces the computational time and the complexity of the technique while maintaining almost the same high-performance results.

5 . 6 Conclusion

We developed a framework for surgemes classification directly on raw kinematic data captured by RMIS. Also, we present the mean feature reduction of the surgical gestures that is fast, accurate, and reduce the complexity of the proposed framework. Additionally, the combination of the mean feature reduction and the SVM method provides the highest performance results. Though, the feasible reasonable outcomes of the proposed technique make it applicable in another domain like recognition of human activities. Although further studies are needed to draw a solid conclusion with more data and surgical tasks with a balanced number of surgemes in each task.

CHAPTER 6 **Clustering Surgemes using Prototypes from Robotic Kinematic Information**

6 . 1 Introduction

The innovations in surgical robotic platforms have wide opened new capabilities in training and education for surgeons to provide high quality surgical care in the operation room. The information captured by robotic minimally invasive surgery (RMIS) deliver program-based insights that could potentially help enhance patient outcomes and care cost [93]. Also, the accessibility of the driven data representing movement of the surgeons give the opportunities to create and build models for objective method and assessments that deliver feedback during a surgical task [52].

Figure 6.1 presents an example of 3D movements of both hands of trainee and expert surgeon for one trial during a surgical suturing session from JIGSAWS dataset [11] where the surgemes are highlighted using different colors. In this figure we can noticed the difference in traces shape between them in which the movements of the expert are smoother and confidence in transition than the novice surgeon.

Figure 6.2 illustrates an example comparison of the raw kinematic data for the right and left hands among the novice and expert surgeon of the same trial in previous figure for each dimension of the cartesian coordinates (x-y-z coordinates) individually. Also, the sequences of the surgical gestures for each trial are provided on Figure 6.2 for each surgeon separately.

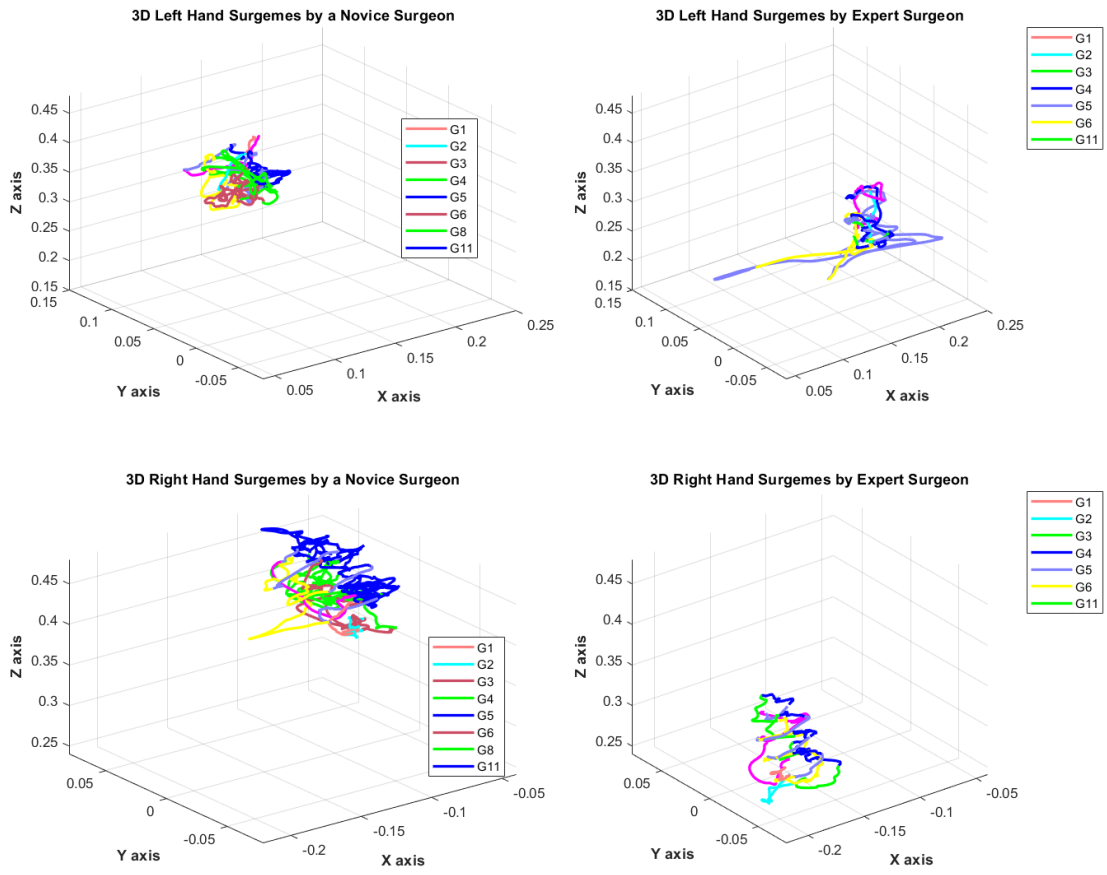
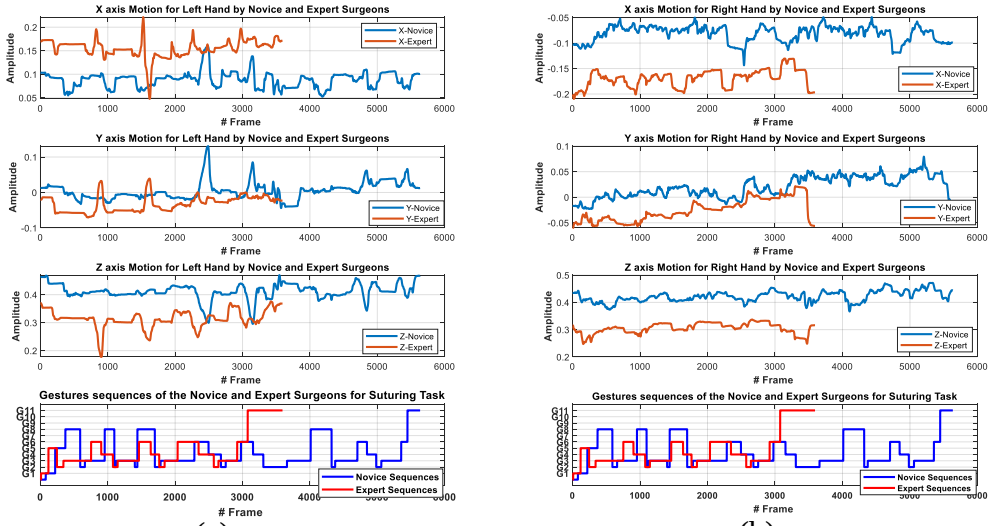


Figure 6.1: Example of 3D Movements for (a) the left and (b) the right hand of the novice and expert surgeons during a suturing task.



(a) (b)

Figure 6.2 Kinematic time series for (a) the left and (b) right hands of the novice and the expert surgeons.

In the literature, most of the techniques are supervised classification based on predefined or pre-segmented surgeme data. These surgical gestures (surgemes) are annotated manually by chief surgeons, consuming more time and being susceptible to human mistakes by missing parts (surgemes) or inconsistently criteria applied throughout a surgical task [12, 50]. Several works intended to identify surgical activities from unsupervised viewpoints without prior knowledge of gestures [51-53]. In [51], they proposed a framework for segmentation and recognition surgemes from kinematic data. They first applied unsupervised segmentation by finding a relevant selection of dexemes (a numerical representation of subgestures to perform a surgemes). Secondly, they used learning features from dexemes to associate them to corresponding surgemes (composed of a set of dexemes) [51]. Another approach introduced by [54] is known as soft boundary unsupervised surgemes segmentation. The temporal sequence of surgemes segment and merge based on some criteria, and then the boundaries between parts are smoothed. A

recent deep-learning approach was proposed by [55], based on a deep convolution network using both kinematic and video data for surgical gestures segmentation.

The main objective in this chapter is to build and assess an unsupervised model to identify the surgemes of the surgeon directly from raw kinematic data, which is a step forward toward segment the surgemes from the surgical traces data. The proposed algorithm is comprised of three major mechanisms: 1) representatives surgemes based on clustering the surgemes of one expert surgeon from every single surgical training trial performed by this expert, 2) mapping the other 'surgeons' surgemes to the nearest representative prototypes, and finally 3) measure the clustering accuracy using the rand-index.

The aim of this chapter is two folds are as follows: i) we propose an unsupervised method to identify the surgemes of the surgeons based on clustering algorithms. ii) we re-represent the segments by utilizing the mean of the feature instead of using all the time frames to reduce the complexity and computational time and make the proposed approach more feasible.

6 . 2 Methodology

6. 2. 1 Surgemes Clustering Framework

Figure 6.3 shows the overall flow diagram of our proposed approach for surgical gestures clustering based on raw kinematic data. First, the surgemes prototypes were obtained from the expert surgeon by clustering their surgemes from all trials using unsupervised algorithms. In the second step, we utilize the medoid method to locate the representative surgeme for each surgeme prototype individually. Next, we mapped every

gesture of the trainee surgeon per trial on the representative surgemes by measuring the distance with all the representative gestures and assigned it to the cluster with the smallest distance. Finally, we assess the performance of our unsupervised approach using the rand-index between the ground truth and the predicted labels of the clustering approach. More details about each step will be discussed in the following sections.

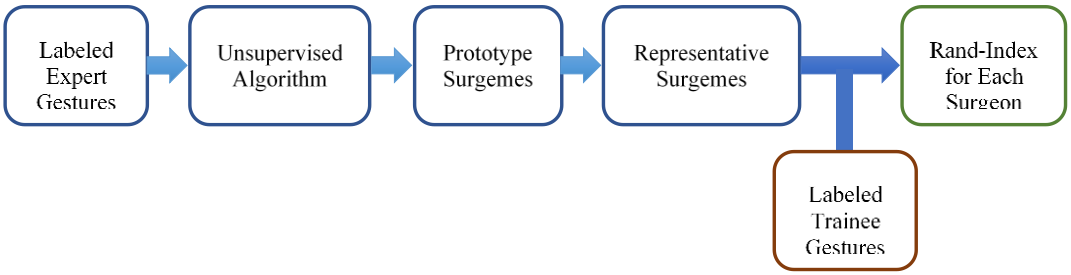


Figure 6.3: Overview of our Clustering Surgemes approach using Rand-Index for each surgeon.

This approach starts by normalizing each surgeme using mean and variance to ensure that the data are scale and shift invariants which allow reasonable comparison between them.

Let X_i be a time series, then the corresponding normalized signal $\hat{X}_i$ is:

$$\hat{X}_i = \frac{(X_i - \mu_i)}{\sigma^2{}_i} \qquad (6-1)$$

Where μ_i and σ_i are the arithmetic mean and standard deviation of time series i, respectively. Next, to form the prototype surgemes, we employ unsupervised methods on one expert surgeon gestures using hierarchical and Fuzzy c-means (FCM) Algorithms.

The hierarchical clustering method has been shown to be effective and efficient at separating human activities, which is well-suited for time series clustering. [94, 95]. Then, we employ the minimum variance algorithm (Ward) on a pairwise distance matrix which is obtained by computing the distance between two segments to create the prototype

surgemes. The distance d_{rs} between two clusters C_r and C_s is defined as the distance of their centroid is equivalent to the following equation:

$$d_{rs} = \frac{n_r n_s}{n_r + n_s} \left\| \overline{X}_r - \overline{X}_s \right\|^2 \qquad (6-2)$$

Where $\overline{X}_r$ and $\overline{X}_s$ are the centroids of the two clusters, $n_r$and n_s are the number of objects in cluster C_r and C_s, respectively [96].. Additionally, we used fuzzy c-means (FCM) to partition the expert surgeon's surgemes into a predefined number of clusters equal to the number of distinct surgemes in each surgical task.

The FCM partition membership needs to meet the following constraint to prevent the trivial solution by allocating all the cluster memberships to zero:

$$\sum_{i=1}^{C} u_{ij} = 1 \qquad \forall j = 1,2 \ldots n \qquad (6-3)$$

The objective function of the FCM that meets the criteria can be formulated as follow:

$$J(U,V) = \sum_{j=1}^{n} \sum_{i=1}^{C} u_{ij}^{m} \, d^2(x_j, v_i) \qquad (6-4)$$

The parameter *m*>1 is the fuzzifier that controls the rate of the membership value. The values of partition membership u_{ij} and prototype centers that require to minimize J and the distance between data sample x_j and the set of cluster centers v_i can be determined by the following equations [63]:

$$u_{ij} = \frac{(1/d(x_j, v_i))^{2/(m-1)}}{\sum_{k=1}^{C} (1/d(x_j, v_k))^{2/(m-1)}} \qquad (6-5)$$

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (6-6)$$

The distance measure plays an important role in time series clustering, where the data are grouped based on their similarity. The Euclidean distance and dynamic time warping (DTW) are the most frequently used similarity measurements in this domain due to their effectiveness and efficiency in determining the similarity between objects. Euclidian distance is simple, fast, and parameter-free. However, it is sensitive to noise and shift in the time axis. It also requires both signals to have equal time lengths (one-to-one matching) [90]

On the other hand, DTW employs a one-to-many matching between time axes without considering local and global shifting issues in the time series data which overcome these restrictions. By resolving this time scale issue (local shift), it is possible to match time series data that are similar in pattern but have a different time axis.[91].

DTW distance can be implemented using dynamic programming in which the accumulated distance formula recursively computes the optimal warp path between two segments $X_i = [x_1 \dots, x_M]$ and $Y_i = [y_1 \dots, y_N]$:

$$D(X_i, Y_j) = \delta(x_i, y_j) + \min\{D(x_{i-1}, y_{j-1}), D(x_i, y_{j-1}), D(x_{i-1}, y_j)\} \qquad (6-7)$$

where $\delta(x_i, y_j)$ is the Euclidean distance between the two aligned segments of the warp path (that give the minimum distance) [92].

To evaluate how well our proposed approach works for clustering surgemes, we compare the resulting predicted labels with ground truth labels using the rand-index. The higher the rand-index value, the better performing.

6. 2. 2 Rand-Index Performance Evaluation

It is not an easy task for unsupervised performance evaluation to assess the cluster results with the absence of data labels. However, for the JIGSAWS dataset, the manually segmented references were provided by a senior specialist in robotic surgery.[11]. The most common quality measure in the domain of time series clustering is the Rand index [97]. We used the Rand index criteria between the predicted result labels of our proposed framework and the ground truth surgeme labels. The Rand index values are between 0 and 1, where one indicates the two surgemes are identical or precisely the same. The Rand index criteria between the ground truth labels and the predicted labels is defined as the measure of the ratio of the correct decisions taken by the approach. In other viewpoints, it can be defined as the number of agreements between two groups G and Y over the total number of pairs (agreements and disagreements), which can be calculated using the following equation [97, 98]:

$$\mathrm{RI}(\mathrm{G},\mathrm{Y}) = \frac{\mathrm{T_P} + \mathrm{T_N}}{\mathrm{T_P} + \mathrm{F_p} + \mathrm{T_N} + \mathrm{F_N}} \qquad (6-8)$$

where T_P, F_p, T_N and F_N are the corresponding number of true positives, false positives, true negatives, and false negative results, respectively.

6. 2. 3 Calinski Harabasz Index

The variance ratio criterion (VRC), known as the Calinski Harabasz (CH) index, is computed for *K* clusters and *N* data point as:

$$VRC = \frac{trace_B}{trace_W} \times \frac{(N-K)}{(K-1)} \qquad (6-9)$$

Where $trace_B$ and $trace_W$ are the overall between-cluster variance and within cluster variance, respectively.

The overall between-cluster $trace_B$ can be written as [99]:

$$trace_B = \sum_{i=1}^{K} n_i \|C_i - C\|^2 \qquad (6-10)$$

Where $C_i$is the centroid of cluster i, n_i is the number of observations in cluster i and C is the centroid of the entire sample data.

The overall within-cluster variance $trace_W$ is defined as:

$$trace_W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \|x_j - C_i\|^2 \qquad (6-11)$$

Clusters that are well-defined clusters will have a high variance between-clusters variance $trace_B$ and a small variance within-cluster $trace_W$. The higher the CH ratio, the better the data partitioning is going to be. The solution with the highest Calinski-Harabasz index value is the one that has the optimal number of clusters [99, 100].

6. 2. 4 Xie-Beni Validity Index

The ratio of the compactness of the fuzzy c-partition to its separation is called the compactness and separation validity function or well-known as Xie-Beni index, which can be computed as [101]:

$$XB = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^2 \|V_i - X_j\|^2}{n \min_{i,j} \|V_i - V_j\|^2} \qquad (6-12)$$

Where n is the number of data points. The fuzzy centroid V_i $(i = 1, 2 \ \ldots c)$ and the fuzzy membership u_{ij} of X_j belonging to cluster i are calculated using

$$V_i = \frac{\sum_{j=1}^{n} u_{ij}^m \, X_j}{\sum_{j=1}^{n} u_{ij}^m} \qquad (6-13)$$

$$u_{ij} = \frac{\left(\frac{1}{\|X_j - V_i\|}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{c}\left(\frac{1}{\|X_j - V_i\|}\right)^{\frac{1}{m-1}}} \qquad (6-14)$$

A lower value of XB implies a partition in which all the clusters are compact and separate. Thus, the smaller values of XB correspond to the optimal number of clusters [99, 101].

6 . 3 Experimental Results

6. 3. 1 Dataset

The experiment on the proposed approach is conducted with a general, and public robotic surgery dataset indicated to it as JIGSAWS The proposed method is tested using a widely used and publicly available robotic surgery dataset dubbed JIGSAWS. [11]. This dataset includes both kinematic and video information from eight different expertise surgeon levels: expert, intermediate, and novice surgeons. Each surgeon performed three basic surgery tasks (suturing, needle passing, knot tying) five times (known as a trial) repetitively. We used only the raw kinematic data captured at 30Hz from the da Vinci robotic surgery system with different trial frame lengths. There are 76 dimensions or variables information to describe the kinematics for all four manipulators. Each manipulator has 19 variables that consist of 3 cartesian positions, 9 rotation matrices, 3

linear velocities, 3 angular velocities, and 1 gripper angle. Table 6.I describes the details of the variables included in the kinematic dataset [11].

Table 6.I: JIGSAWS Kinematic variables [3].

	Variable indices	Number of variables	Description	Manipulator name
1	1-3	3	Tool tip position (x, y, z)	Left MTM
2	4-12	9	Tool tip rotation matrix (R)	Left MTM
3	13-15	3	Tool tip linear velocity $(\hat{x}, \hat{y}, \hat{z})$	Left MTM
4	16-18	3	Tool tip rotational velocity $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$	Left MTM
5	19	1	Tool tip gripper angle velocity (θ)	Left MTM
6	20-38	19	All 1-5 rows in this table	Right MTM
7	39-41	3	Tool tip position (x, y, z)	PSM1
8	42-50	9	Tool tip rotation matrix (R)	PSM1
9	51-53	3	Tool tip linear velocity $(\hat{x}, \hat{y}, \hat{z})$	PSM1
10	54-56	3	Tool tip rotational velocity $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$	PSM1
11	57	1	Tool tip gripper angle velocity (θ)	PSM1
12	58-76	19	All 7-11 rows in this table	PSM2

The JIGSAWS has manually annotated ground truth segments (surgemes) for each trial at every task. Each annotation provides the label of the surgeme, the start, and the end frames in the kinematic data allocated for each trial. In particular, the common vocabulary of potential surgemes is made up of 15 elements and is listed with their description in Table 6.II for the three surgical tasks. Some surgemes seem to be in more than a single task; however, the background environment differs between tasks [11]. Even though there are a combined total of 15 surgemes, not necessarily all of them show up in one surgical task.

That is why suturing, needle passing, and knot tying include 10, 8, and 6 of the 15 surgemes, respectively.

Table 6.II. Surgemes Vocabulary for all the surgical tasks [11].

Surgeme index	Gesture description	Suturing	Needle Passing	Knot Tying
G1	Reaching for needle with right hand	☑	☑	☑
G2	Positioning needle	☑	☑	
G3	Pushing needle through tissue	☑	☑	
G4	Transferring needle from left to right	☑	☑	
G5	Moving to center with needle in grip	☑	☑	
G6	Pulling suture with left hand	☑	☑	
G8	Orienting needle	☑	☑	
G9	Using right hand to help tighten suture	☑		
G10	Loosening more suture	☑		
G11	Dropping suture at end and moving to end points	☑	☑	☑
G12	Reaching for needle with left hand			☑
G13	Making C loop around right hand			☑
G14	Reaching for suture with right hand			☑
G15	Pulling suture with both hands			☑
Number of Gestures in each task		**10**	**8**	**6**

6. 3. 2 Surgemes Representation and Visualization

We re-represent the surgemes of each trial for every surgical task by computing the mean of each 76 variables separately (rather than using the whole segment time frames for each component, as shown in Table 6.II.

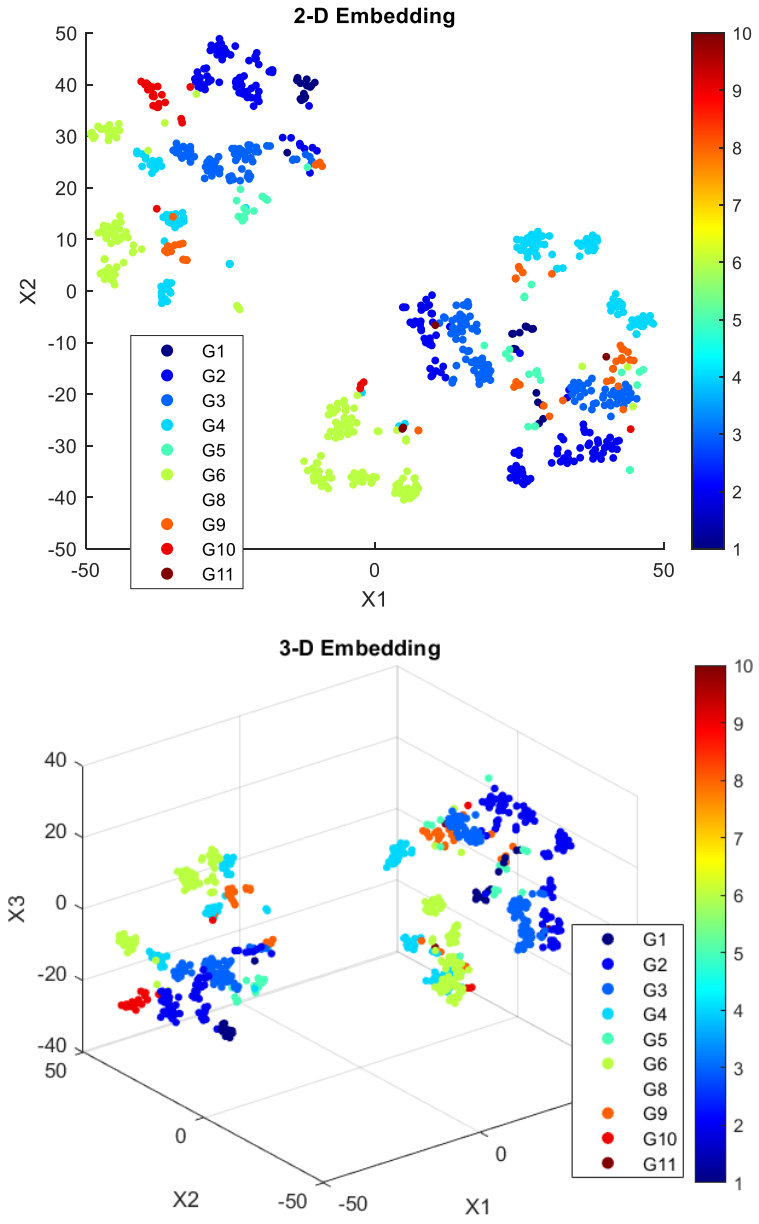
Let $\mathbf{X} = [\boldsymbol{x_{ij}}]_{\boldsymbol{NxP}}$ be a surgeme of length *N* and has *P* dimension variables (features). Then the averaging of surgical gestures will be as the following:

$$\widehat{\mathrm{X}} = [\hat{x}_{ij}]_{1xP} \; = \; Average(\mathrm{X}) = \frac{1}{N}\sum_{i=1}^{N} x_{ip} \quad , \forall \, p \in P \qquad (6-9)$$

Averaging each segment simplifies the computation and eliminates the bias of the time. Furthermore, this approach allows us to use any similarity measure rather than only DTW since surgemes have identical dimensions.

We visualize the surgemes for all the trials of the suturing task in Figure 6.4 using the t-Distributed Stochastic Neighbor Embedding (t-SNE) by reducing the high-dimensional space of the surgemes to a low-dimensional map of two or three dimensions [102]. Figure 6.4 illustrates a reasonable and precise separation of the surgemes, even though some gestures occur in multiple locations. This is because it is also dependent on the surgeon's skill level.

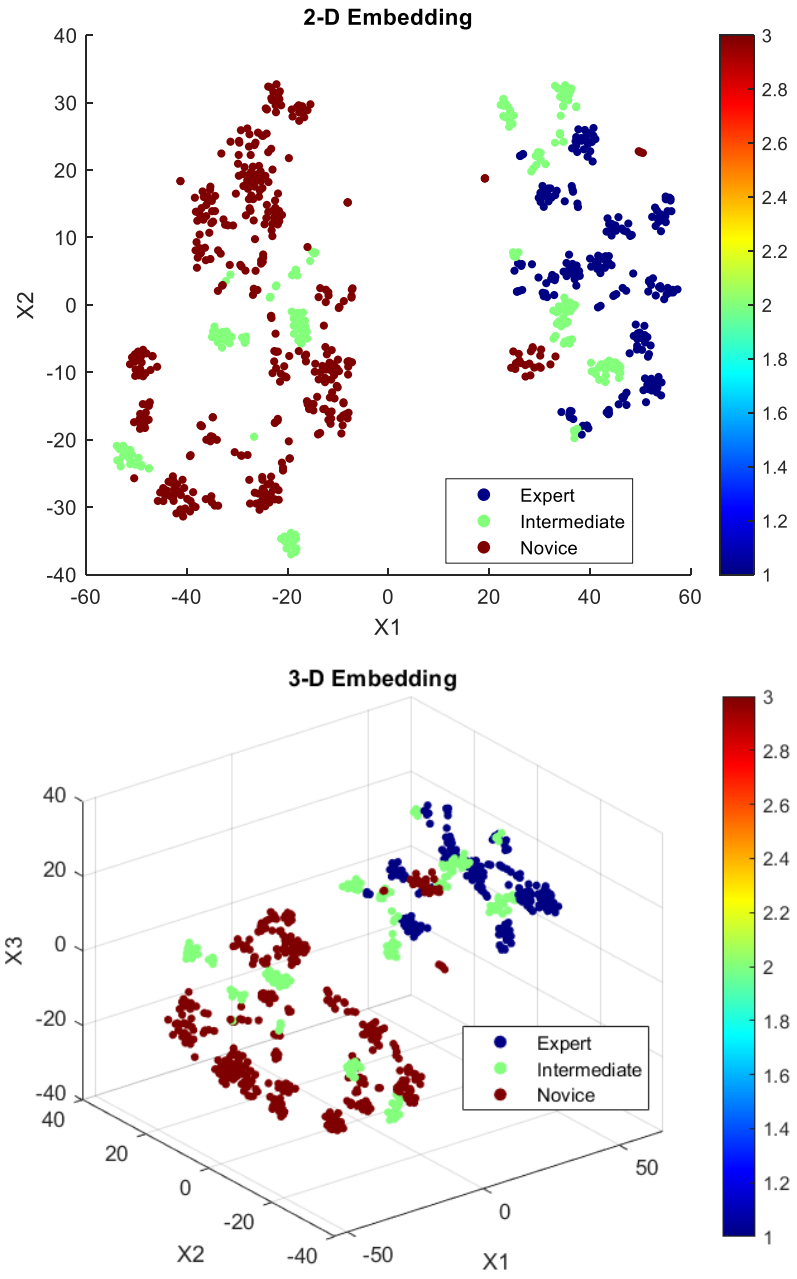
Figure 6.5 visualizes the surgemes labeled with surgeon skill levels instead of the gesture labels. Figure 6.6 illustrates the average modified global rating score (GRS) annotated by a vast experience robotic surgeon for every single surgeon trial during a suturing surgical training task [11]. Although the average score of some expert surgeons was lower than that of some intermediate and novice surgeons, the expert surgeon's consistency is evident through attempts. Back to Figure 6.5, this explains why there are some surgemes of the same class but located in different positions in the reduced feature space.



(a)

(b)

Figure 6.4. t-SNE visualization of surgeme labels in Suturing task (a) 2-dimension, and (b) 3-dimension embedding.



(a)

(b)

Figure 6.5. t-SNE visualization of surgeme using surgeon's skill levels as labels in Suturing task (a) 2-dimension, and (b) 3-dimension embedding.

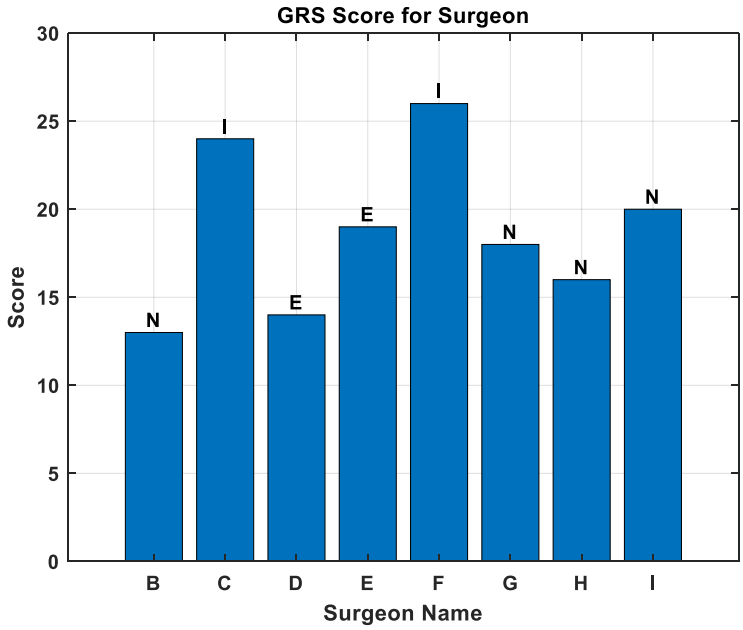


Figure 6.6. GRS Average scores for each surgeon in Suturing task showing the expertise levels on the top of each bar.

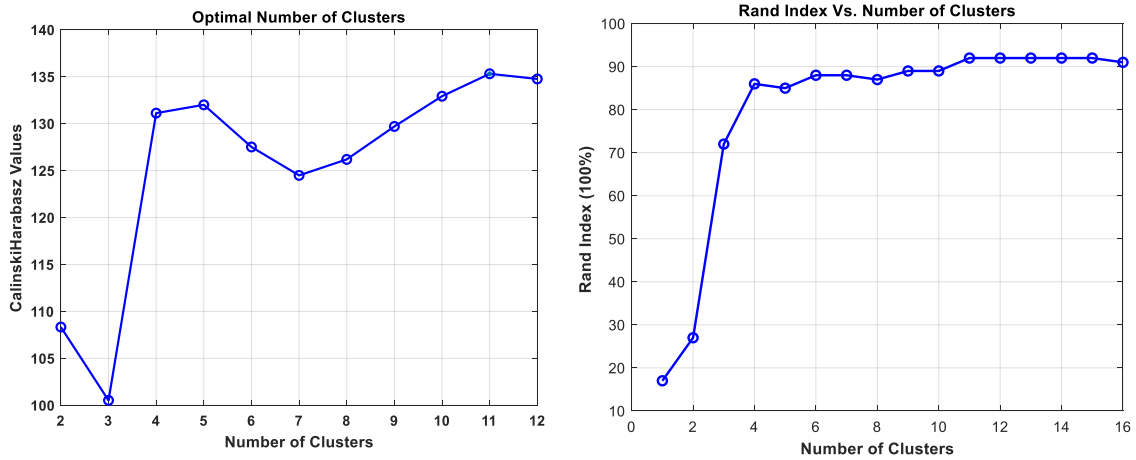
6. 3. 3 Hierarchical Clustering Results

We develop our surgeme clustering approach directly onto the raw kinematic data to prevent excessive pre-processing. We run three experiments based on hierarchical clustering (with Ward linkage) on the JIGSAWS dataset. In the first two sets, we used the DTW as a distance measure between the surgemes, while in the third set, we employed the Euclidean distance. The results are reported based on the framework discussed earlier in Figure 6.3 by applying the unsupervised learning approach.

In the first experiment, we cluster all the surgemes of one expert surgeon utilizing the Ward Hierarchical clustering to find the prototype surgical gestures for each cluster. The raw surgemes are represented as a time series with 76-dimensional and different time lengths. We have chosen the surgemes of an expert surgeon who has the highest GRS

scores among the expert surgeons displayed in Figure 6.6. This is because the higher scores of expert surgeons have resulted from the consistency and fluidity of their trajectories during the surgical task, leading to a better clustering outcome. Hence, we employed DTW as a practical and feasible pairwise distance measure between any two surgemes in this case.

The rand-index is used to measure the agreement between the ground-truth and clustered gestures. But first, we evaluate the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion as illustrated in Figure 6.7(a). The plot shows that the highest Calinski-Harabasz value occurs at eleven, suggesting that the optimal number of clusters is eleven in this case of Ward linkage clustering. Figure 6.7(b) shows the Rand-Index values change with the number of clusters. We can observe that the highest value of the Rand-Index is achieved when we use eleven clusters in the Hierarchical algorithm that is fitting with the Calinski-Harabasz criterion. The rand index resulted from clustering the expert data intended for finding the prototypes is 92%.

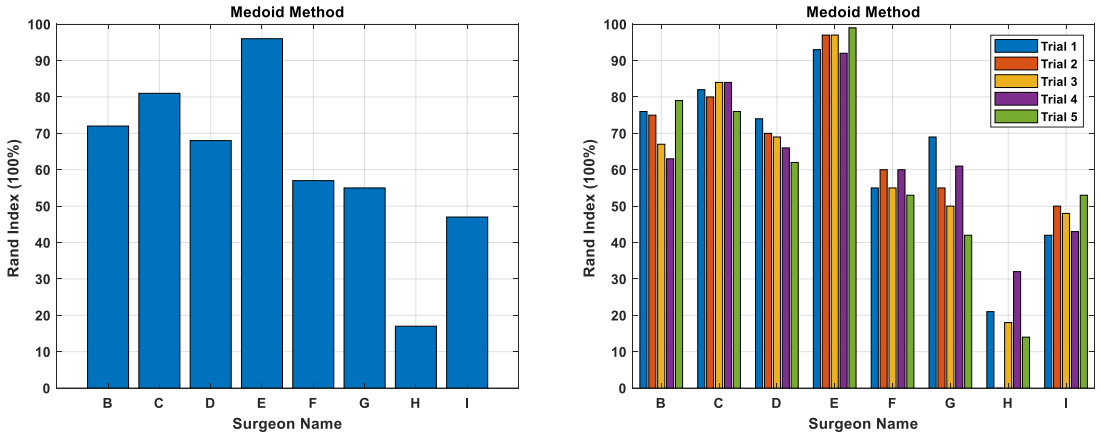


(a) (b)

Figure 6.7: (a) Calinski-Harabasz clustering evaluation criterion, (b) Rand-Index plot as a function of the number of clusters.

At this point, each prototype has candidate surgemes to be a representative surgical gesture within the cluster. We used the medoid to locate the representative for each prototype by computing the DTW distance among the same cluster members and then locate the minimum distance relevant to the elected representative surgeme. The Medoid technique is used here because the surgemes have different lengths. Therefore, employing centers instead of the medoid to assign representatives are not allowable. Now, each cluster has a group of surgemes and one representative surgeme that is representing this group. Finally, we stream every sample point (surgeme) of each trainee surgeon per trial. We measure the DTW distance between each arriving new data sample and the prototypes representative and assign it to the nearest cluster.

Figure 6.8(a) presents the average rand-index results of the proposed method for clustering the surgemes for each trainee surgeon intended for the suturing task. At the same time, the rand-index results per trial for each surgeon are shown in Figure 6.8(b). As expected, the surgeon “E” has the highest average rand-index of 96% accuracy because this surgeon was the prototype clustering surgeon. Also, from this figure, we observed that our proposed method could cluster the surgemes of the surgeon who mapped each trial to the representative surgemes.



(a) (b)

Figure 6.8. Rand-Index Results for Ward clustering using medoid (a) Average Rand Index (b) Rand-Index per trial.

We implemented another experiment by considering each expert's surgeme as a representative member. This can be done by using the expert's ground truth surgemes as a prototype which results in grouping them according to their labels rather than clustering them.

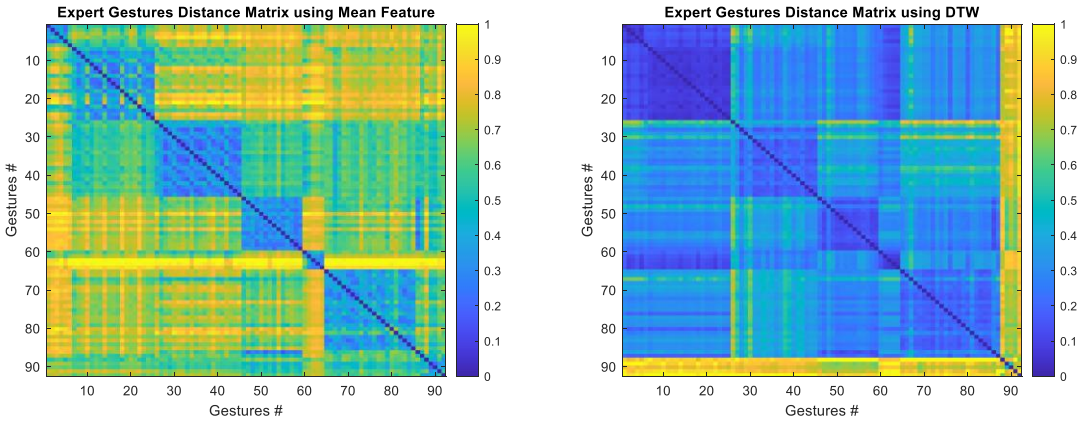
First, we calculate the pairwise distance between all the surgemes of a query surgeon and each group of surgemes that belong to the expert separately. For example, let us have ten surgemes of different labels belonging to the query surgeon, and the expert surgeon has three clusters of different surgemes, each with a cluster size of 2, 4, and 5. Then, we measure the distance between the ten surgemes and the members of each cluster, which results in three distance matrices of the size of 2x10, 4x10, and 5x10 dimensions, respectively. Secondly, we average the distance matrix to each cluster which results in an array of 1x10. Then, we concatenate the resulted mean distance arrays in one matrix and map each surgemes to its closest group. In all the steps of this experiment, we applied DTW as our distance measure. Also, note that the same number of clusters were used in both experiments.

Figure 6.9 illustrates the comparison utilizing the average Rand index between the first experimental results that use the medoid gesture as representative for each cluster and the second experiment that uses all cluster members as representatives for that cluster to assign the labels of the test samples (surgemes). We can observe that the results are very close to each other for most surgeons. Still, the second approach performs better than the medoid because it considers all the members in the cluster, and it reduces the possibility of having a lousy cluster representative. Besides, the second approach uses the ground truth labels to build the prototypes of the expert surgeon instead of clustering, which results in an average Rand index of around 100%, as seen in Figure 6.9, surgeon ("E").

Figure 6.9. Comparison of average Rand-Index between using medoid and mean GT in representing the surgemes in suturing task.

We also investigate the performance of our proposed approach by employing the mean features of the surgeme mentioned before as an alternative to using all the time frames. A surgeme of length (Nx75) using this representation will be mapped to (1x75), making all the surgemes having the same size in the 75-dimensional space. Thus, we can employ

Euclidean distance as our measure instead of DTW. The distance between time series is calculated using Euclidean distance because the sequences are identical in length. Using the DTW is impractical here due to its high complexity. Figure 6.10 compares dissimilarity matrices among gestures by utilizing the mean feature with ED in (a) and using all the time observations along with DTW as a distance. This figure indicates the ability to cluster better with the mean feature rather than using DTW. The histogram of the expert surgemes at the SU task is presented in Figure 6.11. We can observe that the expert surgeon never performed G10, and two gestures (G8 and G9) were performed just one time during the entire five trials. Consequently, it appears that the optimal number of clusters is seven instead of nine for the expert gestures.



(a) (b)

Figure 6.10: Pairwise distance matrices comparison of the expert surgeon surgemes between (a) Mean feature using ED (b) different surgemes lengths using DTW distance.

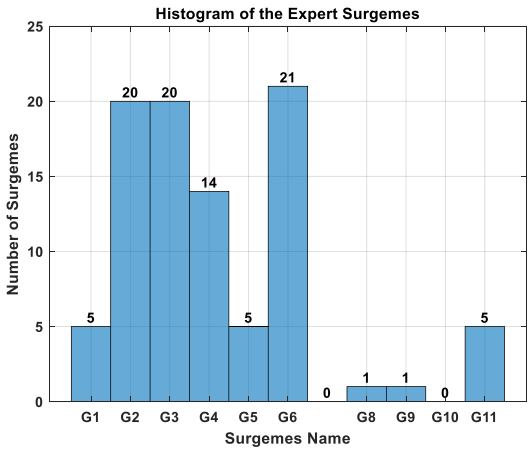
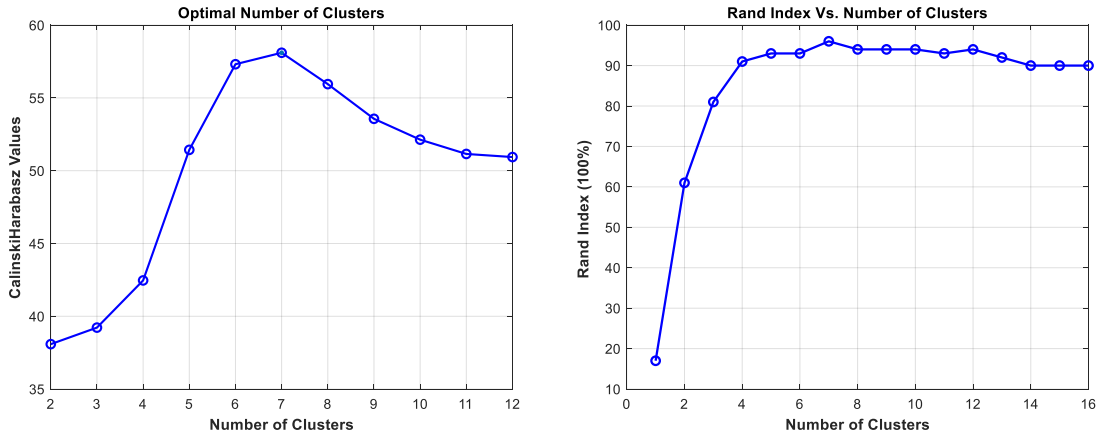


Figure 6.11: Histogram of the expert surgeon surgemes in SU task.

We utilize the Calinski-Harabasz clustering evaluation criterion to find the optimal number of clusters using the mean feature, as illustrated in Figure 6.12. The number of clusters here chosen to be seven will give a higher average Rand-Index of 96%.



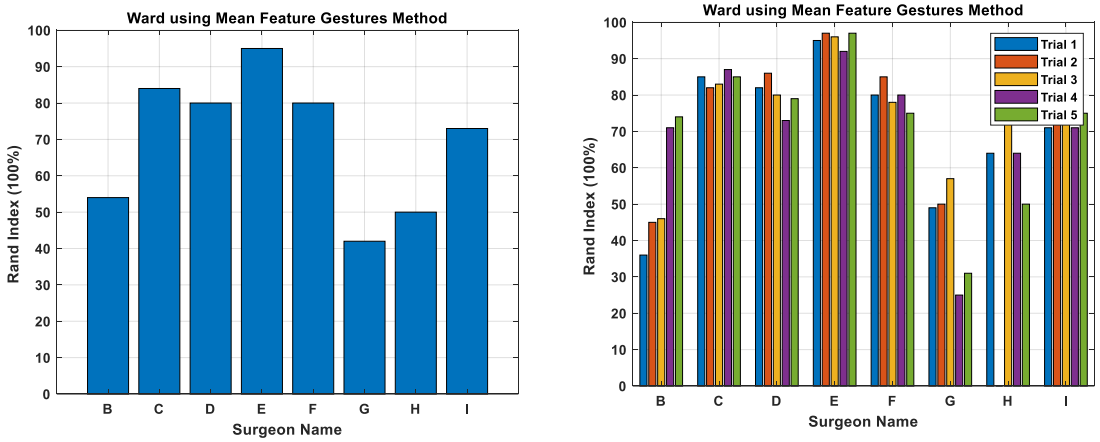
(a) (b)

Figure 6.12: Mean Feature of the gestures (a) Calinski-Harabasz clustering evaluation criterion, (b) Rand-Index plot as a function of the number of clusters.

Figure 6.13 demonstrates the result of using the mean feature technique in implementing the proposed framework through average rand-index in (a) and per trial in (b) of Figure 6.13. We can observe from the results in this figure that using the mean of

the feature is more accurate than the results of the previous experiments. Another point worth mentioning is that the Rand-index results of the expert and intermediate surgeons are distinct from those of the novice surgeons.

This indicates the ability to distinguish their surgemes pattern, which is close to the model of clustering the expert surgeon. Also, the result reveals that enhanced clustering quality is reachable without reducing the time-series features by using the mean feature technique.



(a) (b)

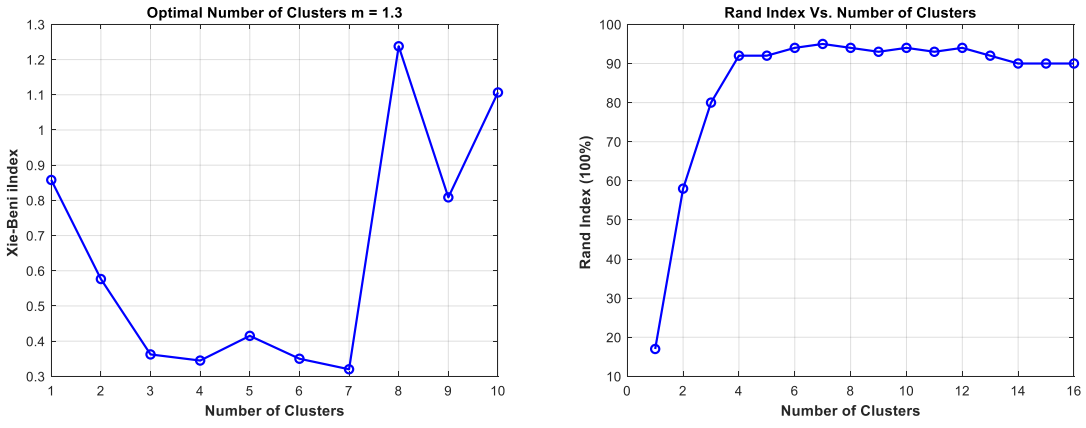
Figure 6.13: Rand-Index results for the Ward clustering using mean features for each surgeon (a) Average Rand Index (b) Rand-Index per trial in suturing task.

6. 3. 4 Fuzzy C-Mean (FCM)

We conducted another experiment to investigate the use of the FCM algorithm to cluster the prototype surgical gestures. As we mentioned previously, FCM is a clustering method wherein each surgeme belongs to multiple groups by a membership grade. In this experiment, we applied the FCM to obtain the prototype surgemes by clustering the surgemes of the expert surgeon (clustering model). We develop our proposed framework on raw kinematic data from the JIGSAWS dataset, which was used to perform suturing

surgical tasks. Here, we employed the mean features representation technique of the surgemes before clustering due to the time computation and low complexity. Therefore, Euclidean distance is used as a distance measure in this case rather than the DTW. We compare the clustering results of the surgeon surgemes with that the manually annotated by a senior expert surgeon.

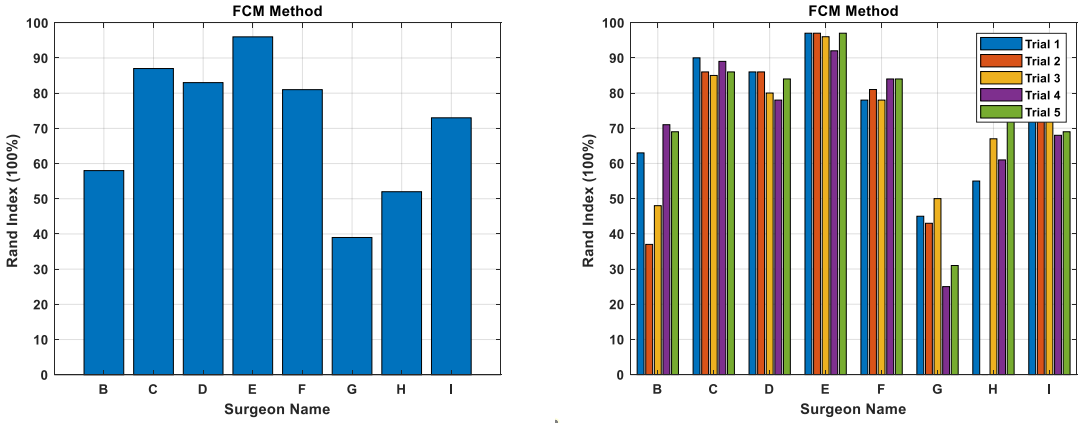
We employ a popular validity index in FCM, the Xie-Beni index criteria [99], to measure the optimal number of clustering as shown on the left of Figure 6.14 and the Rand-Index accuracy to the right of the exact figure. Therefore, the number of clusters chosen is seven that reached both the high Xie-Beni index and Rand-Index. The fuzzifier controlling the partitioning overlap, the small value of m approaches one means more crisp boundaries and less overlap. For the FCM algorithm, we run an experiment for different fuzzy membership m with the Xie-Beni validity index, and we set m to 1.3. For prototype surgemes, we observed that clustering of the expert surgemes achieved %95 of the rand-index accuracy compared with the ground truth.



(a) (b)

Figure 6.14: FCM using Mean Feature of the gestures (a) Xie-Beni validity index, (b) Rand-Index plot as a function of the number of clusters.

Using the clustering method to build prototype surgemes and employing similarity measures to select representative surgeme significantly impacts the Rand index outcomes. Figure 6.15 shows the best results of the surgemes clustering using our proposed method based on the FCM algorithm. Consequently, we can see, the overall accuracy improved of the FCM compared to the experiments-based hierarchical ward clustering algorithm using DTW distance.



(a) (b)

Figure 6.15: Rand-Index results for The FCM clustering using mean features for each surgeon (a) Average Rand Index (b) Rand-Index per trial in suturing task.

In Figure 6.16, we compared the results obtained by the proposed method using all the techniques mentioned in the methodology section through rand-index. We can observe that our approach based on using the mean feature method performs better than the other clustering approaches in most cases. Additionally, the clustering methods based on the mean feature surgemes representation with Euclidean distance outperform the clustering techniques that use DTW as a distance measure. It is also important to mention that the enhanced clustering quality is achievable even with the reduction in the time instance while preserving the dimension of the variable unchanged.

Furthermore, it is crucial to consider the effect of the gesture time to accomplish the surgical task, where short surgeme are challenging to discriminate, resulting in decreased clustering performance. Additionally, the insufficient data for some surgeons in specific trials makes it difficult for any model to differentiate the surgeon surgemes from the prototypes of the expert surgeon.

Figure 6.16: Comparison of the Rand-Index between different surgemes clustering methods for each surgeon in suturing task.

6 . 4 Conclusions and Future Work

Surgical gestures are the key elements in a surgeon's training system, and they can offer a quantitative measurement and feedback to the trainee during the robotic surgical session. We proposed a new unsupervised approach for surgemes clustering by utilizing four techniques based on Hierarchical and FCM algorithms. We evaluated our method on a real dataset by analyzing raw kinematics data from suturing tasks performed by individuals with varying levels of expertise. Also, we demonstrated the benefits of re-

representing the time series data before clustering in terms of computation time reduction and system complexity.

One of the most challenging tasks in unsupervised learning is to deal with outliers. In our case, some surgeons perform surgemes that the expert surgeons generally do not operate. This might adversely affect the quality of surgeme assignment to the appropriate cluster, thereby influencing the clustering algorithm's outcome.

In addition, we used a predetermined number of clusters. Therefore, future research should focus on using a clustering technique to deal with unknown groups derived from expert prototypes. This effort constitutes a step forward toward surgemes segmentation. One of the future work challenges will be to build a clustering method based on using one of the experts' surgemes for model initialization and then map the surgemes from other subjects.

CHAPTER 7 **Summary and Future Research Lines**

This chapter highlights the contributions achieved and the most noticeable results obtained from our current methodologies and summarizes the research directions to be pursued in the future.

7 . 1 Summary

7. 1. 1 Surgical Skill Assessment

Objective evaluation of a surgeon's skill level is a crucial step toward automatic surgical training. The surgical skills of surgeons with different expertise levels have traditionally been evaluated by direct observation of senior expert surgeons inside the surgical room. However, this method suffers from being subjective, costly, and lacking any quantitative indication of the appropriate skills level. Instead, an automatic evaluation would provide an objective and quantitative measure of skill levels. If the surgical activity is captured using a set of sensors, then the problem becomes a task to define an evaluation framework for motion analysis and comparison.

To address this challenge, we propose a new surgical distance, known as PDTW, to learn similarity measures based on dynamic time warping (DTW) and Procrustes analysis that considers the multi-dimensional measure. The DTW method aligns two timeseries with different lengths by contracting/dilating both signals such that their lengthwise becomes equal. The Procrustes analysis, which includes reflection, scaling, and translation, can then be used as a distance measure between two aligned sequences.

We also develop a framework based on a novel surgery skill distance, PDTW, for automated surgical assessment that uses traces representing the surgeon's motion. Furthermore, we employed the k-nearest neighbor as a classifier in the proposed approach and reported the surgical skill assessment classification results. We examined our approach's performance on three surgical datasets using kinematic, Vicon, and wearable sensor data. The obtained results have shown significant assessment improvements of PDTW over the traditional distance measures using raw data. Also, we observed that the experimental results are comparable with relatively high accuracy to the state-of-art methods in automatically classifying surgical skills into expert, intermediate, and novice on different tasks.

7. 1. 2 Surgery Task Classification

Surgeons do various advanced surgical procedures on robotic surgery systems. Recognizing surgical tasks that the surgeon performs is crucial toward automatic surgical training in robotic surgery training. To tackle this problem, we develop a classification framework using kinematic information to recognize among the three different Robot-assisted minimally invasive surgery (RMIS) tasks: suturing, needle passing, and knot tying. This approach is based on using PDTW distance and Fuzzy k-nearest neighbor (FkNN). The FkNN classifier is applied to distinguish between different tasks by assigning a fuzzy class membership based on their distances.

Two validation schemes have been conducted: leave one supertrial out (LOSO) and leave one user out (LOUO). The first set is comparable with state-of-the-art methods based on deep architectures for the LOSO validation technique. In the second set, the results obtained show improvements in the classification of the three surgical tasks. The

performance of our approach surpasses published works for LOUO validation by at least 4.5% accuracy improvement. Furthermore, we consider another scenario by utilizing 3D cartesian motion trajectories of the right and left hands. The results have shown that it reduces time consumption and emphasizes the potential toward using low-cost wrist wearable devices instead of using expensive tools for surgical task recognition and surgeon skill assessment.

7. 1. 3 Surgemes Classification

Regarding supporting trainee surgeons by giving quantifiable feedback about their patterns and improving their skills to be close to the expert surgeons. We developed a framework for surgical gesture classification directly on raw kinematic data. We utilize the k-NN and SVM algorithms to classify the surgemes in our proposed method. Also, we present the mean feature reduction of the surgical gestures that are fast, precise, and decrease the complexity of the proposed framework.

In Addition, we use two distance measures, the DTW with the k-NN method and Euclidean distance with the SVM algorithm and performing the LOSO and LOUO cross-validation schemes. The fusion of the mean feature reduction method along with the SVM technique resulted in the best performance. The results of our proposed approach outperform the existing method based on raw kinematic data from JIGSAWS for both LOSO and LOUO cross-validation setup.

7. 1. 4 Surgemes Clustering

Surgeons need skill and operating knowledge to make proper and safe surgical procedures. However, existing training techniques limit us from conducting in-depth analyses of surgical motions to evaluate these skills accurately.

We develop a method to identify the surgemes by applying unsupervised methods to cluster the surgical activities learned directly from raw kinematic data on the JIGSAWS dataset [11]. We build an unsupervised surgemes model based on predefined surgical gestures. The first step is to find the prototypes by clustering the surgemes of the expert surgeon from all the same expert trials. Then, we map the other surgeons surgemes to the nearest representative of the prototypes and report the clustering accuracy by employing the rand index technique.

We utilize four techniques in our proposed unsupervised approach for surgemes clustering based on Hierarchical and FCM algorithms. Using a real dataset, we tested our methods by assessing raw kinematics data from a suturing task performed by participants with various skill levels. In addition, we highlight the advantages of representing time series data before clustering in terms of computation time saving and system complexity reduction, respectively.

7 . 2 Future Research Directions

Despite the encouraging results exposed through previous chapters in surgical task recognition, surgeon's skill assessment, gestures classification, and clustering, it is still a challenging problem, and more research is required. We think that this dissertation will inspire additional studies in this field, and there are several promising directions to work further in the future to be pursued summarized below:

Transfer learning (TL) algorithms aim to improve the prediction performance in a target task via transferring knowledge from auxiliary data of a related task. The distribution and even the feature space of the data of the tasks can be different. It is interesting to employ the transfer learning based on supervised learning for skill assessment of the surgeon through training a model on the JIGSAWS dataset and performing on our dataset captured by wrist wearable sensors.

Utilizing an unsupervised algorithm on clustering the surgemes could form a basis toward employing the streaming clustering on surgical gestures. One of the expert gestures can initialize the model and stream the remaining surgeons' data. Then map the raw data from other subjects, where surgemes that do not match the existing prototypes will create a new cluster.

The surgical field has few datasets available, and those that do are limited. Thus, a more significant number of data samples with various tasks are required. We need to acquire and annotate new data to construct machine learning models to comprehend and obtain solid surgical activity and skill recognition results.

Final thought inspired by using wearable sensors to detect the stress level of the surgeon while performing unusual during a routine surgery task. This will indicate the expertise skill, where the expert surgeon is expected to act more comfortably than the novice surgeon.

BIBLIOGRAPHY

[1] C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surgical endoscopy*, vol. 25, no. 2, pp. 356-366, 2011.

[2] B. N. Carter, "The fruition of Halsted's concept of surgical training," *Surgery*, vol. 32, no. 3, pp. 518-527, 1952.

[3] A. Peracchia, "Surgical education in the third millennium," *Annals of surgery*, vol. 234, no. 6, p. 709, 2001.

[4] A. Darzi, S. Smith, and N. Taffinder, "Assessing operative skill: needs to become more objective," ed: British Medical Journal Publishing Group, 1999.

[5] M. Levin, T. McKechnie, S. Khalid, T. P. Grantcharov, and M. Goldenberg, "Automated Methods of Technical Skill Assessment in Surgery: A Systematic Review," (in eng), *J Surg Educ*, vol. 76, no. 6, pp. 1629-1639, Nov-Dec 2019, doi: 10.1016/j.jsurg.2019.06.011.

[6] N. Ahmidi *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025-2041, 2017.

[7] R. K. Reznick and H. MacRae, "Teaching surgical skills—changes in the wind," *New England Journal of Medicine*, vol. 355, no. 25, pp. 2664-2669, 2006.

[8] K. E. Roberts, "Evolution of surgical skills training," *World Journal of Gastroenterology*, vol. # 12, no. 20, p. 3219, 2006, doi: 10.3748/wjg.v12.i20.3219.

[9] F. Lalys and P. Jannin, "Surgical process modelling: a review," *International journal of computer assisted radiology and surgery*, vol. 9, no. 3, pp. 495-511, 2014.

[10] "Intuitive Surgical." Intuitive Surgical. https://www.intuitive.com/en-us (accessed.

[11] Y. Gao *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *Miccai workshop: M2cai*, 2014, vol. 3, p. 3.

[12] M. J. Fard, "Computational modeling approaches for task analysis in robotic-assisted surgery," Wayne State University, 2016.

[13] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," *International journal of computer assisted radiology and surgery*, vol. 13, no. 5, pp. 731-739, 2018.

[14] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220-230, 2006.

[15] Z. Wang and A. M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International journal of computer assisted radiology and surgery*, vol. 13, no. 12, pp. 1959-1970, 2018.

[16] J. Martin *et al.*, "Objective structured assessment of technical skill (OSATS) for surgical residents," *British journal of surgery*, vol. 84, no. 2, pp. 273-278, 1997.

[17] V. Datta, S. Mackay, M. Mandalia, and A. Darzi, "The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model 1 1No competing interests declared," *Journal of the American College of Surgeons*, vol. 193, no. 5, pp. 479-485, 2001, doi: 10.1016/s1072-7515(01)01041-9.

[18] T. N. Judkins, D. Oleynikov, and N. Stergiou, "Objective evaluation of expert and novice performance during robotic surgical training tasks," *Surgical endoscopy*, vol. 23, no. 3, pp. 590-597, 2009.

[19] C. Richards, J. Rosen, B. Hannaford, C. Pellegrini, and M. Sinanan, "Skills evaluation in minimally invasive surgery using force/torque signatures," *Surgical endoscopy*, vol. 14, no. 9, pp. 791-798, 2000.

[20] Y. Yamauchi *et al.*, "Surgical skill evaluation by force data for endoscopic sinus surgery training system," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2002: Springer, pp. 44-51.

[21] V. Datta, S. Mackay, M. Mandalia, and A. Darzi, "The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model," *Journal of the American College of Surgeons*, vol. 193, no. 5, pp. 479-485, 2001.

[22] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1-8, 2019.

[23] J. J. H. Leong, M. Nicolaou, L. Atallah, G. P. Mylonas, A. W. Darzi, and G.-Z. Yang, "HMM Assessment of Quality of Movement Trajectory in Laparoscopic Surgery," Springer Berlin Heidelberg, 2006, pp. 752-759.

[24] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *International conference on information processing in computer-assisted interventions*, 2012: Springer, pp. 167-177.

[25] M. J. Fard, S. Ameri, R. B. Chinnam, A. K. Pandya, M. D. Klein, and R. D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," *arXiv preprint arXiv:1611.05136*, 2016.

[26] M. J. Fard, S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1850, 2018.

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

[28] G. Forestier *et al.*, "Surgical motion analysis using discriminative interpretable patterns," *Artificial intelligence in medicine*, vol. 91, pp. 3-11, 2018.

[29] Z. Wang and 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)*, 2018: IEEE, pp. 1793-1796.

[30] Y. Sharma *et al.*, "Video based assessment of OSATS using sequential motion textures," 2014: Georgia Institute of Technology.

[31] Y. Sharma *et al.*, "Automated surgical OSATS prediction from videos," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 2014: IEEE, pp. 461-464.

[32] N. Ahmidi, M. Ishii, G. Fichtinger, G. L. Gallia, and 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*, 2012, vol. 2, no. 6: Wiley Online Library, pp. 507-515.

[33] A. L. Trejos, R. V. Patel, M. D. Naish, and C. M. Schlachta, "Design of a sensorized instrument for skills assessment and training in minimally invasive surgery," in *2008 2nd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2008: IEEE, pp. 965-970.

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

[35] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *International journal of computer assisted radiology and surgery*, vol. 13, no. 3, pp. 443-455, 2018.

[36] L. Bouarfa, P. P. Jonker, and J. Dankelman, "Discovery of high-level tasks in the operating room," *Journal of Biomedical Informatics*, vol. 44, no. 3, pp. 455-462, 2011, doi: 10.1016/j.jbi.2010.01.004.

[37] M. J. Fard, A. K. Pandya, R. B. Chinnam, M. D. Klein, and R. D. Ellis, "Distance-based time series classification approach for task recognition with application in surgical robot autonomy," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 13, no. 3, p. e1766, 2017.

[38] M. J. Bani and S. Jamali, "A new classification approach for robotic surgical tasks recognition," *arXiv preprint arXiv:1707.09849*, 2017.

[39] D. Sarikaya, K. A. Guru, and J. J. Corso, "Joint surgical gesture and task classification with multi-task and multimodal learning," *arXiv preprint arXiv:1805.00721*, 2018.

[40] S. Sefati, N. J. Cowan, and R. Vidal, "Learning shared, discriminative dictionaries for surgical gesture segmentation and classification," in *MICCAI Workshop: M2CAI*, 2015, vol. 4.

[41] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali, and R. Vidal, "End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018: IEEE, pp. 1558-1567.

[42] B. Béjar Haro, L. Zappella, and R. Vidal, "Surgical Gesture Classification from Video Data," Springer Berlin Heidelberg, 2012, pp. 34-41.

[43] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 966-976, 2011.

[44] L. Zappella, B. Béjar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Medical Image Analysis*, vol. 17, no. 7, pp. 732-745, 2013, doi: 10.1016/j.media.2013.04.007.

[45] R. DiPietro *et al.*, "Recognizing surgical activities with recurrent neural networks," in *International conference on medical image computing and computer-assisted intervention*, 2016: Springer, pp. 551-558.

[46] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *European Conference on Computer Vision*, 2016: Springer, pp. 36-52.

[47] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*, 2016: Springer, pp. 47-54.

[48] L. Ding and C. Xu, "Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation," *arXiv preprint arXiv:1705.07818*, 2017.

[49] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," in *International conference on medical image computing and computer-assisted intervention*, 2018: Springer, pp. 247-255.

[50] J. Reason, *Human error*. Cambridge university press, 1990.

[51] F. Despinoy *et al.*, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280-1291, 2015.

[52] Y. Gao, S. S. Vedula, G. I. Lee, M. R. Lee, S. Khudanpur, and G. D. Hager, "Unsupervised surgical data alignment with application to automatic activity annotation," 2016: IEEE, doi: 10.1109/icra.2016.7487608. [Online]. Available: https://dx.doi.org/10.1109/icra.2016.7487608

[53] R. Dipietro and G. D. Hager, "Unsupervised Learning for Surgical Motion by Learning to Predict the Future," Springer International Publishing, 2018, pp. 281-288.

[54] M. J. Fard, S. Ameri, R. B. Chinnam, and R. D. Ellis, "Soft Boundary Approach for Unsupervised Gesture Segmentation in Robotic-Assisted Surgery," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 171-178, 2017, doi: 10.1109/lra.2016.2585303.

[55] A. Murali *et al.*, "TSC-DL: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with Deep Learning," 2016: IEEE, doi: 10.1109/icra.2016.7487607. [Online]. Available: https://dx.doi.org/10.1109/icra.2016.7487607

[56] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan, "Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model," *Computer Aided Surgery*, vol. 7, no. 1, pp. 49-61, 2002.

[57] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Evaluating surgical skills from kinematic data using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018: Springer, pp. 214-221.

[58] J. Lin, S. Williamson, K. Borne, and D. DeBarr, "Pattern recognition in time series," *Advances in Machine Learning and Data Mining for Astronomy*, vol. 1, no. 617-645, p. 3, 2012.

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

[60] J. T. Kent, "New directions in shape analysis," *The art of statistical science*, vol. 115, 1992.

[61] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009.

[62] I. L. Dryden and K. V. Mardia, *Statistical shape analysis: with applications in R*. John Wiley & Sons, 2016.

[63] J. M. Keller, D. Liu, and D. B. Fogel, *Fundamentals of computational intelligence: neural networks, fuzzy systems, and evolutionary computation*. John Wiley & Sons, 2016.

[64] M. Popescu, C. J. Cooper, and S. Barnes, "Automated Operative Skill Assessment Using IR Video Motion Analysis," in *AMIA*, 2014.

[65] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861-874, 2006.

[66] A. B. Haynes *et al.*, "A surgical safety checklist to reduce morbidity and mortality in a global population," *New England Journal of Medicine*, vol. 360, no. 5, pp. 491-499, 2009.

[67] G. Forestier, F. Petitjean, L. Riffaud, and P. Jannin, "Optimal sub-sequence matching for the automatic prediction of surgical tasks," in *Conference on Artificial Intelligence in Medicine in Europe*, 2015: Springer, pp. 123-132.

[68] A. Cuschieri, "Whither minimal access surgery: tribulations and expectations," *The American Journal of Surgery,* vol. 169, no. 1, pp. 9-19, 1995.

[69] S. W. Smith, "The scientist and engineer's guide to digital signal processing," 1997.

[70] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proceedings of the 2011 SIAM international conference on data mining*, 2011: SIAM, pp. 699-710.

[71] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition,* vol. 44, no. 3, pp. 678-693, 2011.

[72] H. Sakoe, "Dynamic-programming approach to continuous speech recognition," in *1971 Proc. the International Congress of Acoustics, Budapest*, 1971.

[73] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing,* vol. 26, no. 1, pp. 43-49, 1978.

[74] S. Z. Li, *Encyclopedia of Biometrics: I-Z*. Springer Science & Business Media, 2009.

[75] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait recognition based on procrustes shape analysis," in *Proceedings. International Conference on Image Processing*, 2002, vol. 3: IEEE, pp. III-III.

[76] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580-585, 1985.

[77] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management,* vol. 45, no. 4, pp. 427-437, 2009.

[78] M. B. Stegmann and D. D. Gomez, "A brief introduction to statistical shape analysis," *Informatics and mathematical modelling, Technical University of Denmark, DTU,* vol. 15, no. 11, 2002.

[79] B. J. Dlouhy and R. C. Rao, "Surgical skill and complication rates after bariatric surgery," *The New England journal of medicine,* vol. 370, no. 3, pp. 285-285, 2014.

[80] M. G. Katos and D. Goldenberg, "Emergency cricothyrotomy," *Operative Techniques in Otolaryngology-Head and Neck Surgery,* vol. 18, no. 2, pp. 110-114, 2007.

[81] A. MacIntyre, M. K. Markarian, D. Carrison, J. Coates, D. Kuhls, and J. J. Fildes, "Three-step emergency cricothyroidotomy," *Military medicine,* vol. 172, no. 12, pp. 1228-1230, 2007.

[82] M. INC. MBIENTLAB INC. https://mbientlab.com/metamotionr/ (accessed.

[83] A. Zia, L. Guo, L. Zhou, I. Essa, and A. Jarc, "Novel evaluation of surgical activity recognition models using task-based efficiency metrics," *International journal of computer assisted radiology and surgery*, vol. 14, no. 12, pp. 2155-2163, 2019.

[84] V. N. Vapnik and C. AY, "On a class of pattern-recognition learning algorithms," *Automation and Remote Control*, vol. 25, no. 6, pp. 838-&, 1965.

[85] V. Vapnik, "The nature of statistical learning theory springer New York google scholar," *New York*, 1995.

[86] H. Byun and S.-W. Lee, "Applications of support vector machines for pattern recognition: A survey," in *International Workshop on Support Vector Machines*, 2002: Springer, pp. 213-236.

[87] V. Agarwal, "Ridge regression approach to color constancy," 2005.

[88] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[89] B. Salehi, A. H. Ghanbaran, and M. Maerefat, "Intelligent models to predict the indoor thermal sensation and thermal demand in steady state based on occupants' skin temperature," *Building and Environment*, vol. 169, p. 106579, 2020.

[90] S. Albasri, M. Popescu, and J. Keller, "A Novel Distance for Automated Surgical Skill Evaluation," in *2019 E-Health and Bioengineering Conference (EHB)*, 2019: IEEE, pp. 1-6.

[91] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561-580, 2007.

[92] S. Albasri, M. Popescu, and J. Keller, "Surgery Task Classification Using Procrustes Analysis," in *2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2019: IEEE, pp. 1-6.

[93] I. Surgical. https://www.intuitive.com/en-us (accessed.

[94] S. Hirano and S. Tsumoto, "Empirical comparison of clustering methods for long time-series databases," in *Active Mining*: Springer, 2005, pp. 268-286.

[95] T. Oates, M. D. Schmill, and P. R. Cohen, "A method for clustering the experiences of a mobile robot that accords with human judgments," in *AAAI/IAAI*, 2000, pp. 846-851.

[96] S. Theodoridis and K. Koutroumbas, "Pattern recognition, edition," ed: Academic Press, fourth edition Edition, 2009.

[97] M. Chiş, S. Banerjee, and A. E. Hassanien, "Clustering time series data: an evolutionary approach," in *Foundations of Computational, IntelligenceVolume 6*: Springer, 2009, pp. 193-207.

[98] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846-850, 1971.

[99] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650-1654, 2002.

[100] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1-27, 1974.

[101] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

[102] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

VITA

Safaa Albasri was born in Baghdad, the capital of Iraq in 1977. He graduated from the Science school with B.Sc. degree in Mathematics in 1999 and studied at the University of Baghdad. He earned B.Sc. degree in Electrical Engineering and M.Ss. in Communication and Electronic Engineering from the Al-Mustansiriyah University in 2006 and 2009, respectively. In 2017, he received his M.E. in Electrical Engineering at the University of Missouri-Columbia. He is currently pursuing toward the Ph.D. degree in the Electrical Engineering and Computer Science Department at the University of Missouri. His current research interests focus on machine learning and computational intelligence applications in medical field and Surgical skill assessment using wearable sensors and depth image. In 2006, he joined Al-Mustansiriyah University in Baghdad as one of the faculty members in the Department of Electrical Engineering. Albasri joined to the University of Missouri-Columbia in August 2013 for a Ph.D. program in Electrical Engineering and Computer Science. He was a sponsored student by the Higher Committee of Education Development in Iraq (HCED).