This is a post-print of an article published in Artificial Intelligence in Medicine. The final authenticated version is available online at: https://doi.org/10.1016/j.artmed.2018.08.002.

Surgical motion analysis using discriminative interpretable patterns

Germain Forestier[a,b,*], François Petitjean[b], Pavel Senin[d], Fabien Despinoy[c], Arnaud Huaulmé[c], Hassan Ismail Fawaz[a], Jonathan Weber[a], Lhassane Idoumghar[a], Pierre-Alain Muller[a], Pierre Jannin[c]

[a] *IRIMAS, Université de Haute-Alsace, Mulhouse, France*
[b] *Faculty of Information Technology, Monash University, Melbourne, Australia*
[c] *Univ Rennes, Inserm, LTSI - UMR_S 1099, F35000, Rennes, France*
[d] *Los Alamos National Laboratory, University Of Hawai'i at Mānoa*

Abstract

Objective. The analysis of surgical motion has received a growing interest with the development of devices allowing their automatic capture. In this context, the use of advanced surgical training systems makes an automated assessment of surgical trainee possible. Automatic and quantitative evaluation of surgical skills is a very important step in improving surgical patient care.
Material and Method. In this paper, we present an approach for the discovery and ranking of discriminative and interpretable patterns of surgical practice from recordings of surgical motions. A pattern is defined as a series of actions or events in the kinematic data that together are distinctive of a specific gesture or skill level. Our approach is based on the decomposition of continuous kinematic data into a set of overlapping gestures represented by strings (bag of words) for which we compute comparative numerical statistic (tf-idf) enabling the discriminative gesture discovery via its relative occurrence frequency.
Results. We carried out experiments on three surgical motion datasets. The results show that the patterns identified by the proposed method can be used to accurately classify individual gestures, skill levels and surgical interfaces. We also present how the patterns provide a detailed feedback on the trainee skill assessment.
Conclusions. The proposed approach is an interesting addition to existing learning tools for surgery as it provides a way to obtain a feedback on which parts of an exercise have been used to classify the attempt as correct or incorrect.

Keywords: Temporal Analysis, Dynamic Time Warping, Surgical Process Modelling, Surgery

1. Introduction

In recent years, analysis of surgical motion has received a growing interest following the development of devices enabling automated capture of surgeon motions such as tracking, robotic and training systems. Surgical training programs now often include surgical simulators which are equipped with sensors for automatic surgical motions recording [1, 2, 3]. The ability to collect surgical motion data brings unprecedented opportunities for automated objective analysis and assessment of surgical trainees progression. The main goal of this effort is to support surgeons in technical skills acquisition, as these are shown to correlate with a reduction of patient complications [4]. Hence, automated evaluation of surgical skill level is an important step in surgical patient care improvement and is related to the more general initiative of *surgical data science* [5].

This article tackles the issue of identifying discriminative and interpretable patterns of surgical practice from recordings of surgical motions. We define a *pattern* as a series of actions or events in the kinematic data that together are distinctive of a specific gesture or a skill level. We show, that by using these patterns, we can reach beyond the simple classification of observed surgeons into categories (*e.g.*, Expert, Novice) by providing a quantitative evidence-supported feedback to the trainee as per where he or she can improve. The proposed approach, based on SAX-VSM algorithm [6], considers surgical motion as continuous multi-dimensional time-series and starts by discretizing them into sequence of letters (*i.e.*, strings) using Symbolic Aggregate approXimation (SAX) [7]. In turn, SAX sequences are decomposed into subsequences of few consecutive letters via sliding window. The relative frequencies of these subsequences, *i.e.*, the number of times they appear in a given sequence or in a set of sequences,

*Corresponding author – IRIMAS - Université de Haute Alsace, 12 rue des freres Lumiere, 68093 Mulhouse, France, Tel.:+33 3 8933 6963, Fax.:+33 3 8942 3282
Email addresses: germain.forestier@uha.fr (Germain Forestier), francois.petitjean@monash.edu (François Petitjean), senin@hawaii.edu (Pavel Senin), fabien.despinoy@chu-rennes.fr (Fabien Despinoy), arnaud.huaulme@univ-rennes1.fr (Arnaud Huaulmé), hassan.ismail-fawaz@uha.fr (Hassan Ismail Fawaz), jonathan.weber@uha.fr (Jonathan Weber), lhassane.idoumghar@uha.fr (Lhassane Idoumghar), pierre-alain.muller@uha.fr (Pierre-Alain Muller), pierre.jannin@univ-rennes1.fr (Pierre Jannin)

are then used to identify discriminative patterns that characterize specific surgical motion. To discover the patterns, we rely on the Vector Space Model (VSM) [8] which has been originally proposed as an algebraic model for representing collection of text documents. The identified discriminative patterns are then used to perform classification by identifying them in to-be-classified recordings. Furthermore, by highlighting discriminative patterns in the visualization of original motion data, we are able to provide an intuitive visual explanation about *why* a specific skill assessment is provided. We evaluated our method on the kinematic data from the JHU-ISI Gesture and Skill Assessment Dataset (JIGSAWS) [9] (the largest publicly accessible database for surgical gesture analysis) and two other surgical motions datasets. The main contributions of this paper are:

- A framework for identifying discriminative and interpretable patterns in surgical activity motion based on SAX [7] and VSM [6].
- Experimental evaluation highlighting the relevance of the proposed method for gestures classification, skill assessment and surgical interface comparison.
- A visualization technique enabling self-assessment of trainee skills.

2. Background

2.1. Related work

Previous surgical skills assessment methods focused on evaluating the trainees by a senior surgeon who used a dedicated check-list [10, 11]. These methods depend on the senior surgeon's work hours with too many subjective variables: the checklists development process, inter-rater reliability and the rater bias [12]. Another method consists on evaluating the patients' outcome after several surgeries [13]. This type of methods suffers from two impediments: it needs a huge amount of patients' outcome data - which is very difficult to acquire for trainees, and depends largely on the patient's condition before, during and after the surgery. Considering these disadvantages, several researchers considered evaluating surgical skills using surgical motion analysis which is mainly based on kinematic data recorded by surgical robots [14, 15] and video data [16, 17, 18]. By assuming that the features of a dynamic scene (video data) are the output of a Linear Dynamical System (a set of linear equation with latent variables [19]), new video clips were classified in order to show that skills and gestures classification, based on video data, can achieve the state of the art performance of methods based on kinematic data [16, 20]. Sequential feature selection was used in [17] to reduce the number of features extracted from the videos. The resulting feature vector was fed to a nearest neighbor classifier with cosine distance metric. Kinematic data usually include multiple attributes such as the position of robot's tools, rotations, and velocities. From such data, significant amount of work has been devoted to the segmentation of surgical tasks into more detailed gestures [21, 22, 23] and to the study of teleoperation and its effect on the spatiotemporal characteristics of the surgeon's movements [24, 25]. Segmenting surgical motion into gestures makes it possible to obtain a finer description of surgical tasks leading to a more detailed feedback on skill assessment [26, 27, 28]. The need of new skill assessment techniques with a detailed feedback during the surgical residency programs has been emphasized in [29]. Previous work concerned with gesture segmentation and surgical skill assessment using kinematic and video data use Hidden Markov Models [14, 15, 30, 31, 32], Conditional Random Fields [33] and Linear Dynamical Systems [20]. In [14], a surgeon's skill level was identified by discretizing the velocity data into discrete symbols which are used to train the Hidden Markov Models. [15] used the kinematic data from a surgical robot to generate smooth trajectories in order to capture the underlying structure of experts' trajectories. [30] compared different statistical modeling techniques (Latent Dirichlet Allocation, Hidden Markov Models and Gaussian Mixture Models) for surgical skills evaluation. In [31], Sparse Hidden Markov Models were introduced for the classification of gestures and skills in surgical tasks. Markov Models of force and torque signals (applied by the surgeons on on their instruments) were used in [32] to characterize surgical skills. [33] proposed a method that combines Markov and semi-Markov Conditional Random Fields for surgical gestures segmentation and classification. They also showed how their proposed approach [33] allows the use of both kinematic and video data. However, studies in [34, 18, 35] showed that feedback on medical practice allows surgeons to improve their performance and achieve even higher skill levels. Hence the main drawback of the previous approaches is the difficulty for the trainees to understand the output and to use it as a feedback to improve their performance. To tackle this problem, a web-based surgical skill training and evaluating tool was proposed in [18]. This method is based on a computer vision algorithm that analyzes the video of a surgeon's hand and surgical tool movements in order to extract features that could be passed to a multivariate linear regression model to find the relationship between the extracted features and the surgeon's score. Although this technique provides live skill evaluation and feedback for the trainees, the main drawback is that it uses for each task a pre-defined area of error that the surgeon should not enter. In [36], the approach is perhaps the closest to our work, where they viewed the surgical task as a sequence of sub-tasks. For each surgical task, they designed an optimal trajectory that was compared to the surgeons' trajectories using Dynamic Time Warping. The problem is similar to [18] with the fact that for complex trajectories (e.g. suturing) finding one optimal trajectory is not simple and maybe not fair. In contrast, our approach seeks not only to identify that a surgical motion has been performed

by a novice surgeon, but also to explain *why* it has been classified as such without the need to predefine any *perfect* trajectory that should be followed by the trainees. This step is critical in justifying the reasons why the trainee is still considered as a novice and to help him or her focus on the specific steps that require improvement.

2.2. Previous work on JIGSAWS

In this subsection, we review related work related to the largest publicly available dataset for surgical gesture analysis, JIGSAWS [9]. Recent work on this benchmark dataset has focused mainly on four tasks: (1) surgical skill evaluation, (2) surgical gesture classification, (3) surgical gesture segmentation and (4) surgical task recognition. In this paper we tackle the first and second problems, but for completeness, the other complementary approaches validated on the JIGSAWS dataset are detailed in this subsection.

Several existing work have focused on analyzing the movements' spatiotemporal characteristics by extracting features from the kinematic data using the two-thirds power law [37] and the one-sixth power law [24]. Another approach based on reinforcement learning [38] showed how intermediate surgeons had lower trajectory errors than novice surgeons. Deep learning frameworks have been proposed in [39, 40, 41] to estimate the position and velocities of the surgical robot's tool-tips using the recorded video data.

As for recognizing surgical tasks,the goal is to identify the surgical task (Suturing, Needle passing, Knot tying) by analyzing the kinematic data. Two approaches [42, 43] applied a k-nearest neighbor classifier with Dynamic Time Warping to identify the three surgical tasks.

A lot of approaches have been proposed for the automatic segmentation of surgical tasks into more detailed gestures. Variants of Hidden Markov Models (HMM) and Conditional Random Fields have been tested in [44]. In [45], using the kinematic data, a soft unsupervised gesture segmentation framework has been proposed to automatically segment surgical gestures based on their gradual transitions. Two temporal subspace clustering methods were proposed in [46] for unsupervised action segmentation based on the kinematic data. A recent approach [47] introduced an end-to-end algorithm for jointly learning the weights of a Conditional Random Field model to classify gestures based on the kinematic captured data. In [48] they designed a deep learning architecture to segment surgical gestures using the surgical tasks' recorded video data. Another deep learning framework was proposed in [49, 50] that uses a an encoder-decoder convolutional neural network on video data to segment the surgical tasks into fine-grained gestures.

As for the gesture classification task, the gestures' boundaries are supposed to be known and the goal is to determine which class the gestures belong to. In the benchmark conducted in [44], the authors tested the Bag of Spatio-Temporal Features and the Linear Dynamical System (LDS) to classify surgical gestures using the already segmented video data. LDS was also tested along a variant of the HMM (Gaussian Mixture Models), to classify surgical gestures using only the kinematic data. A k-nearest neighbor classifier with Dynamic Time Warping applied over the kinematic data was also used in [42] for surgical gesture classification. Another approach for surgical gesture classification was proposed in [23]. They used an auto-encoder over the kinematic followed by a variant of Dynamic Time Warping to align the extracted features.

For the surgical skill evaluation task, we distinguish between two types of approaches based on their output. The first type of approaches [51, 52, 53] aims to predict the modified Objective Structured Assessment of Technical Skills (OSATS) [54] scores. The second type of surgical skill evaluation is to predict the self-proclaimed skill level (Novice, Intermediate, Expert) of the subjects performing the surgical tasks. In [31] Sparse Hidden Markov Models (SHMM) were proposed to classify surgical skills using the kinematic data and its corresponding gesture boundaries and labels. Another recent approach uses Approximate Entropy (ApEn) [53] to extract global features from the kinematic data and feed it to a nearest neighbor classifier. As already mentioned in the introduction, the main drawback of these approaches [53, 31] is that no interpretable feedback is provided for the trainees. This type of feedback could help them improve and achieve even higher surgical skill levels.

3. Method

3.1. Symbolic Aggregate approXimation (SAX)

We propose to use Symbolic Aggregate approXimation (SAX) [7] to discretize the input time series [55]. For time series T of length n, SAX obtains a lower-dimensional representation by first performing a z-normalization then dividing the time series into s equal-sized segments. Next, for each segment, SAX computes a mean value and maps it to a symbol according to a pre-defined set of breakpoints dividing the data space into α equiprobable regions, where α is the user specified alphabet size. While dimensionality reduction is a desirable feature for exploring global patterns, the high compression ratio (n/s) significantly affects performance in cases where localized phenomena are of interest. Thus, for the local pattern discovery, SAX is typically applied to a set of subsequences that represent local features – a technique called subsequence discretization [56] which is implemented via a sliding window. Note that other time-series discretization approaches could have been used at this step [57, 58, 59]. The choice of SAX is motivated by its unsupervised nature which relies on predefined cutoffs. These cutoffs (given an alphabet size) allows us to have the exact same discretization configuration for each step of the experiments. Thus, the identified patterns are comparable throughout the experiments and can be compared across population of trainees. An example of SAX representation of time series is provided in Figure 1. In this example, raw kinematic data corresponding

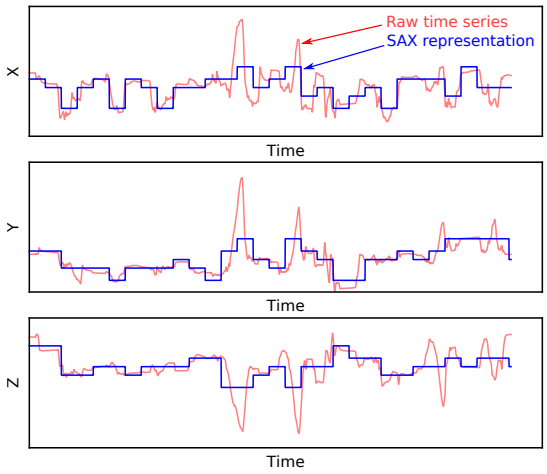


Figure 1: Example of converting three time series of raw kinematic data corresponding to X, Y and Z coordinates of the right hand of the surgeon into a five characters alphabet. The raw time series are in red and the SAX levels are in blue.

to X, Y and Z coordinates of the surgeon's right hand are converted into a five characters alphabet [60].

3.2. Bag of words representation of kinematic data

Following the approach proposed in [6], a sliding window technique is used to convert a time series T of length n into the set of m SAX words, where $m = (n - l_s) + 1$ and l_s the sliding window length. A sliding window of length l_s is applied across the time series T and the overlapping extracted subsequences are converted into SAX words and then put in a collection. This collection is a *bag of words* representation of the original time series T.

In the case of kinematic data, this process is performed independently for each dimension of the data (*e.g.*, x coordinate, y coordinate, etc.). All features are normalized on a per-trial per-feature basis. Each word extracted in each dimension of the data is postfixed with the name of the dimension (*e.g.* x, y, etc.). We assume that depending on the gesture or the skill level to classify, different kinematic features can be relevant. Note, that this methodology can be used regardless of the available kinematic data (*e.g.* number of features, etc.). Figure 2 illustrates the conversion of kinematic data for one trial into a bag of words using SAX.

3.3. Vector Space Model (VSM)

We rely on the original definition of vector space model as it is known in Information Retrieval (IR) [8, 6]. The *tf∗idf* weight for a term t is defined as a product of two factors: term frequency (*tf*) and inverse document frequency (*idf*). The first factor corresponds to logarithmically scaled term frequency [61].

$$\mathrm{tf}_{t,d} = \begin{cases} \log(1 + \mathrm{f}_{t,d}), & \text{if } \mathrm{f}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where t is the term, d is a bag of words (a document in IR terms), and $\mathrm{f}_{t,d}$ is the frequency of t in d. The inverse document frequency [61] is defined as

$$\mathrm{idf}_{t,D} = \log \frac{|D|}{|d \in D : t \in d|} = \log \frac{N}{\mathrm{df}_t} \quad (2)$$

where N is the cardinality of a corpus D (the total number of classes) and the denominator df_t is the number of bags where the term t appears. Then, *tf∗idf* weight value for a term t in the bag d of a corpus D is defined as

$$\mathrm{tf{*}idf}(t, d, D) = \mathrm{tf}_{t,d} \times \mathrm{idf}_{t,D} = \log(1 + \mathrm{f}_{t,d}) \times \log \frac{N}{\mathrm{df}_t} \quad (3)$$

for all cases where $\mathrm{f}_{t,d} > 0$ and $\mathrm{df}_t > 0$, or zero otherwise.

Once all frequencies are computed, the term frequency matrix becomes the term weight matrix, whose columns are used as *class term weight* vectors to perform classification using Cosine similarity. For two vectors $\mathbf{a}$ and $\mathbf{b}$, the Cosine similarity is based on their inner product and defined as

$$\mathrm{similarity}(\mathbf{a}, \mathbf{b}) = cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{||a|| \cdot ||b||} \quad (4)$$

To classify an unlabeled document, its word frequencies vector $freq_{unlabeled}$ is compared to the *tf∗idf* weight vectors of each of the i classes:

$$\text{class label} = \arg\max_i \{tf{*}idf_i * freq_{unlabeled}\} \quad (5)$$

3.4. Training and classifying kinematic data

The training step starts by transforming the kinematic data into SAX representation using two parameters: the size of the sliding window l_s, and the size of the alphabet α. Then, the algorithm builds a corpus of N bags corresponding to the subsequences extracted from the N classes of kinematic data, i.e. same skill level or same gesture depending on the application. The *tf∗idf* weighting is then applied to create N real-valued weight vectors of equal length, representing the different class of kinematic data.

In order to classify an unlabeled kinematic data, the method transforms it into a terms frequency vector using exactly the same sliding window and SAX parameters used for the training part. It computes the cosine similarity measure (Eq. 4) between this term frequency vector and the N *tf∗idf* weight vectors representing the training classes. The unlabeled kinematic data is assigned to the class whose vector yields the maximal cosine similarity value.

4. Experiments

The proposed method has been evaluated on three surgical motion datasets. The first one is the JIGSAWS dataset [9] which is the largest publicly accessible database

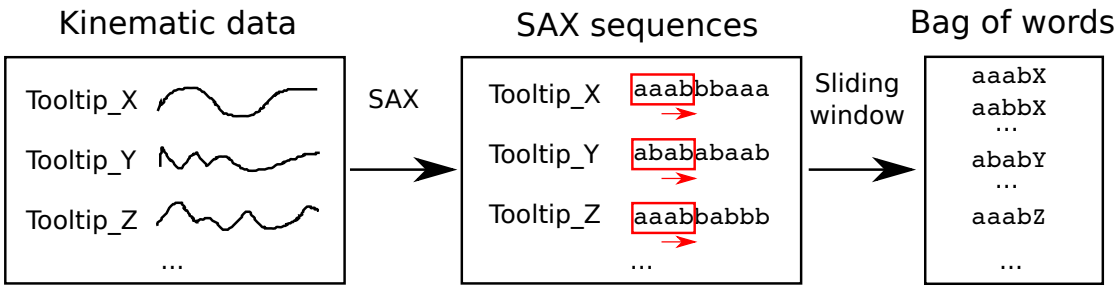


Figure 2: Conversion of kinematic data for one trial into a bag of words using SAX [7] and a sliding window of size 4 (in red).

for surgical gesture analysis. This dataset was used for gesture and skills classification. The second one is a dataset used in a recent study [22] to compare two surgical motion interfaces (Sigma.7 and Leap Motion) to control a RAVEN-II robot [62]. This dataset was used for motion interface classification and skill classification. Finally, the third one is a dataset of micro-surgical suturing tasks captured using a dedicated robot [63]. This dataset was used for skill classification.

4.1. Datasets

4.1.1. JIGSAWS dataset

This dataset includes 8 subjects with 3 different skill levels (Novice, Intermediate and Expert) performing 3–5 trials of three tasks (suturing, knot tying, and needle passing) [9]. Figure 3 illustrates the three tasks. Each trial lasts about 2 minutes and is represented by the kinematic data of both master and slave manipulators of the da Vinci robotic surgical system recorded at a constant rate of 30 Hz. Kinematic data consists of 76 motion variables including positions and velocities of both master and slave manipulators. All trials in the JIGSAWS dataset were manually segmented into 15 surgical gestures. Video of the trials are also available and synchronized with the kinematic data. A detailed description of the dataset is available in [44].

4.1.2. RAVEN-II dataset

This dataset of trajectories has been acquired on a teleoperation platform composed of a RAVEN-II robot [62] and two separated human-machine interfaces which are the Sigma.7 from *Force Dimension* and the Leap Motion device. The objective of this dataset was to compare surgical motion performance from operators using both interfaces in order to validate the use of contactless interface as a relevant control input for teleoperation purpose [22].

For each interface, the same training task was executed 10 times by 3 different participants with different skill levels: an urologist expert who regularly performs surgeries with the da Vinci system (named 'C'), a last year resident who only used few times a robotic training system (named 'B') and a teleoperation system engineer (named 'A'). All participants were right-handed. At the end, we kept the last five trials as relevant in order to remove the learning phase in the different gesture executions. It allows us to obtain 30 trials with an average duration of about 68 seconds for the Leap Motion and about 37 seconds for the Sigma.7. This surgical dataset contains 27 kinematic data acquired at 100Hz into the robot reference, corresponding to the surgical tools information (transformation matrix of each tooltip and graspers angle) as well as console status. Regarding the exercise, the surgical training task was directly inspired by the FLS guidelines [64]. This task involved peg transfers to several target locations using bimanual manipulation (see [22] for a precise description of the task). Figure 5 shows the contactless control interface (Leap Motion) and the RAVEN-II robot.

4.1.3. Microsurgery dataset

This dataset has been collected at the University of Tokyo Hospital. It consists of micro-surgical suturing tasks of $0.7mm$ artificial blood vessel performed using a master-slave robotic platform [63]. The dataset includes 6 participants with different surgical expertise and different robotics skills. Three of them, called experts, are surgical experts and robotics novices, the three others, called engineering students, are surgical novices and robotics experts. Each participant performed 3-6 trials for a total of 11 trials for experts and 16 for engineering students. The mean duration of suturing task is about 3 minutes. Figure 6 shows snapshots of this task. For each trial, the video and the kinematic data have been recorded at 30Hz. Kinematic data consist of 16 motion variables including, for booth surgical instrument, the positions (x, y, and z), the rotations, the tool grip and the output voltage of the grip. The videos have been manually annotated using the Surgery Workflow Toolbox [65].

4.2. Parameters and evaluation

4.2.1. Parameters

The training step of our method first transforms the kinematic data time series into SAX representation configured by two parameters: the sliding window length (l_s) and SAX alphabet size (α). The number of segments per window was kept equal to the length of the window which means that every point of the time series was transformed

Figure 3: Snapshots of the three surgical tasks in the JIGSAWS dataset (from left to right): suturing, knot-tying, needle-passing [9].

into a letter. This choice was made to allow us to map back the patterns on the original time series. Parameters l_s and α were optimized using cross-validation on the training data. As they can differ for each specific classification problem, their values are provided along with the experimental results.

4.2.2. Gesture classification evaluation

Gesture classification has been performed only on the JIGSAWS dataset because it is the only one with annotated gesture boundaries. We considered the gesture boundaries to be known and we used the kinematic data alone. We present results for two cross-validation configurations provided with the JIGSAWS data [9]. In the first configuration – leave one supertrial out (LOSO) – for each iteration of cross-validation (five in total), one trial of each subject was left out for the test and the remaining trials were used for training. In the second configuration – leave one user out (LOUO) – for each iteration of the cross-validation (eight in total), all the trials belonging to a particular subject were left out for the test. These are the standard benchmark configurations provided in [9]. We report micro (average of total correct predictions across all classes) and macro (average of true positive rates for each class) performance results as defined in [44]. For each of the F cross-validation folds, a confusion matrix C_f of size $n \times n$ is computed as : $C_f[i,j]$ = number of *class* i samples predicted as *class* j. The complete confusion matrix, C, is the sum of all of the confusion matrices:

$$C = C_1 + C_2 + \ldots + C_F \tag{6}$$

Given the complete confusion matrix, the Micro average is computed as the average of total correct predictions across all classes:

$$Micro = \frac{\sum_{i=1}^{n} C[i,i]}{\sum_{i,j=1}^{n} C[i,j]} \tag{7}$$

and the Macro average is the mean of true positive rates for each class:

$$Macro = \frac{1}{n} \frac{C[i,i]}{\sum_{ij1}^{n} C[i,j]} \tag{8}$$

4.2.3. Skills classification evaluation

For the JIGSAWS dataset, we performed experiments to identify the skill level (Novice, Intermediate or Expert) at the trial level. In this experiment, we used the leave one supertrial out (LOSO) cross-validation configuration provided in [9].

For the RAVEN-II dataset, we used a leave-one-out cross-validation approach to first to predict the interface used to control the robot (Leap Motion or Sigma.7) and then to predict the skill level of the trainee (A, B and C). Table 3 (left) presents the performance of our method on this dataset.modif

For the Microsurgery dataset, we used a leave-one-out cross-validation approach to predict the skill level of the trainee (expert or novice). Table 3 (right) presents the performance of our method on this dataset.

4.3. Results

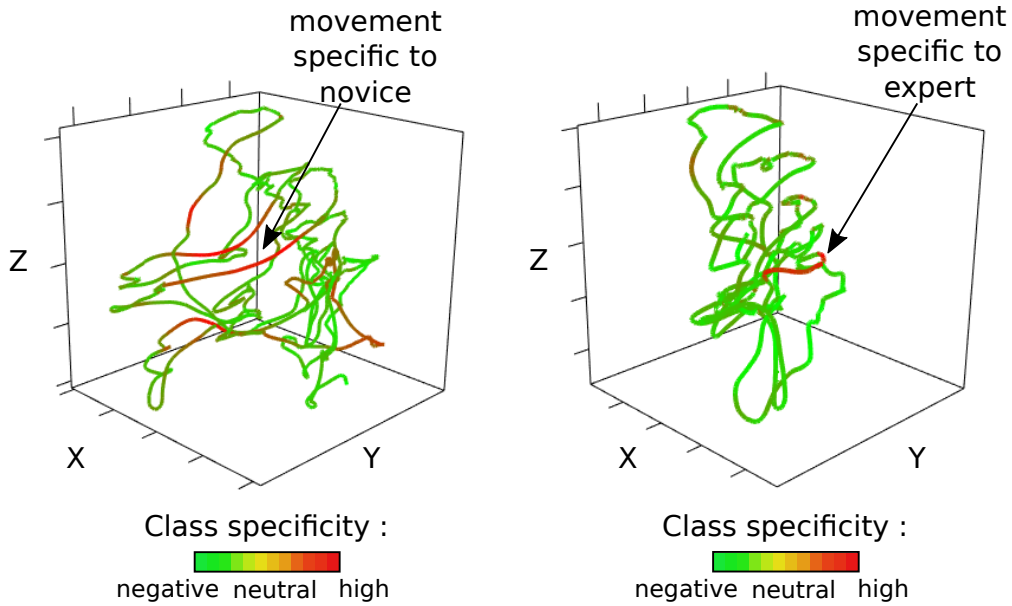
4.3.1. Gesture classification

Table 1 presents the results for gesture classification assuming known boundaries and using kinematic data only for the JIGSAWS dataset. For comparison purposes, we also report state-of-the-art results for Linear Dynamical Systems (LDS) and Hidden Markov Models (HMM). The results have been taken from [44]. The proposed method outperforms both LDS and HMM methods in terms of micro and macro performances for the three tasks and the two cross-validation configurations. These results show that our method accurately identifies patterns that are specific to a gesture motion. One of the interesting features of the proposed method is the ability to use different kinematic data depending on the gesture. As our method computes the frequencies for each component of the kinematic data for each gesture independently, the most discriminative attributes of a given gesture naturally stand out. Furthermore, the *tf∗idf* regularization discards the motion patterns that are common to every gesture (*i.e.*, irrelevant for classification as not distinctive of any class).

The LOUO configuration is known to be particularly challenging, because we attempt to classify gestures of a subject without having any of his or her other attempts.

Table 1: Gesture classification performance, assuming known boundaries and using kinematic data only of JIGSAWS dataset.

JIGSAWS		Leave-one-supertrial-out			Leave-one-user-out		
Method	Metric	Suturing	Needle Passing	Knot Tying	Suturing	Needle Passing	Knot Tying
	(l_s,α)	(8,19)	(13,18)	(15,7)	(8,19)	(14,18)	(10,12)
Proposed	Micro	**93.69**	**81.08**	**92.45**	**88.27**	**75.29**	**89.76**
	Macro	**79.95**	**74.67**	**89.78**	**68.77**	**67.54**	**82.29**
LDS [44]	Micro	84.61	59.76	81.67	73.64	47.96	71.42
LDS [44]	Macro	63.87	46.55	74.51	51.75	32.59	63.99
HMM [44]	Micro	92.56	75.68	89.76	80.83	66.22	78.44
HMM [44]	Macro	79.66	72.36	87.29	65.03	62.70	72.68



(a) Trial 5 of Suturing task of subject B (novice) using Novice class *tf∗idf* vector weights of 5th fold (best viewed in color).

(b) Trial 5 of Suturing task of subject E (expert) using Expert class *tf∗idf* vector weights of 5th fold (best viewed in color).

Figure 4: Example of interpretable feedback using a heat-map visualization of subsequence importance to a class identification. The value corresponds to the combination of the *tf∗idf* weights of all patterns which cover the point.

The good performance of our approach can be explained by its ability to identify highly discriminative patterns that are the most distinctive of each gesture. These results also indicate that our method generalizes well, as shown by the fact that it can accurately classify gestures from unobserved trainees.

4.3.2. Skills classification

Table 2 presents the results for the JIGSAWS dataset for the three tasks and reports micro and macro performances. The results are better for Suturing and Needle Passing tasks than for Knot Tying task. The poor performance on the Knot Tying task can be explained by the minor difference between the Expert and Intermediate subjects for this task (mean GRS is 17.7 and 17.1 for expert and intermediate respectively). We also report the state-of-the-art results from [31] for the Suturing task. The SHMM approach gives better results for the per trial classification configuration as it uses global temporal information, whereas our method is focusing on the local patterns regardless of their location within larger time series. Furthermore, the SHMM approach [31] uses gestures boundaries to learn the temporal model while our method is not using this information.

Table 2: Skill classification performance per trial using kinematic data only of JIGSAWS dataset.

JIGSAWS		Leave-one-supertrial-out		
Method	Metric	Suturing	Needle Passing	Knot Tying
	(l_s,α)	(10,9)	(12,13)	(5,14)
Proposed	Micro	89.74	**96.30**	61.11
	Macro	86.67	95.83	53.33
SHMM [31]	Micro	**97.40**	96.20	**94.40**

Table 3 (left) presents the performance of our method on the RAVEN-II dataset. For interface prediction, the performance is very high as our method successfully predicts the used interface with a 100% accuracy. This means that the method was able to identify inner patterns that are specific to each human-machine interface (Leap Motion and Sigma.7), validating results obtained by the authors. This is particularly interesting as it could allow us to better understand the specificities of each interface and how they differ from each other. This might be used to compare different interfaces and understand which one is more suitable for a specific training task or for a specific training purpose. For skill classification, our method obtained a precision of 83.33% on this dataset (25 out of 30 tasks were correctly classified). Most of the errors were in the B class (resident who only used few times a robotic training system) which is consistent with previous results on this data set (see [22]).

Table 3 (right) presents the performance of our method on Microsurgery dataset. The precision of our method on this dataset for skill level prediction is 85.19% (23 out of 27 are correctly classified). These results can be explained by the difficulty of the studied task of microsurgery suturing. Indeed, the kinematic data contains many artifacts in the movements due to the use of the microscope. Furthermore, as the amount of data is limited, it is very challenging to learn a discriminative pattern in these data.

4.4. Interpretable patterns visualization

An interesting and unique feature of our approach is to output a set of discriminative patterns weighted by the class specificity for each of the input class. These lists of ranked patterns can be studied to better understand what makes each class distinctive. As the use of *tf∗idf* (Eq. (3)) discards patterns that are common to all classes, only patterns having discriminative power remain.

The list of weighted discriminative patterns can be used to visualize, on a given trial, the location of the areas that are specific to the current skill level of the trial. We propose to use a heat map-like visualization technique that provides immediate insight into the layout of the "important" class-characteristic patterns (as described in [6]).

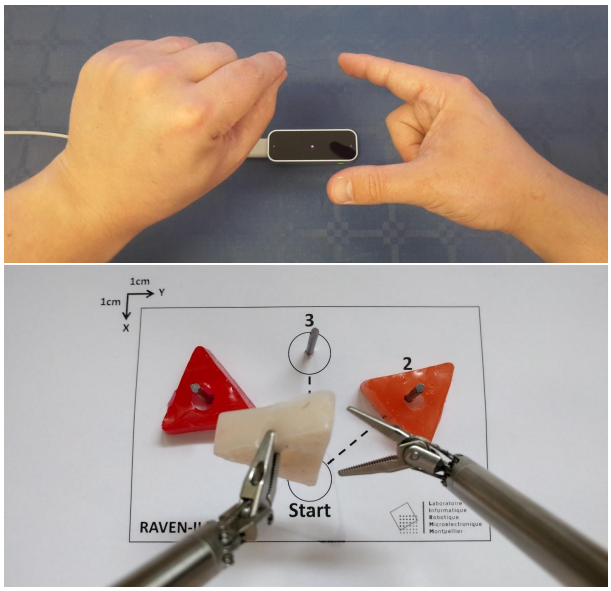


Figure 5: The contactless control interface (Leap Motion) (top) and the RAVEN-II robot (bottow) for surgical training [22].

Figure 4 shows, for the Suturing task of the JIGSAWS dataset, the two individual 5$^{\text{th}}$ trials of subjects B (Novice) and E (Expert), using (x, y, z) coordinates for the right hand. In this figure, we used respectively the *tf∗idf* weights vectors of the 5$^{\text{th}}$ fold for the Novice on subject B and for the Expert on subject E. The red areas correspond to specific motions that are correlated with a skill level. For Subject B (Figure 4a), these areas correspond to motions that were only observed among the novices. By contrast, green areas correspond to motions that are common to all subjects regardless of their skill. This visualization provides a rich information about what makes a specific skill level distinctive and can also be used to provide individual and personalized feedback. As the videos of the trials are also available, this result has to be displayed side-by-side with the videos in order to show to the trainee the movements that are specific. Note that a more detailed analysis could be performed by observing which kinematic data features are specific in these areas or by performing the analysis on a per gesture basis. Visualization (like Figure 4) for all the subject trials for the Suturing task are available on the companion webpage[1].

Note that as the *tf∗idf* weight vectors are computed prior to the classification step, it is possible to display this heat-map visualization in real-time during the trial. We provide a video in the supplementary material attached to this paper that shows the real-time computation of this visualization while a trainee performs a suturing task.

[1] http://germain-forestier.info/src/aiim2018/

Table 3: Performance for interface and skills classification for the RAVEN-II and Microsurgery datasets.

		RAVEN-II Leave-one-out		Microsurgery Leave-one-out
Method	Metric	Interface	Skills	Skills
	(l_s,α)	(6,20)	(6,5)	(6,5)
Proposed	Micro	100	83.33	85.19

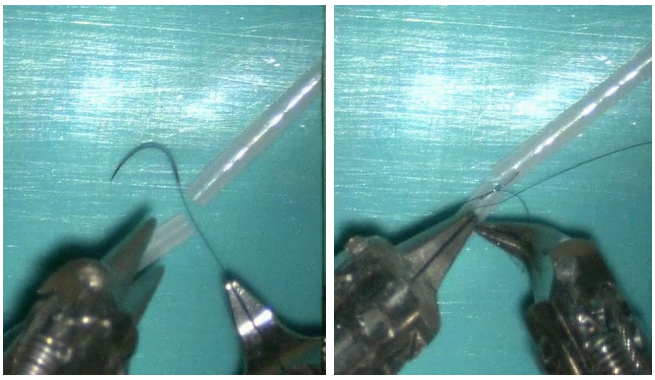


Figure 6: Snapshots of micro-surgical suturing tasks performed using a master-slave robotic platform [63].

5. Discussion

The results presented in the previous section show that the proposed method can be used to successfully classify surgical gesture, surgical skill, and recording interface obtained from different recording systems.

An interesting feature of our approach compared to other existing approaches is the ability of the method to explain the classification by providing discriminative and interpretable patterns. The method not only classify, it explains the reasons behind this classification. Thus, we believe that the proposed approach is a valuable addition to existing learning tools for surgery as it provides a way to obtain a constructive feedback on which parts of an exercise have been used to classify the attempt as correct or incorrect. It is however difficult to evaluate its effect on the teaching processes as it would require to conduct an empirical study with two trainee populations – one with an access to the tool and a control one. Yet another way to evaluate the system would be to use a questionnaire that a trainee would be asked to fill after using the system. In our future work, we plan to conduct these type of studies in order to evaluate how our method could influence the acquisition of technical skills.

The method could also be integrated into a clinical education platform in order to guide trainees in the acquisition of skills. The heat-map visualization computed by the method (Figure 4) could be provided to the trainee along with a system allowing to watch videos of the most discriminative gestures performed by experts.

The analysis of identified patterns should also be conducted in order to evaluate their importance in the skill level classification since it is currently rather difficult to evaluate the influence of a single pattern because the cosine similarity averages out all the patterns. The final affectation is thus an accumulation of small or important "hints" that the trainee belongs to a given skill level. It is not guaranteed that changing a single pattern will automatically change the class of a trainee. However, our future work includes a closer analysis of single patterns in order to better direct the trainee to the most important pattern.

The way the method works also leads to a loss of the temporal order of the patterns. The patterns frequency are indeed computed throughout the surgery regardless of their position in time-line. This part could be improved, as the position of a pattern in the course of a surgery could be informative. On this context, the analysis could be performed for each phase of the surgery [66] instead of on the whole surgery. Finally, while the method currently relies on raw kinematic data to perform the analysis, additional information like force and torque could be used to further improve the results. The method also showed a great potential to compare different surgical motion systems in order to extract useful insights on differences and similarities of surgical motion interfaces.

Finally, we should note that the method is generic enough so that it could be used to process kinematic data generated in domains other than surgery. As sensors (*e.g.*, accelerometers and gyroscopes) are getting broadly used, our approach could benefit other domains to analyze patterns of movements that discriminate between different classes of gestures. For example in [67] a deep convolutional neural network was proposed to classify the motor state of Parkinson's Disease using patients' accelerometer coordinates captured by wearable sensors. Our method could be directly applied using patients' data, thus replacing the black-box model proposed in [67] with an interpretable analysis that highlights which movements contributed the most to a certain diagnosis. Another related area is the Human Activity Recognition where data from wearables [68] could constitute an input time series to our SAX based method thus enabling the interpretation of human motion identification.

6. Conclusion

In this paper, we presented a new method for discovery of discriminative and interpretable patterns in surgical activity motion. Our method uses SAX to discretize the kinematic data into sequence of letters. A sliding window is then used to build bag of words. Finally, $tf{*}idf$ framework is applied to identify motion class-characteristic patterns. Experiments performed on the three surgical

motion datasets have shown that our method successfully classifies gestures, skill levels and recording interfaces. The strong advantage of the proposed technique is the ability to provide a precise quantitative feedback for the classification results. The proposed method is thus a good addition to existing automatic feedback methods. Of course, the evaluation of our visualization approach needs to be performed within curriculum.

Acknowledgement

Dr Petitjean is the recipient of an ARC Discovery Early Career Award (DE170100037) funded by the Australian Government. This work was also funded by ImPACT Program of Council for Science, Technology and Innovation, Cabinet Office, Government of Japan. This work was also supported by the French ANR within the Investissements d'Avenir Program (Labex CAMI, ANR-11-LABX0004); by the Equipex ROBOTEX Program (ANR-10-EQPX-44-01); and by the Région Languedoc-Roussillon.

References

[1] Tsuda S, Scott D, Doyle J, Jones DB. Surgical skills training and simulation. Current problems in surgery 2009;46(4):271–370.

[2] Forestier G, Petitjean F, Riffaud L, Jannin P. Optimal subsequence matching for the automatic prediction of surgical tasks. In: AIME 15th Conference on Artificial Intelligence in Medicine; vol. 9105. Springer; 2015, p. 123–32.

[3] Forestier G, Petitjean F, Riffaud L, Jannin P. Automatic matching of surgeries to predict surgeons' next actions. Artificial intelligence in medicine 2017;81:3–11.

[4] Dlouhy BJ, Rao RC. Surgical skill and complication rates after bariatric surgery. The New England Journal of Medicine 2014;370(3):285–.

[5] Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. Nature Biomedical Engineering 2017;1(9):691.

[6] Senin P, Malinchik S. SAX-VSM: Interpretable time series classification using SAX and vector space model. In: International Conference on Data Mining. IEEE; 2013, p. 1175–80.

[7] Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery 2007;15(2).

[8] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM 1975;18(11):613–20.

[9] Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. Modeling and Monitoring of Computer Assisted Interventions (M2CAI)–MICCAI Workshop 2014;:1–10.

[10] Scott DJ, Valentine R, Bergen PC, Rege RV, Laycock R, Tesfay ST, et al. Evaluating surgical competency with the american board of surgery in-training examination, skill testing, and intraoperative assessment. Surgery 2000;128(4):613 –22.

[11] Cuschieri A, Francis N, Crosby J, Hanna GB. What do master surgeons think of surgical competence and revalidation?11see appendix for participating surgeons. The American Journal of Surgery 2001;182(2):110 –6.

[12] Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the objective structured assessment of technical skills (osats): a systematic review of validity evidence. Advances in Health Sciences Education 2015;20(5):1149–75.

[13] Bridgewater B, Grayson AD, Jackson M, Brooks N, Grotte GJ, Keenan DJM, et al. Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. BMJ 2003;327(7405):13–7.

[14] Reiley CE, Hager GD. Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop. 2009,.

[15] Reiley CE, Plaku E, Hager GD. Motion generation of robotic surgical tasks: Learning from expert demonstrations. In: International Conference on Engineering in Medicine and Biology Society. IEEE; 2010, p. 967–70.

[16] Haro BB, Zappella L, Vidal R. Surgical gesture classification from video data. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer; 2012, p. 34–41.

[17] Zia A, Sharma Y, Bettadapura V, Sarin EL, Ploetz T, Clements MA, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. International journal of computer assisted radiology and surgery 2016;11(9):1623–36.

[18] Islam G, Kahol K, Li B, Smith M, Patel VL. Affordable, web-based surgical skill training and evaluation tool. Journal of Biomedical Informatics 2016;59(Supplement C):102 –14.

[19] Chaudhry R, Vidal R. Recognition of Visual Dynamical Processes: Theory, Kernels, and Experimental Evaluation. Technical Report 2009;(April).

[20] Zappella L, Béjar B, Hager G, Vidal R. Surgical gesture classification from video and kinematic data. Medical Image Analysis 2013;17(7):732–45.

[21] Reiley CE, Hager GD. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. Springer; 2009, p. 435–42.

[22] Despinoy F, Bouget D, Forestier G, Penet C, Zemiti N, Poignet P, et al. Unsupervised trajectory segmentation for surgical gesture recognition in robotic training. IEEE Transactions on Biomedical Engineering 2016;:1280–91.

[23] Gao Y, Vedula SS, Lee GI, Lee MR, Khudanpur S, Hager GD. Unsupervised surgical data alignment with application to automatic activity annotation. In: International Conference on Robotics and Automation. IEEE; 2016, p. 4158–63.

[24] Sharon Y, Nisky I. What Can Spatiotemporal Characteristics of Movements in RAMIS Tell Us? ArXiv e-prints 2017;.

[25] Nisky I, Okamura AM, Hsieh MH. Effects of robotic manipulators on movements of novices and surgeons. Surgical Endoscopy 2014;28(7):2145–58.

[26] Zhou Y, Ioannou I, Wijewickrema S, Bailey J, Kennedy G, O'Leary S. Automated segmentation of surgical motion for performance analysis and feedback. In: International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer; 2015, p. 379–86.

[27] Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet R, Wright A, et al. Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 2014;192(2):329–38.

[28] Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery 2006;11(5):220–30. PMID: 17127647.

[29] Reznick RK, MacRae H. Teaching surgical skills — changes in the wind. New England Journal of Medicine 2006;355(25):2664–9. PMID: 17182991.

[30] Reiley CE, Lin HC, Varadarajan B, Vagvolgyi B, Khudanpur S, Yuh D, et al. Automatic recognition of surgical motions using statistical modeling for capturing variability. Studies in Health Technology and Informatics 2008;132:396.

[31] Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R. Sparse hidden markov models for surgical gesture classification and skill evaluation. In: Information Processing in Computer-Assisted Interventions. Springer; 2012, p. 167–77.

[32] Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov

modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Transactions on Biomedical Engineering 2001;48(5):579–91.
[33] Tao L, Zappella L, Hager GD, Vidal R. Surgical gesture segmentation and recognition. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer; 2013, p. 339–46.
[34] Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Academic Medicine 2004;79(10).
[35] Forestier G, Riffaud L, Petitjean F, Henaux PL, Jannin P. Surgical skills: Can learning curves be computed from recordings of surgical activities? International journal of computer assisted radiology and surgery 2018;13(5):629–36.
[36] Jiang J, Xing Y, Wang S, Liang K. Evaluation of robotic surgery skills using dynamic time warping. Computer Methods and Programs in Biomedicine 2017;152(Supplement C):71 – 83.
[37] Shafiei SB, Cavuoto L, Guru KA. Motor skill evaluation during robot-assisted surgery. 2017.
[38] Li K, Burdick JW. A Function Approximation Method for Model-based High-Dimensional Inverse Reinforcement Learning. ArXiv e-prints 2017;arXiv:1708.07738.
[39] Marban A, Srinivasan V, Samek W, Fernandez J, Casals A. Estimating position & velocity in 3d space from monocular video sequences using a deep neural network. In: The IEEE International Conference on Computer Vision (ICCV). 2017,.
[40] Rupprecht C, Lea C, Tombari F, Navab N, Hager GD. Sensor substitution for video-based action recognition. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016, p. 5230–7.
[41] Sarikaya D, Corso JJ, Guru KA. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Transactions on Medical Imaging 2017;36(7):1542–9.
[42] Fard MJ, Pandya AK, Chinnam RB, Klein MD, Ellis RD. 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 2017;13(3):e1766–n/a. E1766 RCS-16-0026.R2.
[43] Bani MJ, Jamali S. A New Classification Approach for Robotic Surgical Tasks Recognition. ArXiv e-prints 2017;arXiv:1707.09849.
[44] Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Bejar B, et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering 2017;.
[45] Fard MJ, Ameri S, Chinnam RB, Ellis RD. Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery. IEEE Robotics and Automation Letters 2017;2(1):171–8.
[46] Li S, Li K, Fu Y. Temporal subspace clustering for human motion segmentation. 2015 IEEE International Conference on Computer Vision (ICCV) 2015;:4453–61.
[47] Mavroudi E, Bhaskara D, Sefati S, Ali H, Vidal R. End-to-End Fine-Grained Action Segmentation and Recognition Using Conditional Random Field Models and Discriminative Sparse Coding. ArXiv e-prints 2018;.
[48] Ding L, Xu C. TricorNet: A Hybrid Temporal Convolutional and Recurrent Network for Video Action Segmentation. ArXiv e-prints 2017;arXiv:1705.07818.
[49] Lea C, Vidal R, Reiter A, Hager GD. Temporal convolutional networks: A unified approach to action segmentation. In: Hua G, Jégou H, editors. Computer Vision – ECCV 2016 Workshops. Cham: Springer International Publishing. ISBN 978-3-319-49409-8; 2016, p. 47–54.
[50] Lea C, Reiter A, Vidal R, Hager GD. Segmental spatiotemporal cnns for fine-grained action segmentation. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016. Cham: Springer International Publishing. ISBN 978-3-319-46487-9; 2016, p. 36–52.
[51] Fard MJ, Ameri S, Chinnam RB, Pandya AK, Klein MD, Darin Ellis R. Machine Learning Approach for Skill Evaluation in Robotic-Assisted Surgery. ArXiv e-prints 2016;.
[52] Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery 2018;14(1):e1850–n/a. E1850 RCS-16-0174.R4.
[53] Zia A, Essa I. Automated Surgical Skill Assessment in RMIS Training. ArXiv e-prints 2017;arXiv:1712.08604.
[54] Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (osats) for surgical residents. British Journal of Surgery 1997;84(2):273–8.
[55] Höppner F. Time series abstraction methods-a survey. In: GI Jahrestagung. 2002, p. 777–86.
[56] Patel P, Keogh E, Lin J, Lonardi S. Mining motifs in massive time series databases. In: International Conference on Data Mining. IEEE; 2002, p. 370–7.
[57] Moskovitch R, Shahar Y. Classification-driven temporal discretization of multivariate time series. Data Mining and Knowledge Discovery 2015;29(4):871–913.
[58] Schäfer P. Scalable time series classification. Data Mining and Knowledge Discovery 2015;:1–26.
[59] Sefati S, Cowan NJ, Vidal R. Learning shared, discriminative dictionaries for surgical gesture segmentation and classification. In: Modeling and monitoring of computer assisted interventions (M2CAI)–MICCAI Workshop. Citeseer; 2015,.
[60] Tavenard R. tslearn: A machine learning toolkit dedicated to time-series data. 2017. https://github.com/rtavenar/tslearn.
[61] Manning CD, Raghavan P, Schütze H, et al. Introduction to information retrieval; vol. 1. Cambridge University Press; 2008.
[62] Hannaford B, Rosen J, Friedman DW, King H, Roan P, Cheng L, et al. Raven-ii: an open platform for surgical robotics research. IEEE Transactions on Biomedical Engineering 2013;60(4):954–9.
[63] Mitsuishi M, Morita A, Sugita N, Sora S, Mochizuki R, Tanimoto K, et al. Master–slave robotic platform and its feasibility study for micro-neurosurgery. The International Journal of Medical Robotics and Computer Assisted Surgery 2013;9(2):180–9.
[64] Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. Development of a model for training and evaluation of laparoscopic skills. The American journal of surgery 1998;175(6):482–7.
[65] Garraud C, Gibaud B, Penet C, Cazuguel G, Dardenne G, Jannin P. An ontology-based software suite for the analysis of surgical process model. Surgetica 2014;.
[66] MacKenzie L, Ibbotson J, Cao C, Lomax A. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minimally Invasive Therapy & Allied Technologies 2001;10(3):121–7.
[67] Um TT, Pfister FMJ, Pichler D, Endo S, Lang M, Hirche S, et al. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In: ACM International Conference on Multimodal Interaction. 2017, p. 216–20.
[68] Hammerla NY, Halloran S, Plötz T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In: International Joint Conference on Artificial Intelligence. 2016, p. 1533–40.