Finding discriminative and interpretable patterns in sequences of surgical activities

Germain Forestier[a,b,*], François Petitjean[b], Pavel Senin[c], Laurent Riffaud[d,e],
Pierre-Louis Henaux[d,e], Pierre Jannin[d]

[a] *MIPS EA 2332, University of Haute-Alsace, Mulhouse, France*
[b] *Faculty of Information Technology, Monash University, Melbourne, Australia*
[c] *Los Alamos National Laboratory, Los Alamos, NM, 87544*
[d] *INSERM MediCIS, Unit U1099 LTSI, University of Rennes 1, Rennes, France*
[e] *Department of Neurosurgery, Pontchaillou University Hospital, Rennes, France*

Abstract

Objective. Surgery is one of the riskiest and most important medical acts that is performed today. Understanding the ways in which surgeries are similar or different from each other is of major interest to understand and analyze surgical behaviors. This article addresses the issue of identifying discriminative patterns of surgical practice from recordings of surgeries. These recordings are sequences of low-level surgical activities representing the actions performed by surgeons during surgeries.
Material and Method. To discover patterns that are specific to a group of surgeries, we use the Vector Space Model (VSM) which is originally an algebraic model for representing text documents. We split long sequences of surgical activities into subsequences of consecutive activities. We then compute the relative frequencies of these subsequences using the *tf∗idf* framework and we use the Cosine similarity to classify the sequences. This process makes it possible to discover which patterns discriminate one set of surgeries recordings from another set.
Results. Experiments were performed on 40 neurosurgeries of anterior cervical discectomy (ACD). The results demonstrate that our method accurately identifies patterns that can discriminate between (1) locations where the surgery took place, (2) levels of expertise of surgeons (*i.e.*, expert vs. intermediate) and even (3) individual surgeons who performed the intervention. We also show how the *tf∗idf* weight vector can be used to both visualize the most interesting patterns and to highlight the parts of a given surgery that are the most interesting.
Conclusions. Identifying patterns that discriminate groups of surgeon is a

[*] Corresponding author – ENSISA - Universite 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), `psenin@lanl.gov` (Pavel Senin), `laurent.riffaud@chu-rennes.fr` (Laurent Riffaud), `pierrelouis.henaux@chu-rennes.fr` (Pierre-Louis Henaux), `pierre.jannin@univ-rennes1.fr` (Pierre Jannin)

very important step in improving the understanding of surgical processes. The proposed method finds discriminative and interpretable patterns in sequences of surgical activities. Our approach provides intuitive results, as it identifies automatically the set of patterns explaining the differences between the groups.

Keywords: Temporal Analysis, Vector Space Model, Bag of Words, Surgical Process Modelling, Surgical Technical Skills, Surgery

1. Introduction

More than half a million surgeries are performed every day worldwide [1], which makes surgery one of the most important component of global health care. Competing demands are motivating a better understanding of surgical processes, including: surgical procedures are getting more complex [2], residents now have to be trained while performing less procedures [3], the surgical interventions need increasingly thorough justification [4] and the costs have to be reduced [5]. A better understanding of surgical practices is key to addressing these issues. Surgical Process Modelling (SPM) [6] is the general process that aims at understanding surgeries, in order to improve the quality of care and the training of surgeons. SPM is part of *surgical data science* [7], which targets the development of data-driven methods to support surgery. SPM traditionally considers surgeries as sequences of activities that are performed by the surgeon over the course of the surgery.

Previous work on the analysis of surgeries considered the comparison of entire sequences of surgical activities. For example, Forestier et al. [8] used Dynamic Time Warping (DTW) as a dissimilarity measure between sequences of surgical activities. This measure was used to create groups of similar surgeries and made it possible to cluster surgeons according to their expertise. This approach was later used in [9] to perform a multi-site study comparing the surgical behaviors in France, Germany and Canada. This study revealed differences in surgical practice depending on the expertise of the surgeon and the location where the surgery took place. Forestier et al. [10] also proposed Non-Linear Temporal Scaling (NTLS), a new approach for realigning a set of surgeries on the same timeline. This method calculates an average surgery that is used as a reference for the realignment. Using the realigned sequences, NTLS offers a visualization that makes it possible to understand the differences and common parts in a set of surgeries. Neumuth et al. [11] also investigated different similarity metrics for surgical process models. Five different similarity metrics were compared with the objective to deal with several dimensions of process compliance in surgery, including granularity, content, time, order, and frequency of surgical activities. These approaches have limitations because of their global approach: any unusual event in a given surgery has to be matched to the element of another surgery, which make these methods sensitive to noise. They are also difficult to understand because the only explanation about the prediction that they provide is the most similar surgery that was found in the database; as surgeries are complex processes, this is often not informative enough to understand the why of the predictions. Furthermore, existing approaches have been mostly used to evaluate similarities between surgeries and not for finding and describing important differences between them.

In this article, we address the issue of identifying discriminative patterns from recordings of surgeries in order to better understand surgical practice. These recordings are sequences of low-level surgical activities representing the actions performed by surgeons during surgeries. Our objective is to analyze these recordings to find discriminative patterns that characterize specific behavior of a group of surgeries over a baseline. Identifying patterns that separate groups of surgeries is a very important step in improving the understanding of surgical processes. It makes possible to easily explain the main differences in the way multiple surgeries were performed: e.g. what makes the behavior of senior surgeons unique compared to junior surgeons, or what makes the behavior of French surgeons different from the one of German surgeons. Comparing the practice of surgeons according to their experience is of major interest from a teaching perspective [12].

The rest of the paper is organized as follows: Section 2 presents our method to find discriminative patterns using a sliding window technique in conjunction with the Vector Space Model (VSM). Section 3 presents the assessment of our method on a dataset composed of 40 neuro-surgeries of anterior cervical discectomy (ACD) surgeries. Finally, we discuss the results in Section 4 as well as the advantages and drawbacks of our method. Section 5 concludes the paper.

2. Method

2.1. Surgeries as sequences of activities

We consider surgeries as sequences of activities that are performed by a surgeon during an intervention. Mehta et al. [13] proposed to represent surgical activities as triplet composed of an *action*, an *anatomical structure* and an *instrument*. For example, the surgeon can *cut* the *skin* using a *scalpel* with his/her right hand. In this paper, we use this formalization which was introduced in [8].

Let $\mathbb{S} = \{S_1, \cdots, S_N\}$ be the a set of surgeries. A surgery S can be modeled as a sequence of surgical activities $S =< a_1, ..., a_n >$ where a_i denotes the i^{th} activity. An activity a_i belongs to $\mathcal{A}$, the set of all possible activities, and has a start time and a stop time within the time-line of the surgery. In general, activities that are performed by both hands are recorded, as well as the use of the microscope. In this paper, we focus on the activities that were performed by the right hand (*i.e.*, the dominant hand in our dataset), as previous studies [8] showed that they are the activities that carry the most important information. Figure 1 illustrates one sequence of activities, where each activity is in a different color.

Given multiple sets of surgeries ($\{\mathbb{S}_1 \cdots \mathbb{S}_N\}$) our goal is to find subsequences of activities that are specific to each set. These sets are defined according to

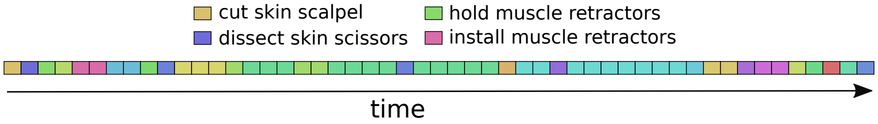

Figure 1: Illustration of one surgery recording as a sequence of surgical activities.

the targeted application, for example by regrouping surgeries performed in the same location or performed by surgeons having the same level of expertise (*e.g.*, junior or senior). The subsequences are expected to be present in most of the sequences of a given set and absent from the other sets. The underlying idea is to discover what makes a set of surgeries unique compared to other surgeries.

2.2. Proposed method

The proposed method starts by splitting sequences into subsequences of consecutive activities. It then computes the relative frequencies of these subsequences, *i.e.*, the number of times they appear in a given sequence and in a set of sequences. We extract these subsequences from a set of surgeries, and use their relative frequencies to find discriminative patterns that characterize specific behavior of a group of surgeries over another.

To discover the patterns that are specific to a group of surgeries, we use the VSM [14] framework which is originally an algebraic model for representing text documents. Using this paradigm, the subsequences extracted from a set of surgeries are interpreted as a single *bag of words*, where *words* here represent subsequences of surgical activities. Bags of words are extracted from the sequences of surgical activities using a sliding window. We then use the well-know *tf∗idf* weighting scheme [14], which ranks patterns based on their relative frequencies. This weighting scheme makes it possible to discard patterns that are frequent across all classes/groups; the idea being that even if a pattern is frequent, if it is so in all groups, then it will not be discriminant. We can then analyze these patterns to better understand the specificities of a set of surgeries over another; for example identify subsequences that are *only* present in a given set of surgeries. We can then also reproject these patterns over the sequences themselves to outline portions that are more or less characteristic of one class. Finally, we show that these patterns can also be used to predict: we transform the "query surgery" into its VSM representation and predict if its bag of words resemble more to one group of surgeries or another (typically using the cosine similarity). The classification process is able to average out the subsequences that are common to most surgeries, in order to focus on the most discriminant ones.

2.3. Vector Space Model

We use the vector space model exactly as it is known in Information Retrieval (IR) [14, 15]. The *tf∗idf* weight for a term t is defined as a product of two factors:

term frequency (*tf*) and inverse document frequency (*idf*). For the first factor, we use logarithmically scaled term frequency [16]:

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

where t is the term, d is a bag of words (a document in IR terms), and $\mathrm{f}_{t,d}$ is a frequency of the term in a bag. The inverse document frequency we compute as usual [16]:

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

where N is the cardinality of a corpus D (the total number of classes) and the denominator df_t is a 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

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

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

Once all frequency values are computed, the term frequency matrix becomes the term weight matrix, whose columns used as *class term weight* vectors that facilitate the classification using Cosine similarity. For two vectors **a** and **b** Cosine similarity is based on their inner product and defined as

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

2.4. Vector Space Model for surgeries analysis

The first step to apply *tf∗idf* scheme to sequences of surgical activities is to convert them into bags of words. The sliding window size (w) is a parameter of this step. It defines the length of the words that will be present in the bags. In our case, this corresponds to considering sets of w consecutive surgical activities. Figure 2 illustrates the computation of the subsequences using an overlapping sliding window of size 5 ($w = 5$). The influence of the size of the sliding window will be discussed in Section 3.2. This process is performed for all of the N sets that will be used in the analysis (*e.g.*, set of junior surgeries, set of senior surgeries, etc.) leading to N bags (*i.e.*, one per group). This technique was previously used in activity recognition from video [17, 18] where this process is referred as extracting *n-grams* frequency histograms, n being the width of the sliding window.

Once we have constructed the N bags of words, we compute the frequency of each word in every bag (Eq. 1), and apply the *tf∗idf* weighting (Eq. 3). This step makes it possible to reduce the importance of frequent patterns that are so in most groups, because patterns that appear frequently in all groups cannot discriminate between them. Figure 3 illustrates the computation of the bag of words for two sets of surgeries and the computation of the *tf∗idf* weight vectors.

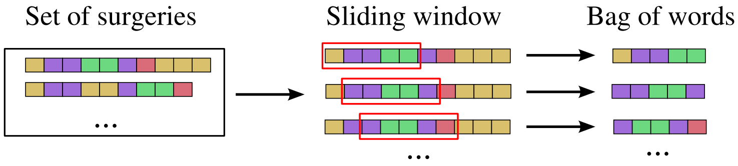

Figure 2: Illustration of the generation of a bag of words from a set of surgeries.

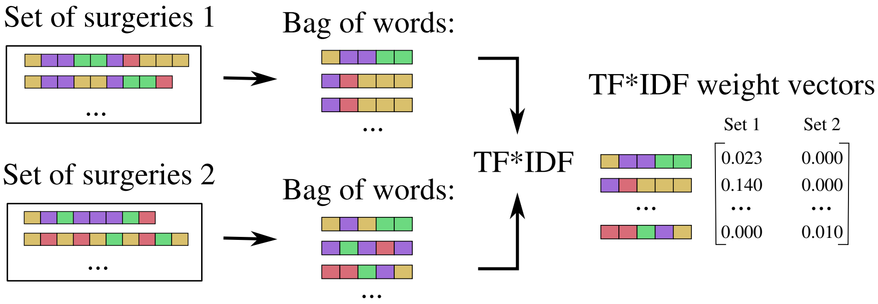

Figure 3: Illustration of the generation of the *tf*∗*idf* vectors from two sets of surgeries.

Once the *tf*∗*idf* weight vectors computed, it is possible to rank the patterns according to their relative weight. The weight of a pattern can be naturally interpreted as its *importance* for discriminating of a specific group (*i.e.*, a set of surgeries). The term frequency (*tf*) allows to highlight the patterns that are often present in the set, while the inverse document frequency (*idf*) weighting allows to increase the importance of the patterns that are only present in this set and reduce the importance of patterns that are present in multiple sets.

2.5. Performing classification using VSM

In order to classify an unlabeled sequence of surgical activities, we first transform the input sequence into its VSM representation using exactly the same sliding window used to learn the model. We then compute the Cosine similarity values between its term frequency vector and the N *tf*∗*idf* weight vectors representing the N different groups of surgeries. The unlabeled sequence is assigned to the group whose vector yields maximum cosine similarity value (Eq. 4).

2.6. Visualizing the importance of a pattern

Since the vector space model approach outputs *tf*∗*idf* weight vectors for all subsequences extracted within a group a surgery, it is possible to find the weight of any arbitrary selected subsequence. This feature makes it possible to visualise the results using a heat-map, which provides an immediate insight into the layout of important discriminative subsequences.

Figure 4: Illustration of the on-line recording of the data in the operating room.

3. Experiments and results

3.1. Surgical dataset used in the experiments

Experiments were performed on one-level anterior cervical discectomy (ACD) surgeries [9]. During this procedure, a cervical disc can be removed through an anterior approach. This means that surgery is done through the front of the neck as opposed to the back of the neck. A one-level ACD surgery can usually be decomposed into four major phases: the approach, the discectomy, the arthrodesis, and the closure phases. An additional phase of hemostasis may be mandatory in certain cases. Forty surgeries were recorded on-line using the Surgical workflow Editor [19] resulting in the creation of forty sequences of activities. Figure 4 illustrates the recording of the data in the operating room. Surgeries were performed at the Neurosurgery departments of: (1) the Rennes University Hospital, France, (2) the Leipzig University Hospital, Germany, and (3) the Montreal Neurological Institute and Hospital, McGill University, Canada. Among the 40 surgeries, 11 were performed at site A, 18 were performed at site C, and 11 at site B (we used site A, B and C as anonymized site names). As for the expertise level of the attending surgeon, site C had two expert and two intermediate surgeons participating in the study, site A had one intermediate and three expert surgeons participating, while in site B, all participating surgeons were considered to be expert surgeons. Table 1 presents the information for each surgeon involved in the study: the location of the acquisition (sites A, B and C), the index of the surgeon (1 to 11) and his/her level of expertise (E: Expert, I: Intermediate). Expert surgeons were defined as those who already performed more than 200 ACD surgeries, whereas intermediate surgeons were fully trained neurosurgeons but who performed less than 100 ACD procedures. SPMs were acquired on-line by the same operator (an expert neurosurgeon) in site A and site C, whereas SPMs of site B were acquired by an intermediate surgeon, both having the same training on the software. Figure 5 presents boxplots of the duration of the interventions according to the location and expertise of the surgeon. Figure 5 reveals that it is not possible to rely on the duration of the surgery to accurately classify according to location or experience levels.

Surgeon ID	Expertise	Location	# Surgeries recorded
1	Intermediate	Site A	3
2	Expert	Site A	3
3	Expert	Site A	3
4	Expert	Site A	2
5	Expert	Site B	6
6	Expert	Site B	2
7	Expert	Site B	3
8	Expert	Site C	6
9	Expert	Site C	6
10	Intermediate	Site C	2
11	Intermediate	Site C	4

Table 1: List of the surgeons involved in the study with their location and expertise level.

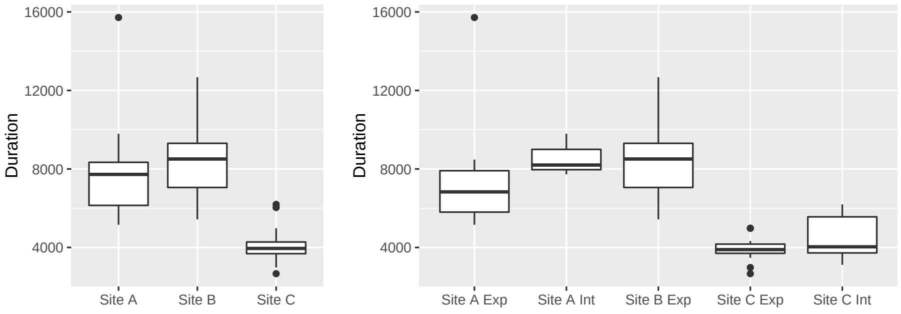

Figure 5: Distribution of intervention duration according to the location (left) and surgeon expertise (right).

For example, while the overall duration is significantly different in site C, they are very similar in sites A and B (confirmed by a Welch t-test comparison of the two distributions with $p = 0.7087$). The same conclusion applies to experience levels where the duration between expert and intermediate surgeons is not statistically different in site A ($p = 0.5742$) or in site C ($p \approx 0.326$).

3.2. Selection of the sliding window size parameter

To apply our method, we have to set the size of the sliding window (w) used to create the bag of words. In all the experiments, this parameter was learnt by cross-validation on the learning set, using a greedy search: we started with $w = 1$ and increased its value as long as the (cross-validated) accuracy of the classifiers increased. To provide an intuition about the influence of w on the accuracy, we present in Figure 6 the evolution of the accuracy on the learning set according to different window sizes for the first three experiments. In these experiments, the best values was always between 3 to 5 with no important

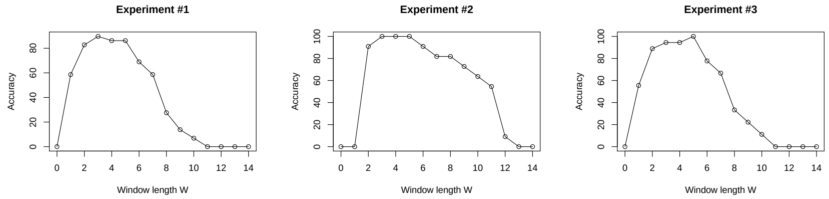

Figure 6: Evolution of the accuracy according to the window length.

variations (see the plateau on Figure 6). This means that cross-validation works well to automatically set the value of w, but also that if a value between 3 and 5 would work well if the user decides to set it manually. A detailed description of the experiments is given in the following section.

3.3. Description of the experiments

In this section we present the experiments performed to evaluate the ability of the proposed method to identify patterns that are specific to a group of surgeries and how these patterns can be used to classify the surgery accurately. Six different experiments were performed to predict alternatively:

- the location where the surgery took place (experiment #1).
- the expertise of the surgeon (experiments #2 to #5)
- which surgeon performed the surgery (experiment #6)

In experiment #1, the goal was to identify patterns that are specific of a surgery department of a hospital. In this experiment, we used the data from sites A, B and C. The identified patterns are useful to understand the differences in practice between different countries or surgery departments.

In experiments #2 to #5, we looked for patterns that are specific to either expert or intermediate surgeons. As sites A and C are the only sites that contain expert and intermediate surgeons (B contains only experts), we first carry out experiments for these sites independently. We then combine the data from A and C and repeat the experiments of trying to identify skill level. Finally, we pool the data from sites A, B and C together and study the same question; given that site B only adds expert-performed surgeries, this allows us to observe the influence of class imbalance on the results. This study was designed to evaluate if the differences between expert and intermediate surgeons were related to the location site. The identified patterns are useful to understand the characteristics of an expert surgeon and what are the main differences between expert and intermediate surgeons. These patterns are also useful to support the automatic assessment of surgical skills.

Finally, in experiment #6 the goal is to identify patterns that are specific to one given surgeon. By comparing the surgeries performed by one surgeon to the ones performed by all the other surgeons, the method is able to identify the

Table 2: Compared accuracies of the 1-Nearest Neighbor classifier using the Euclidean distance, Dynamic Time Warping and our proposed VSM model. We use leave-one-out cross-validation. Values in boldface represent the best results obtained for each experiment.

Exp	Prediction	Sites			Euclidean	DTW	VSM
		A	B	C			
#1	Location	☒	☒	☒	92.5% ± 8.2	97.5% ± 4.8	**100%** ± 0 ($w = 3$)
#2	Expertise	☒			88% ± 19.2	94.5% ± 13.5	**100%** ± 0 ($w = 3$)
#3	Expertise			☒	81% ± 18.1	**100%** ± 0	**100%** ± 0 ($w = 5$)
#4	Expertise	☒	☒		75.9% ± 15.6	**93%** ± 9.3	89.6% ± 11.1 ($w = 3$)
#5	Expertise	☒	☒	☒	82.5% ± 11.8	**95%** ± 6.7	85% ± 11.1 ($w = 5$)
#6	Surgeon Id	☒	☒	☒	45% ± 15.4	65% ± 14.8	**77.5%** ± 12.9 ($w = 3$)

behavioral characteristics that are unique to this surgeon, which can be seen as a proxy for his or her surgical *signature*. In this experiment, we use the data from all sites.

As competitors, we used a 1-Nearest Neighbor (1-NN) classifier using as similarity metric (1) the Euclidean distance and (2) DTW score. We selected the Euclidean distance and DTW in conjunction with 1-NN classifier as this combination has proven to be extremely efficient for time-series and sequence classification [20], and in particular for SPM [8, 9]. For the proposed method (referred as VSM), we used the Cosine similarity as presented in Eq. 4. We used a leave-one-out cross-validation approach consisting in alternatively taking one surgery out of the set and classifying it using the remaining ones. For the Euclidean and DTW methods, each surgery that is left out is compared to the set of remaining surgeries. The class of the identified nearest surgery is then compared to the actual class. For the VSM approach, each surgery that is left out is compared to the *tf∗idf* weight vectors that are learnt on the set of remaining surgeries. The prediction is performed by taking the maximal cosine similarity value. We used the accuracy to compare the results (*i.e.*, number of correct predictions over the total number of predictions). We also computed confidence intervals with an estimation of the true error at a confidence level of 95% [21]: $1.96 \times \sqrt{\frac{(error)(1-error)}{n}}$ where $error$ is the error-rate (number of incorrect predictions over the number of predictions) and n the number of samples used in the experiment.

3.4. Results

Table 2 presents the results of our six experiments with their associated confidence intervals. We can first observe that the Euclidean distance performs poorly and does not once achieve better accuracy than either DTW or VSM. Our method appears competitive and complementary to DTW. Our VSM model uniformly outperformed DTW for location and surgeon prediction, while DTW seems competitive to predict the level of expertise of surgeons. Furthermore, it is interesting to observe that, when the number of examples per class is limited (which is particularly the case for prediction of the surgeon), VSM seems to

#	Pattern ($w = 3$)	Site A	Site B	Site C
1	`di-fa-cl in-mu-re di-fa-cl`	23.00	0.00	2.00
2	`re-li-ro di-li-ho re-li-ro`	16.00	0.00	7.00
3	`di-fa-cl di-fa-cl di-fa-cl`	15.00	0.00	1.00
4	`in-mu-re di-fa-cl di-fa-cl`	13.00	0.00	0.00
5	`in-mu-re di-fa-cl in-mu-re`	13.00	0.00	2.00
6	`re-di-ro di-di-ho re-di-cu`	13.00	0.00	4.00
7	`in-ve-fl in-ve-re in-ve-fl`	12.00	0.00	0.00
8	`re-di-ro ho-di-su re-di-ro`	12.00	0.00	0.00
9	`in-ve-fl in-ve-fl in-ve-fl`	12.00	0.00	0.00
10	`re-li-ro re-li-ro re-li-ro`	12.00	6.00	10.00

Table 3: Top 10 patterns from site A in experiment #1 sorted by *tf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

cope better with the lack of data. We attribute this behavior to the ability of our VSM model to collect a relatively robust description of the sequences; this is because even after having scanned one sequence, we have already collected several thousands of 'words' and their statistics.

An interesting features of the VSM approach is to provide the set of patterns that are the most distinctive for a given group of surgeries. Analysis of these top patterns allows us to understand what makes a group of surgeries specific. To highlight the influence of using *tf∗idf*, we first present in Table 3 the top 10 subsequences patterns with regards to *tf* alone for the task of predicting the location of the surgery (#1 in Table 2). This corresponds to the 10 most frequent subsequences in the surgeries that were performed in Site A, and the corresponding frequencies of these patterns in site B and C. In this table, the names of the surgical activities were shortened. For example, `di-fa-cl` stands for *dissect* the *fascia* with a *classic-cottonoids-forceps*. Table 9 provides the legend for the abbreviations used in the sub-sequences.

Table 4 presents the top 10 sequences after the application of the *idf* factor (see Eq. 3). We can observe that some patterns from Table 3 were discarded. For example, the pattern 2 disappeared as it also appeared in Site C. The pattern 10 was also discarded as it appeared in Site B and Site C. The only remaining patterns, are the patterns that are frequent in Site A, and not frequent in Site B and C, because of their specificity. Tables 5 and 6 present the most frequent patterns respectively for Site B and site C.

Table 7 and 8 present the patterns for second experiment (#2 in Table 2) on classifying the surgeons of site A according to their experience. The tables present the 10 most discriminative patterns according to *tf∗idf* factor for the two groups (*i.e.*, expert vs. intermediate).

Finally, it is also possible to use the *tf∗idf* weight vectors of all subsequences extracted from a group of surgeries to highlight the important subsequences in a given complete surgery. This feature enables a heat map like visualization technique that provides an immediate insight into the layout of important class-characterizing subsequences. Figure 7 illustrates a sequence of an expert sequence of activities from experiment #2. It shows in red the subsequences that are specific to expert surgeons, while in green the subsequences that are

#	Pattern ($w = 3$)	Site A	Site B	Site C
1	`in-mu-re di-fa-cl di-fa-cl`	6.20	0.00	0.00
2	`in-ve-fl in-ve-re in-ve-fl`	5.72	0.00	0.00
3	`re-di-ro ho-di-su re-di-ro`	5.72	0.00	0.00
4	`in-ve-fl in-ve-fl in-ve-fl`	5.72	0.00	0.00
5	`ho-di-su re-di-ro ho-di-su`	5.24	0.00	0.00
6	`in-ve-re in-ve-fl in-ve-re`	4.77	0.00	0.00
7	`cu-mu-sc co-sk-bi cu-mu-sc`	3.81	0.00	0.00
8	`co-sk-bi cu-mu-sc co-sk-bi`	3.81	0.00	0.00
9	`di-fa-di in-mu-re di-fa-di`	3.81	0.00	0.00
10	`di-fa-di di-fa-di in-mu-re`	3.33	0.00	0.00

Table 4: Top 10 patterns from site A in experiment #1 sorted by *tf∗idf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

#	Pattern ($w = 3$)	Site A	Site B	Site C
1	`di-mu-sc co-fa-bi di-mu-sc`	0.00	60.11	0.00
2	`di-fa-sc co-fa-bi di-fa-sc`	0.00	60.11	0.00
3	`co-fa-bi di-mu-sc co-fa-bi`	0.00	53.43	0.00
4	`co-fa-bi di-fa-sc co-fa-bi`	0.00	50.57	0.00
5	`se-sk-ne cu-sk-sc se-sk-ne`	0.00	14.31	0.00
6	`ho-di-cu dr-di-cu ho-di-cu`	0.00	12.40	0.00
7	`dr-di-cu ho-di-cu dr-di-cu`	0.00	12.40	0.00
8	`ho-de-su ho-de-su ho-de-su`	0.00	11.45	0.00
9	`cu-sk-sc se-sk-ne cu-sk-sc`	0.00	10.97	0.00
10	`se-fa-ne cu-sk-sc se-fa-ne`	0.00	9.54	0.00

Table 5: Top 10 patterns from site B in experiment #1 sorted by *tf∗idf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

#	Pattern ($w = 3$)	Site A	Site B	Site C
1	`in-mu-cl in-ve-fl re-mu-cl`	0.00	0.00	5.24
2	`co-sk-bi in-sk-cl in-sk-re`	0.00	0.00	5.24
3	`cu-sk-sc co-sk-bi in-sk-cl`	0.00	0.00	4.77
4	`in-mu-re in-mu-re co-li-fo`	0.00	0.00	4.77
5	`in-ve-fl re-mu-cl in-mu-re`	0.00	0.00	3.81
6	`in-mu-re co-li-fo ir-li-sa`	0.00	0.00	3.33
7	`di-fa-cl di-fa-sc di-fa-di`	0.00	0.00	2.86
8	`in-ve-ar in-ve-ar in-ve-ar`	0.00	0.00	2.86
9	`in-mu-re co-li-fo cu-li-sc`	0.00	0.00	2.86
10	`ir-di-sa in-di-ar in-di-ar`	0.00	0.00	2.38

Table 6: Top 10 patterns from site C in experiment #1 sorted by *tf∗idf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

#	Pattern ($w = 3$)	Exp	Inter
1	`re-di-ro di-di-ho re-di-cu di-di-ho`	2.10	0.00
2	`di-fa-di di-fa-di in-mu-re di-fa-di`	1.80	0.00
3	`di-di-ho re-li-ro di-di-ho re-li-ro`	1.50	0.00
4	`re-di-ro re-di-ho re-di-ro re-di-ho`	1.50	0.00
5	`re-li-ro re-li-ro re-li-ro re-li-ro`	1.50	0.00
6	`re-di-ro di-di-ho re-di-cu re-di-ro`	1.50	0.00
7	`re-li-ro di-di-ho re-li-ro di-di-ho`	1.50	0.00
8	`di-fa-di in-mu-re di-fa-di di-fa-di`	1.50	0.00
9	`re-di-ro re-di-ho re-di-ho re-di-ro`	1.20	0.00
10	`re-di-ho re-di-ro re-di-ho re-di-ro`	1.20	0.00

Table 7: Top 10 patterns from site A expert surgeons in experiment #2 sorted by *tf∗idf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

#	Pattern ($w = 3$)	Exp	Inter
1	`di-fa-cl in-mu-re di-fa-cl in-mu-re`	0.00	2.40
2	`ho-di-su re-di-ro ho-di-su re-di-ro`	0.00	2.40
3	`co-sk-bi cu-mu-sc co-sk-bi cu-mu-sc`	0.00	2.10
4	`re-di-ro ho-di-su re-di-ro ho-di-su`	0.00	2.10
5	`cu-mu-sc co-sk-bi cu-mu-sc co-sk-bi`	0.00	2.10
6	`in-mu-re di-fa-cl di-fa-cl di-fa-cl`	0.00	1.50
7	`di-fa-cl di-fa-cl di-fa-sc di-fa-cl`	0.00	1.50
8	`di-fa-cl di-fa-sc di-fa-cl di-fa-cl`	0.00	1.20
9	`co-mu-bi di-fa-sc co-mu-bi di-fa-sc`	0.00	1.20
10	`re-di-ro re-di-ro ho-di-su re-di-ro`	0.00	0.90

Table 8: Top 10 patterns from site A intermediate surgeons in experiment #2 sorted by *tf∗idf* score. The highest values (very frequent) are depicted in green while the lowest values (rare) are in red.

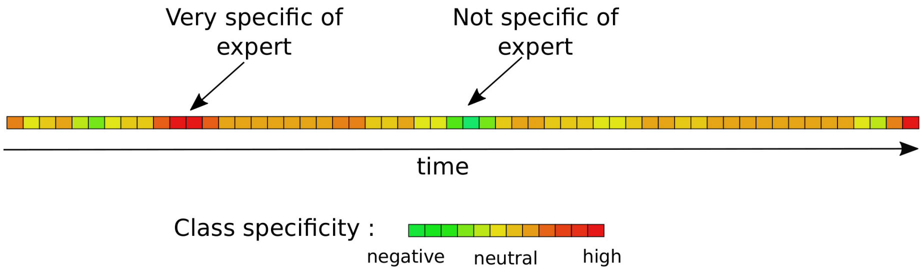

Figure 7: Heat map representation of an expert sequence of right hand activities highlighting the specificity of sub-sequences using *tf∗idf* weights vector.

also common with intermediate surgeons. In this example, the most specific activities (in red) are identified in the dissection phase (*e.g.*, `dissect fascia` using a `dissectors`).

4. Discussion

In experiments #1, #2 and #3 (Table 2), the proposed method (VSM) provides the best results obtaining a perfect classification of 100%. This approach outperformed the state-of-the-art method of 1-NN using DTW as distance measure (except for #3 where they both reached 100%). These results can be explained by the ability of the method to identify patterns that are specific to the location where the surgeries took place (experiment #1) and specific to surgeon experience (experiment #2). The method using 1-NN Euclidean distance is far behind as it does not take into account the temporal distortion.

For experiment #1, the 10 more specific patterns for the different location sites A, B and C are provided respectively in Table 4, 5 and 6. In these tables, one can observe that each location has its own specific patterns. Site B seems to have very specific behaviors compared to Site A and C as the values of *tf∗idf* weight vectors in Table 5 are very high compared to Tables 4 and 5. It means that in Site C, some patterns are very frequent, and absent from Site A and B. Each surgery department has its own way of performing the surgeries.

Furthermore, national recommendations, the context of the operating rooms or the medical equipment that is available, can also influence the way a surgery is performed in a specific location.

In experiments #2 and #3, the goal was to identify patterns that are specific to a group of surgeons of a specific location and having different experience levels (expert vs. intermediate). The experiment #2 focused on surgeons of site A. We presented the top 10 patterns (Table 7 and 8) extracted from this experiment to an expert surgeon for interpretation. Using his expertise, we identified that the discriminative patterns for the two groups belonged to very distinctive phases of the surgery. For the expert, most of the patterns were identified during the discectomy phase (multiple actions using `dissect`), which is a very technical phase where the technical expertise really makes a difference. For the intermediates, the patterns concerned mostly the dissection of the surgical approach, which is also a very technical phase. The patterns from intermediate surgeons also contain more repetitive activities (*e.g.*, *abab*) which is typical when a surgeon learns how to perform a specific type of surgery.

In experiments #4 and #5, we put together the expert and intermediate surgeries of sites A and C (containing expert and intermediate surgeons) and B (containing only expert surgeons). In this configuration, our method provides accuracies that are lower than 1-NN DTW but higher than 1-NN Euclidean. The difficulty to identify specific patterns in this scenario can be explained by the heterogeneity within the two groups of "expert" and "intermediate" surgeries as they come from two different locations. It means that the method was not able to identify patterns that are common to all experts and all intermediates of sites A and C. This result is in favor of the hypothesis that the way surgical skills are transmitted is dependent of the location.

Finally in the last experiment (#6), we looked for patterns that are specific to a given surgeon (11 surgeons in total). We alternatively took all the surgeries performed by one surgeon, and we compared them to all the other surgeries. The goal was to find the patterns that are specific to one surgeon, a subsequences that is the "signature" of one surgeon. In this experiment, our method provided the best result with an accuracy of 77.5 %. It was possible to find patterns that are specific of one surgeon for 31 surgeries out of 40. Four surgeries of the nine misclassified surgeries were the surgeries from surgeons #4 and #10, from which there were only two surgeries available for each surgeon in the dataset. The method was not able to find distinctive patterns of a surgeon from only two surgeries. For the five remaining misclassifications, they are spread out on surgeons having more surgeries, meaning that a specific pattern in their other performed surgeries was still found. Note that compared to 1-NN DTW, the VSM method does not introduce additional misclassifications and only corrected some DTW classification errors.

These results reveal that our method makes it possible to identify precisely what are the subsequences of activities that are highly frequent in the behavior of senior surgeons and absent from the behavior of junior surgeons. These specific patterns could be used to better understand what “makes” a senior surgeon, and what are the specific pattern a junior has to learn throughout his training.

Our method makes it possible to, for instance, identify which parts of a surgery was performed like a senior surgeon or like a junior surgeon. It can be used as a teaching tool to provide specific feedback showing to junior surgeons the parts of the surgery where they behaved like a senior surgeons and where they behaved like a junior surgeons.

Our work differs from previous efforts in that instead of looking into a single group of surgeries, we focus on the comparative analysis of multiple groups at the same time. The approach proposed in this paper provides more intuitive results, as it identifies automatically the set of patterns explaining the differences between the groups. Note that the proposed approach is not designed to discriminate *good* or *bad* surgical behaviors, as such high-level interpretation requires years of surgical expertise and practice. We only aim at supporting high-level analyses by identifying what are the most specific patterns in a set of surgeries, as compared to another set.

It is interesting to note some drawbacks of our study. The first one is related to the inherent variability of surgeries. While we focused on a standardized procedure with patients having similar background, there is always a surgery- or patient-specific part to each surgery. Even if we didn't notice this to be impacting our results, it is important to keep in mind that our method naturally discards the subsequences that rarely occur. Second, as the size of the dataset used in the experiments is currently limited, the patterns extracted should not be considered uniformly true. A medical study about actual surgical patterns, using a larger corpus, would be important – this paper introduces the method to perform such a study which we hope will be possible as more data is being collected. Finally, it is important to observe that our VSM approach focuses on the sequence of actions, and not how those actions were performed. This should be taken into account when using our approach in a training system.

We believe that our VSM model is a milestone in surgical process analysis and that there are numerous possible applications. For example, it would be possible to correlate specific patterns with specific practical skills, which would directly support the automatic evaluation of surgical skills. Furthermore, the correlation presence/absence of some patterns with after-surgery complication, or readmission could be studied [22, 23]. In this case, a dictionary of good and bad patterns could be built. Finally, this method could also be used as an addition to surgical activities prediction systems [24, 25, 26, 27] by providing frequencies of most frequent subsequences. Our system could also be used to identify the core set of subsequences activities that are performed by all the surgeons regardless of their countries or skill levels. This would however require a larger dataset to be collected.

To conclude this discussion, we list the main contributions of this paper:

1. We introduced the first integration and application of VSM to the field of SPM.
2. We introduced a framework which makes possible to identify the most discriminative patterns of a given set of surgeries.
3. We assessed our framework on real-world data with the task of predicting

the location where the surgery took place, the experience of the surgeon, and which surgeon performed the surgery.

4. We proposed a visualization of the *tf∗idf* weight vectors as a tool that can support teaching programs to highlight interesting parts of a given surgery.

5. Conclusion

In this paper we presented a method that builds upon *tf∗idf* pattern ranking and VSM in order to identify discriminative patterns in surgeries. We showed how this framework can be applied to identify patterns that can then be used to classify according to the location where the surgery took place or the expertise of the surgeon. The method was also able to identify patterns that are characteristic of a single surgeon. We have also shown that the visualization of the top patterns ranked using their *tf∗idf* weights, along with the visualization of their weights on a sequence, could be a useful tool while teaching surgeries. There are multiple ways to extend this work to the identification of patterns correlating with the acquisition of a specific technical skills, or explaining after-surgery complication, or readmission. In future work, we plan to investigate in more depth the correlation between sub-sequences of activities and skills assessment.

Acknowledgements

This work was supported by the Australian Research Council under awards DE170100037. This material is based upon work supported by the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AOARD) under award number FA2386-16-1-4023.

[1] Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AHS, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. New England Journal of Medicine 2009;360(5):491–9.

[2] Stitzenberg K, Sheldon G. Progressive specialization within general surgery: adding to the complexity of workforce planning. Journal of the American College of Surgeons 2005;201(6):925–32.

[3] Hutter M, Kellogg K, Ferguson C, Abbott W, Warshaw A. The impact of the 80-hour resident workweek on surgical residents and attending surgeons. Annals of surgery 2006;243(6):864.

[4] McPherson K, Bunker J. Costs, risks and benefits of surgery: a milestone in the development of health services research. Journal of the Royal Society of Medicine 2007;100(8):387–90.

[5] Schuhmann T. Hospital financial performance trends to watch. Healthcare Financial Management 2008;62(7):59–66.

Actions	
cu	cut
se	sew
co	coagulate
in	install
di	dissect
ir	irrigate
dr	drill
re	remove
ho	hold

Anatomical structures	
mu	muscle
ve	vertebra
sk	skin
fa	fascia
di	disc
li	ligament

Instruments	
ne	needle-holders
ro	rongeurs
cl	classic-cottonoids-forceps
fl	fluoroscopy
di	dissectors
fo	forceps-monopolar
su	suctiontube
sa	salinesolution
re	retractors
cu	curettes
sc	scalpel
ho	hooks
ar	arthrodesis
co	cottonoids-forceps
cl	classic-drape-forceps
sc	scissors
bi	bipolar-forceps
cu	cup-forceps-forceps

Table 9: Vocabulary used to describe the surgical activities.

[6] Lalys F, Jannin P. Surgical process modelling: a review. International Journal of Computer Assisted Radiology and Surgery 2014;9(3):495–511.

[7] Maier-Hein L, Vedula S, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science: Enabling next-generation surgery. arXiv preprint arXiv:170106482 2017;.

[8] Forestier G, Lalys F, Riffaud L, Trelhu B, Jannin P. Classification of surgical processes using Dynamic Time Warping. Journal of Biomedical Informatics 2012;45(2):255–64.

[9] Forestier G, Lalys F, Riffaud L, Collins DL, Meixensberger J, Wassef SN, et al. Multi-site study of surgical practice in neurosurgery based on surgical process models. Journal of Biomedical Informatics 2013;46(5):822–9.

[10] Forestier G, Petitjean F, Riffaud L, Jannin P. Non-linear temporal scaling of surgical processes. Artificial Intelligence in Medicine 2014;62(3):143–52.

[11] Neumuth T, Loebe F, Jannin P. Similarity metrics for surgical process models. Artificial Intelligence in Medicine 2012;54(1):15–27.

[12] Riffaud L, Neumuth T, Morandi X, Trantakis C, Meixensberger J, Burgert O, et al. Recording of surgical processes: a study comparing senior and junior neurosurgeons during lumbar disc herniation surgery. Neurosurgery 2010;67:325–32.

[13] Mehta N, Haluck R, Frecker M, Snyder A. Sequence and task analysis of instrument use in common laparoscopic procedures. Surgical endoscopy 2002;16(2):280–5.

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

[15] 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.

[16] Manning CD, Raghavan P, Schütze H, et al. Introduction to information retrieval; vol. 1. Cambridge University Press; 2008.

[17] Bettadapura V, Schindler G, Plötz T, Essa I. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2013, p. 2619–26.

[18] Hamid R, Maddi S, Johnson A, Bobick A, Essa I, Isbell C. A novel sequence representation for unsupervised analysis of human activities. Artificial Intelligence 2009;173(14):1221–44.

[19] Neumuth T, Strauß G, Meixensberger J, Lemke H, Burgert O. Acquisition of process descriptions from surgical interventions. In: Database and Expert Systems Applications. 2006, p. 602–11.

[20] Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery 2016;:1–55.

[21] Mitchell TM, et al. Machine learning. wcb. 1997.

[22] Schumann S, Bühligen U, Neumuth T. Outcome quality assessment by surgical process compliance measures in laparoscopic surgery. Artificial Intelligence in Medicine 2015;63(2):85–90.

[23] Huaulmé A, Voros S, Riffaud L, Forestier G, Moreau-Gaudry A, Jannin P. Distinguishing surgical behavior by sequential pattern discovery. Journal of Biomedical Informatics 2017;67:34–41.

[24] Forestier G, Petitjean F, Riffaud L, Jannin P. Automatic matching of surgeries to predict surgeons next actions. Artificial Intelligence in Medicine 2017;.

[25] Forestier G, Petitjean F, Riffaud L, Jannin P. Optimal sub-sequence matching for the automatic prediction of surgical tasks. In: Conference on Artificial Intelligence in Medicine. Springer; 2015, p. 123–32.

[26] Forestier G, Riffaud L, Jannin P. Automatic phase prediction from low-level surgical activities. International Journal of Computer Assisted Radiology and Surgery 2015;10(6):833–41.

[27] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics 2008;77(2):81–97.