Multi-site study of surgical practice in neurosurgery based on surgical process models.

Germain Forestier, Florent Lalys, Laurent Riffaud, D. Louis Collins, Jurgen Meixensberger, Shafik Wassef, Thomas Neumuth, Benoit Goulet, Pierre Jannin

Multi-site study of surgical practice in neurosurgery based on Surgical Process Models

Germain Forestier [1] (1), Florent Lalys (2), Laurent Riffaud (2,4), D. Louis Collins (3), Jurgen Meixensberger (5,6), Shafik N. Wassef (7,8), Thomas Neumuth (5), Benoit Goulet (3), Pierre Jannin (2)

(1) MIPS (EA 2332), University of Haute-Alsace, Mulhouse, France
(2) INSERM MediCIS, Unit U1099 LTSI, University of Rennes 1, Rennes, France
(3) McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada
(4) Department of Neurosurgery, Pontchaillou University Hospital, Rennes, France
(5) Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Germany
(6) Department of Neurosurgery, University Hospital Leipzig, Germany
(7) Department of Radiology, University of Iowa Hospitals and Clinics, Iowa City, IA, USA (8) Department of Neurological Surgery, University of Iowa Hospitals and Clinics, Iowa City, IA, USA

Abstract

Surgical Process Modeling (SPM) was introduced to improve understanding the different parameters that influence the performance of a Surgical Process (SP). Data acquired from SPM methodology is enormous and complex. Several analysis methods based on comparison or classification of Surgical Process Models (SPMs) have previously been proposed. Such methods compare a set of SPMs to highlight specific parameters explaining differences between populations of patients, surgeons or systems. In this study, procedures performed at three different international University hospitals were compared using SPM methodology based on a similarity metric focusing on the sequence of activities occurring during surgery. The proposed approach is based on Dynamic Time Warping (DTW) algorithm combined with a clustering algorithm. SPMs of 41 Anterior Cervical Discectomy (ACD) surgeries were acquired at three Neurosurgical departments; in France, Germany, and Canada. The proposed approach distinguished the different surgical behaviours according to the location where surgery was performed as well as between the categorized surgical experience of individual surgeons. We also propose the use of Multidimensional Scaling to induce a new space of representation of the sequences of activities. The approach was compared to a time-based approach (e.g. duration of surgeries) and has been shown to be more precise. We also discuss the integration of other criteria in order to better understand what influences the way the surgeries are performed. This first multi-site study represents an important step towards the creation of robust analysis tools for processing SPMs. It opens new perspectives for the assessment of surgical approaches, tools or systems as well as objective assessment and comparison of surgeon's expertise.

Keywords: Surgical Process Models, Surgery, Surgical skills, Clustering

1. Introduction

The concept of decomposing a surgical process into a sequence of tasks was first presented by MacKenzie et al. [1] and Jannin et al. [2] who introduced the concept of Surgical Process Modelling (SPM). SPM allows description of a surgical intervention using a formal and structured language to model a Surgical Process (SP). Thus, SPMs represent SPs which are formalized as symbolic structured descriptions of surgical interventions using a pre-defined level of granularity and a dedicated terminology [3, 4].

The development of SPM involves three major processes: modelling, acquisition and analysis [4]. The modelling describes the work-domain of the study and its formalism, *i.e.* what is studied and what is modelled. The level of granularity is defined according to the level of abstraction for describing a surgical procedure. The acquisition describes the collection of data on which the models are built, this step being performed by human observations [2, 3, 4] or sensor systems [5]. The analysis process links acquired

---

[1]Corresponding Author: Germain Forestier `germain.forestier@uha.fr`, ENSISA - Universite de Haute Alsace, 12, rue des freres Lumiere, 68093 Mulhouse FRANCE, tel:+33(0)3 89 3369 63, fax:+33 (0)3 89 42 32 82

data to the studied modelled information. Analysis methods can be divided into three types: methods to create an individual model (iSPM), methods that aggregate/fuse information, and the methods that classify/compare data for extracting a specific parameter.

The methods that help creating individual models are characterized by the levels of granularity of the acquired information and of the modelling. Top-down approaches are described as analyses that go from a global overview of the intervention with patient-specific information and a description of high-level tasks (such as phases or steps) to fine-coarse details (such as activities or motions). On the contrary, a bottom-up approach takes as input low-level information from sensor devices and tries to extract semantic high-level information. From the large number of papers published in this category, input data coming from videos [6, 7, 8, 9, 10] or tracking systems [11, 12] have been of increased attention.

The goal of aggregation/fusion methods is to create a global model (gSPM) of a specific procedure by merging a set of SPMs. One approach is to merge similar paths as well as to filter infrequent ones to create average SPMs [13]. This may provide a global overview of the surgical practice. Another approach is to create gSPMs that represent all possible transitions within SPMs. A step of synchronization may be necessary for both approaches in order to be able to merge all SPMs. For such purpose, probabilistic analysis have been used [7].

Finally, the principle of comparison/classification methods is to use SPMs to highlight a specific parameter (*i.e.* meta-information) that explains differences between populations of patients, surgeons or systems. Two main applications have been considered: comparison of surgical tools/approaches/systems and objective evaluation of surgical skills. For both, different approaches have been employed. For quantitatively describing the similarities among multiple SPMs, similarity metrics were developed.

Time was the first information chosen to evaluate surgical systems, tools, approaches or assess surgeons skills [14, 15]. Many clinical studies adopted the principle of time-motion analysis in the early 90s using off-line observer-based videos recording (installed in the OR, surgeons' head mounted, or in the operating field) [16]. Information regarding phases/steps/activities was then processed through simple statistical analysis such as average, number of occurrence or standard deviation [17, 18, 19, 20, 21, 22, 23]. The principle of time-motion analysis was later used by Riffaud et al. [24] but with on-line (*i.e.* live) SPM acquisition to compare expertise of surgeons. Different metrics were used: the operating time for the whole procedure and for each step, the number of activities performed with either the right or the left hand, the number of changes in microscope position, and the number of gestures performed by the surgeon (instruments used and anatomical structure treated). Furthermore, a set of similarity metrics has been recently proposed by Neumuth et al. [25] to compare different SPMs. In particular, the similarity of granularity, the content similarity, the temporal similarity, the transitional similarity and the transition frequency similarity were defined, each of them representing different aspects of SPMs. Classification focusing on the sequential aspect of SPMs was studied by Forestier et al. [26], where Dynamic Time Warping (DTW) along with K-Nearest Neighbour (KNN) algorithm were used for evaluating surgical skills over a population of surgeons. This method focuses on the different types of activities performed during surgery and their sequencing, by minimizing time differences. For example, if two sequences are composed of the same set of actions in the same order, they will be considered as identical even if they do not last the same amount of time. This approach turned out to be a complementary approach to the classical methods that only focus on differences in the time and the number of activities.

In this study, the surgical practice at three different institutions is studied with SPM methodology based on a comparison/classification analysis method, using on-line observer-based recordings of surgical processes, modelled by SPMs. For this study, we followed the methodology described in [26]. Additionally, a matching process was introduced to make the link between terminologies. It allowed comparing SPMs acquired at different sites. 41 surgeries of anterior cervical discectomy (ACD) SPMs were acquired at the Neurosurgery departments of the Rennes University Hospital (France), the Leipzig University Hospital (Germany), and the Montreal Neurological Institute University Hospital (Canada). SPMs performed at different sites were classified using a similarity metric based on sequencing to 1) distinguish the different surgical behaviours according to the location where surgery was performed, and 2) establish a detailed classification of SPMs according to the level of surgical expertise of the surgeon performing the surgical procedure. Neurosurgery is among the riskiest and most important surgeries that is performed today. The complexities involved in the OR on the human brain mean that the initial training of a neurosurgeon requires extensive one-on-one instruction from a senior neurosurgeon. After that initial training, neurosurgeons still require several further years of experience to themselves reach a senior level. Consequently, comparing the way surgery is performed throughout a population of surgeons in several location increases the understanding of the complexity of the field of surgery. Thus, the main

goal of this paper is to present how a proposed metric can be used to compare SPMs in order to create groups of similar surgical behaviours that can be explained by external parameters, in this paper the location and the expertise of the surgeons.

We also propose the use of Multidimensional Scaling (MDS) to induce a new space of representation of the sequences of activities. Indeed, the similarity computed using DTW is complex and does not allow to easily display the surgeries for visual assessment. Using MDS allows us to plunge the surgeries on 2D Euclidean space, allowing to easily assess the similarity between them.

2. Methods

2.1. Surgical Process Model (SPM) as sequence of activities

A Surgical Process Model (SPM) can be seen in the real world as a sequence of flow objects [27]. According to the Workflow Management Coalition (WFMC) terminology [28], flow objects representing surgical work steps were named as activities $\mathbf{ac}_i$ and a set of activities as $\mathcal{AC}$ with $\mathbf{ac}_i \in \mathcal{AC}$ ($\mathbf{ac}_i$ being the ith activity). Each activity in a SPM corresponds to a surgical work step, which contains several kinds of information. Thus, an activity $\mathbf{ac}_i$ is defined as a triple :

$$\mathbf{ac}_i = <\mathbf{a}; \mathbf{s}; \mathbf{i}> \qquad \mathbf{a} \in \mathcal{A}, \ \mathbf{s} \in \mathcal{S}, \ \mathbf{i} \in \mathcal{I}^{m_i} \tag{1}$$

with $\mathcal{A}$ the set of possible actions (*e.g.* {cut, remove, ...}), $\mathcal{S}$ the set of possible anatomical structures (*e.g.* {skin, dura matter, ...}), $\mathcal{I}$ the set of possible instruments (*e.g.* {scalpel, scissors, ...}) and m_i the number of instruments used in the activity $\mathbf{ac}_i$. An example of one complete activity could be: <cut, skin, scalpel>. Thus, the domain of definition of an activity is given by: $\mathcal{A} \times \mathcal{S} \times \mathcal{I}^{m_i}$. These sets of possible values are generally specific to the type of studied surgery. Let $\mathcal{T} = \{\mathcal{A}, \mathcal{S}, \mathcal{I}\}$ be the terminology used to describe a specific set of SPMs. We address the problem of heterogeneity among these sets on data acquired on different sites, in the next section. Indeed, each site has generally its own terminology $\mathcal{T}$. An ontology can be used to describe the vocabulary for a specific type of surgery [2, 4, 29].

Along with the information of the action ($\mathbf{a}$), the anatomical structure ($\mathbf{s}$) and the used instrument-s ($\mathbf{i}$), each activity has a starting point ($start(\mathbf{ac}_i)$) and a stopping point ($stop(\mathbf{ac}_i)$) which respectively correspond to the time point when the activity started and the time point when the activity stopped ($start(.) \rightarrow \mathbb{R}, stop(.) \rightarrow \mathbb{R}$) on the timeline of the surgeries. Note that $start(\mathbf{ac}_i) < stop(\mathbf{ac}_i)$, induces a partial order among the activities. The last information carried on the activity is the hand used to perform the activity ($hand(\mathbf{ac}_i)$) which can either be right or left.

A Surgical Process Model can be seen as a sequence of activities ($\mathbf{sp}_k$) performed during surgery. Each activity of this sequence belongs to the set of all the different activities performed during surgery ($\mathcal{AC}_k$) :

$$\mathbf{spm}_k = <\mathbf{ac}_1^{(k)}, \mathbf{ac}_2^{(k)}, \ldots, \mathbf{ac}_{n_k}^{(k)}> \ | \ \mathbf{ac}_i^{(k)} \in \mathcal{AC}_k \tag{2}$$

We proposed in our previous work [26] to use the Dynamic Time Warping (DTW) algorithm [30] to compare SPMs. DTW is based on the Levenshtein distance (or edit distance), and was originally used for applications in speech recognition. It finds the optimal alignment between two sequences, and captures flexible similarities by aligning the two sequences. In order to use DTW to compare two sequences, a distance was defined to evaluate the similarity between the different elements composing the sequence. This approach allows us to compare surgeries according to the activities performed and their sequencing in the timeline. Note that the cost of the alignment can be seen as a dissimilarity measure but is not a distance as DTW is a semi-pseudometrics. The term distance is used here as an abuse of language.

2.2. Dealing with terms heterogeneity

One of the main problems, when comparing data acquired at different contexts (*e.g.* different sites), is the heterogeneity within the data. There are several sources of heterogeneity, which lead to bias in the data acquisition step, such as the expertise of the surgeon performing the acquisition (named the *operator*), the error in the acquisition, or the precision of the data. This bias are heavily reduced by the use of a common software for the acquisition. Furthermore, recent work [31] showed that the bias due to the operator is limited. However, another source of heterogeneity is the use of a different terminology to describe the activities performed during the surgeries. Indeed, depending of the parameters of the SPM acquisition software, the operator can use different terminologies (*i.e.*, list of words describing action, anatomical structures and instruments). In this study, to compare surgeries performed at three

different sites (Rennes, Leipzig and Montreal), the terminologies used in the different sites were checked for differences and similarities. Since the approach used [26] is based on binary comparison of the components of the activities (action, anatomical structure and instrument), even a slight difference in the used terms leads to different evaluation of the similarity. For example, the terms `scalpel` and `surgicalKnife` would be considered as different, even if they share the same meaning. Furthermore, even the terms `scalpel` and `myScalpel` would be considered as different. Consequently, if the used terminology is different according to the sites, the comparison is meaningless.

To solve this problem, one solution is to use an ontology as reference. An ontology is defined as an explicit formal specification of a shared conceptualization [32]. According to different level of explicitness, an ontology can be a full description of a domain using complex axioms and taxonomy [33], or as a simple catalog of normalized terms composing a vocabulary [34, 35]. The knowledge stored in an ontology can be used to solve disambiguation [36] as the synonyms of different words can be represented. However, even if some well established resources exist in specific domains (*e.g.* anatomy with the FMA [33]), they are not easily applied for surgical instruments and surgical actions. Indeed, some work has been carried out to use ontological engineering [37, 38] to formalized surgical knowledge, but no recent initiatives exist.

In order to evaluate the heterogeneity of the terms used in Rennes, Leipzig, and Montreal, the terminologies used at each site were compared. Each location was anonymously given a letter, *i.e* Site A, Site B and Site C without providing the identifying key. Thus, the set of terms used in the three sites are $\mathcal{T}_A$, $\mathcal{T}_B$ and $\mathcal{T}_C$. The results of this comparison showed that the terminologies used in the recording of site A and site C were highly similar, more than 90% ($\mathcal{T}_A \cap \mathcal{T}_C$) of the words used for the actions ($\mathcal{A}$), anatomical structure ($\mathcal{S}$) and instruments ($\mathcal{I}$) were similar. However, the terminology used for site B ($\mathcal{T}_B$) was very different with less than 50% of similarity with sites A and C.

Consequently, the terms used in sites A and C were manually matched with the terms used in site B by an expert surgeon. This matching contains simple correspondence (`suctiontip` $\rightsquigarrow$ `suctiontube`) to more complex ones (`tie` $\rightsquigarrow$ `sew`). Using this knowledge, a function Φ which converts the terminology used in site B to the terminology used in sites A/C was defined. This function was applied to the SPMs acquired in site B leading to 9282 transformations (*i.e.* switch from one term to another). These transformations allowed to fairly compare the SPMs. The Φ function was applied to site B data before performing binary comparison of the activities between SPMs performed in site B and sites A/C which reduces heavily the bias due to the use of different terminologies.

2.3. Analysis using Hierarchical Clustering

Clustering [39] is the automatic assignment of a set of objects into subsets (called clusters) so that objects in the same cluster are similar to some extent. This approach was applied to automatically create clusters of similar surgeries. DTW is the similarity measure used to compare the SPMs [26]. This approach allows comparing surgeries according to the different activities performed by the surgeon and their sequencing in the surgery timeline.

Hierarchical clustering is a method of cluster analysis, which seeks at building a hierarchy of clusters. Starting with the objects, the clusters are created iteratively by merging the two most similar clusters. Different criteria exist to choose the clusters to merge. The average-link approach [40] was used, consisting in evaluating the similarity of two clusters according to the average distance between all couple of objects in the two clusters. Thus, the distance between two clusters C_i and C_j composed of SPMs, is defined as:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{k=1}^{|C_i|} \sum_{l=1}^{|C_j|} d(\mathbf{spm}_k, \mathbf{spm}_l) \tag{3}$$

where $|C|$ is the cardinality of the cluster (*i.e.* the number of SPMs in the cluster). Hierarchical clustering approaches are known to be computationally expensive. However, as the number of data we manipulated was limited, using this kind of approach was tractable (*e.g.* less than 10 seconds of computation time for one clustering of the data, a few minutes to compute the distance matrix). The average-link approach was selected for its low sensibility to noise and outliers.

A dendrogram, which is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering, was a useful tool to carry out a multi-level study. Indeed, by cutting the dendrogram at different levels, the clustering results can be analysed in details and can exhibit different patterns across the cuts.

4

Figure 1: Illustration of the on-line recording of the data in the operating room.

2.4. Data

Experiments were performed on one-level anterior cervical discectomy (ACD) surgeries. 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 1-level ACD surgical procedure can be decomposed into four major phases, whereas a fifth one may be necessary. These four phases are: the approach, the discectomy, the arthrodesis, and the closure phases. An additional phase of hemostasis may be mandatory in certain cases. The figure 4 presents an index-plot [26] representing the activities performed by the surgeon using the right hand for one surgery. It also presents the different phases of the surgery. Fourty-one surgeries were recorded on-line using the Surgical workflow Editor [41] resulting in the creation of 41 XML files containing the sequence of activities of each surgery. The figure 1 illustrates the recording of the data in the OR. 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 41 surgeries, 11 were performed at site A, 18 were performed at site C, and 12 at site B. According to level of expertise 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. 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.

3. Results

3.1. Dendrogram analysis

The 41 surgeries composing our dataset (section 2.4) were processed using hierarchical clustering (section 2.3) using the Matlab software. The figure 3 presents the dendrogram for the clustering of the surgeries. The x labels indicates the location of the acquisition (A; site A, B: site B, C: site C), the index of the surgeon (1 to 11) and its level of expertise (E: Expert, I: Intermediate). The table 1 presents the information for each surgeon involved in the study and the table 2 the information on the patients.

At the first level of the study, the dendrogram can be divided to create three clusters C_1, C_2 and C_3 (highlighted in blue, green and red on the figure). One can observe that different surgical behaviours can be identified according to the location where surgery was performed. Indeed, the blue cluster (C_1) contains 95% of the surgeries performed in site C, the green cluster (C_2) contains 100% of the surgeries performed in site A, and the red cluster (C_3) contains 100% of the surgeries performed in site B. This first result showed differences in this same surgery performed at the three sites. Furthermore, the size of the link between clusters in the dendrogram is proportional to the distance between the clusters, which suggests that the surgical behaviour of site C and site A in the dataset are more similar in behaviour than site B.
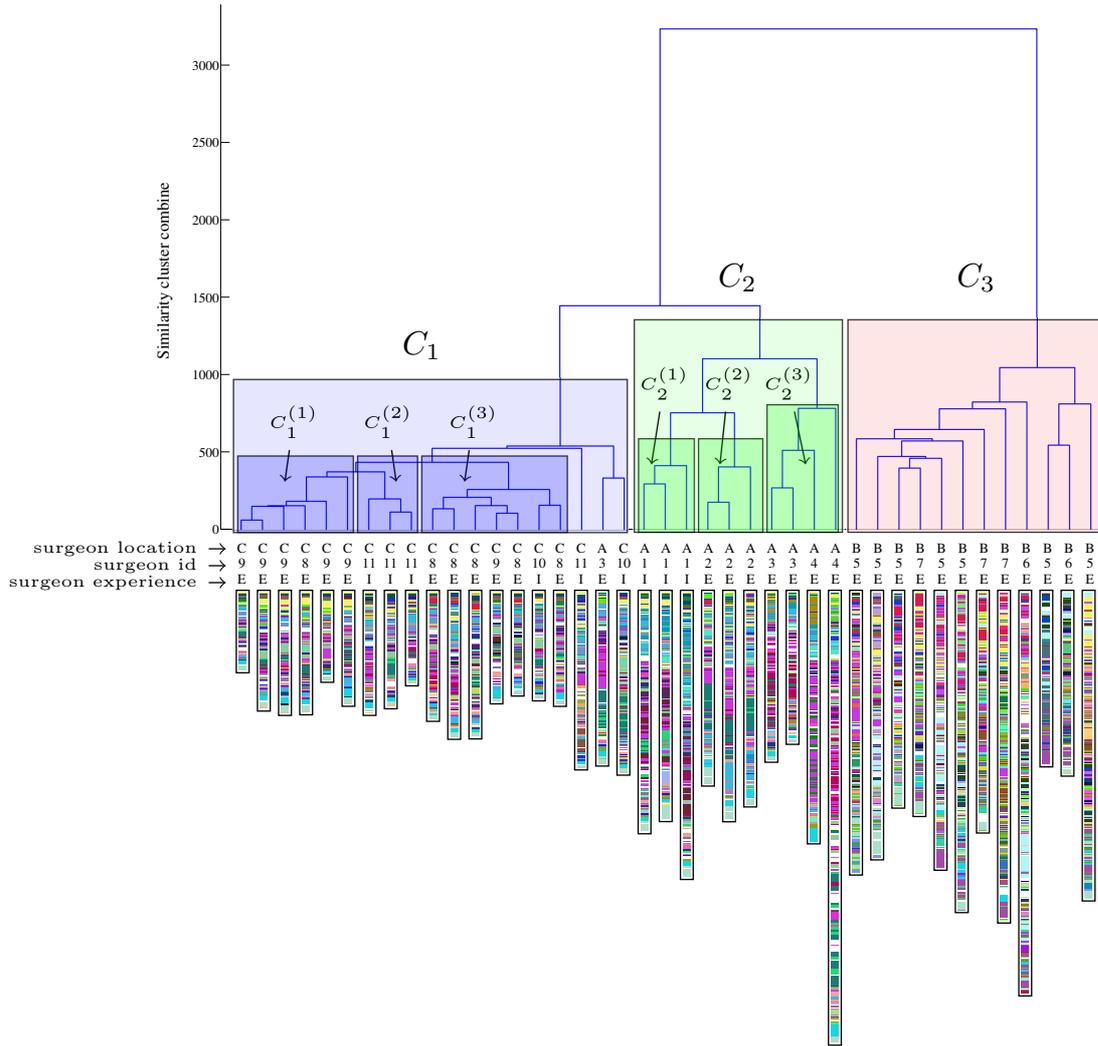
5

Figure 2: Dendrogram representing the hierarchical clustering of the sequence of activities performed during 41 surgeries. For each surgery, the site (A,B,C), the surgeon id (1-11) and the level of expertise (Expert (E), Intermediate (I)) is mentioned on the top of the sequence of activities.

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

Table 1: List of the surgeons involved in the study.

Patient ID	Sex	Age	Patient ID	Sex	Age
1	F	37	2	-	-
3	F	54	4	M	47
5	M	32	6	F	-
7	M	54	8	M	43
9	F	35	10	F	38
11	M	51	12	M	36
13	F	76	14	F	34
15	F	47	16	M	51
17	F	81	18	F	50
19	F	73	20	M	66
21	F	70	22	M	66
23	M	66	24	M	55
25	M	48	26	M	50
27	M	37	28	M	58
29	M	53	30	F	53
31	F	48	32	-	-
33	-	-	34	F	37
35	F	60	36	F	41
37	M	46	38	-	46
39	-	-	40	M	60
41	F	56			

Table 2: List of the patients involved in the study with sex and age. Missing values are represented by a "-" sign.

At a second level of the study, three sub-clusters were identified within the blue cluster (C_1) : $C_1^{(1)}$, $C_1^{(2)}$ and $C_1^{(3)}$. The first one ($C_1^{(1)}$) contains 6 expert surgeons (100%). The second one ($C_1^{(2)}$) contains 3 intermediate surgeons (100%). The third one ($C_1^{(3)}$) contains 6 expert surgeons (86%) and one intermediate surgeon. The remaining three surgeries being mixed up. This result shown that our approach was able to identify different surgical behaviours between expert and intermediate surgeons. Indeed, the surgeries performed by expert surgeons seem more similar to each other than surgeries performed by intermediate as they are clustered together. This can be explained by the experience gained during the formation and the career of a surgeon. Furthermore, if we go even further by observing how surgeries from the same surgeon were clustered, we can observe that most of the time, they are clustered together. For example, the cluster $C_1^{(1)}$ of six experts is composed of five surgeries out of six (83%) performed by the 9th surgeon. In the cluster $C_1^{(2)}$, 100% of surgeries were performed by the 11th surgeon. And in the cluster $C_1^{(3)}$ five out of seven surgeries (71%) were performed by the 8th surgeon. These results highlighted that each surgeon had his own behaviour, and that our approach was able to identify them by clustering together surgeries performed by the same surgeon. An interesting fact to notice is that the cluster $C_1^{(1)}$ containing expert surgeons and the cluster $C_1^{(2)}$ containing intermediate surgeons were merged together in the hierarchy before merging with the cluster $C_1^{(3)}$ containing mainly expert surgeons. It means that the behaviour of intermediate surgeon 11th is closer in the way he operated to a certain group of experts than the behaviour of the two groups of experts. It can be explained by the fact that the intermediate surgeon 11th present in the cluster $C_1^{(2)}$ was trained by the expert surgeon 9th present in $C_1^{(2)}$. The approach used in this study was consequently able to identify similarity in the behaviours of a surgeon and its supervisor, explained in the transposition of surgical skills.

This second-level analysis can also be performed within cluster C_2 (in green in the figure 3). Three sub-clusters can be identified $C_2^{(1)}$, $C_2^{(2)}$, $C_2^{(3)}$. The first one ($C_2^{(1)}$) contains 100% of expert surgeons, the second ($C_2^{(2)}$) and third one ($C_2^{(3)}$) contains both 100% of expert surgeons. Once again, surgeries performed by the same surgeon are clustered together (*i.e.* all surgeries of $C_2^{(1)}$ were performed by the 1st surgeon, all surgeries of $C_2^{(2)}$ were performed by the 2nd surgeon and all surgeries of $C_2^{(3)}$ were performed by the 4th surgeon).

This second-level analysis is less conclusive in the C_3 as no clear sub-clusters emerged. This can be
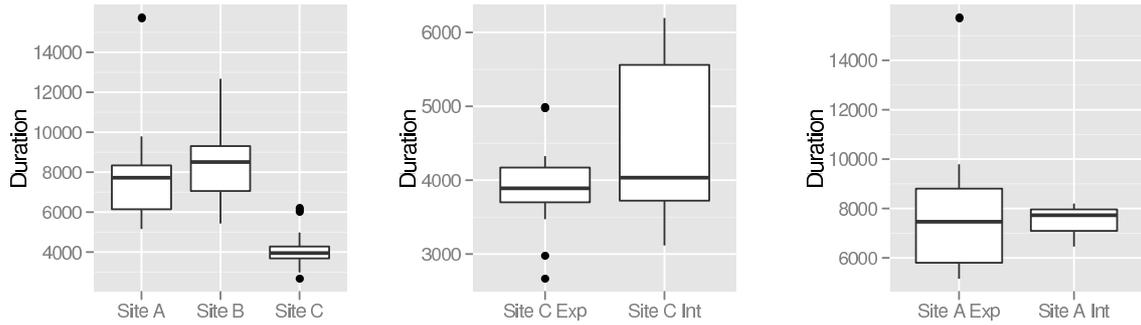
(a) Mean duration distribution in site A,B,C

(b) Mean duration distributions in site C between Expert (Exp) and Intermediate (Int)

(c) Mean duration distributions in site A between Expert (Exp) and Intermediate (Int)

Figure 3: Mean duration of the surgeries in the three sites (a) and between expert (Exp) and intermediate (Int) surgeons in site C (b) and A (c).
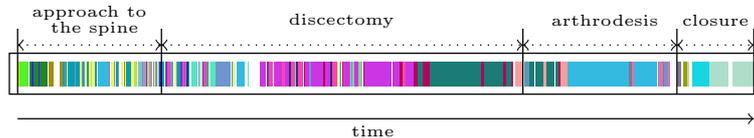

Figure 4: Example of one sequence used in this study. Each color corresponds to a different activity. The different phases of the surgery are displayed above the sequence. The phases are determined by the operator during the acquisition of the data.

attributed to the comparable level of expertise of all the surgeons involved at site B. One other reason why surgeries from the same surgeon were clustered together, could be due to the complexity of the data, as the SPMs recorded at site B were longer, and were consequently prone to error in comparison. While these errors were not of great impact on analysing the dataset at a coarse level (multi-site), they can have much weight in identifying finer grain differences between surgeries performed locally in site B.

3.2. Multidimensional Scaling

The approach proposed in this paper allows us to compute a similarity measure between sequences of activities performed during surgery. Thus, we are able to compute a $N \times N$ similarity matrix representing the similarity of N given surgeries according to each others. This similarity matrix was used in the previous section to perform a clustering of the surgeries using hierarchical clustering. However, it is often convenient to have a way to display the data in low dimension space in order to have a clear and simple grasp of the distribution of the data objects. The similarity provided using DTW induces a complex space of representation as it is based on a warping of the time scale.

In order to find a simpler space of representation, we propose to use Multidimensional Scaling (MDS) [42] to display the sequences in a 2D Euclidean space. Multidimensional Scaling is a set of statistical tools which takes as input an item-item matrix of similarity and provides as output a location of each item in a M-dimensional space (M being chosen as parameter). The basic idea is to optimize the locations of the items in the new space so that they respect the best the constraints represented by the similarity matrix.

In this work, we used non-metric mutidimensioanl scaling [43] to find a non-parametric monotonic relationship between the dissimilarities, as DTW is semi-pseudometric and not a distance. The Figure 5 displays on two dimensions the results of the application of MDS on the similarity matrix computed on our data (we used the R package isoMDS). Each point represents one sequence of activities of a surgery. The colors correspond to the different sites where the surgeries were performed. Even if reducing the complexity of a sequence of activities to single point is challenging, some observations can still be made
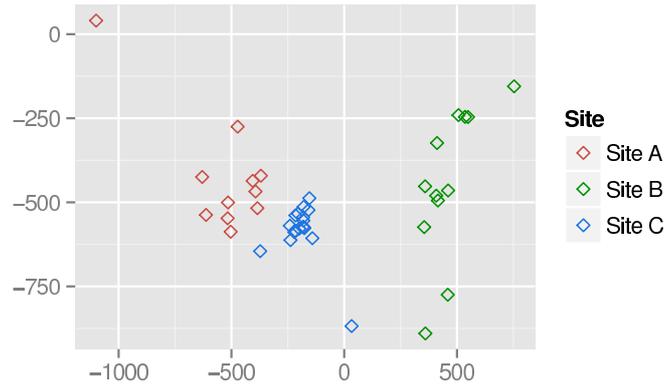
8

Figure 5: Results of the Multidimensional Scaling on 2 Dimension using the similarity matrix of the surgeries.

from this display. For example, one can observe that points (i.e surgeries) from the same location are close to each-other. One can also see that the points of Site A and the points of Site C are closer to each other than the points of site B. This observation backs up the results obtained from the clustering results (see Section 3.1). Furthermore, the points of site A and C seems more compact than site B, results also observed on the clustering result.

This display is a way to easily represent the information and to observe a set of surgeries according to the similarity to each others. An important point to notice is that the coordinates of each surgery is computed according to the similarities to all the other surgeries. Consequently, these coordinates are relative values and not absolute values. If we select one of the surgery and put it within another set of surgeries, its coordinates would have been different. Finally, it would also be possible to apply data mining approaches in this newly created data space instead of using the similarity matrix. However, as this visualization can be seen as a heavy features reduction, it does not grasp the whole complexity of the sequences.

4. Discussion

4.1. Duration of surgeries

The approach used in this paper focused on comparing surgeries based on the different actions performed by the surgeon during surgery and their sequencing. By using Dynamic Time Warping (DTW), we reduced the importance of duration. If two surgeries were composed of exactly the same activities in the same order, DTW disregards the fact that they might not have last the same amount of time. This positioning was made since there is not always a correlation between surgical behaviours and duration of the surgeries. Indeed, several factors can be taken into account, like the complexity of the disease, the extent of the disease, the demographic characteristic of the patient, and so on. Considering such factor, the importance of time was reduced, and more focus was given to the actions performed by the surgeons. However, this could be counter intuitive. For example, in figure 3, one can observe on the bottom of the figure, the sequence of activities performed by the surgeon with their right hand; each colour corresponds to one activity, the height of the index-plot being proportional to the total duration of the surgery. From this figure, one can see that surgeries recorded in site B last much longer than surgeries performed in site A and site C. It could therefore be tempting to base the analysis on the total duration of the surgery. The figure 3 (a) presents box-plots representing the distribution of the mean duration of surgeries according to each site. As foreseen from the figure 3, there are differences in total duration between the three sites. The durations were dramatically shorter in site C, while they were much longer in sites A and B. ($p = 0.709$).

In a finer grain comparison between time duration of expert and intermediate surgeons, figure 3 (b) presents the distributions of the mean duration of the surgery between expert and intermediate surgeons in site C, and figure 3 (c) demonstrates the same analysis for site A. These figures highlight the difficulty to discriminate between these two groups based on the mean duration of surgeries only. The expert

surgeons at site C performed surgeries at a shorter duration of time than the intermediate surgeons, but this difference was not statistically significant ($p = 0.326$). On the other hand, in site A, the intermediate surgeons performed surgeries at a shorter duration of time than the expert surgeons, once again this result was not statistically significant ($p = 0.587$). Again, the duration is not always an accurate measure of skill, as complex cases are often given to experienced surgeons. Duration of surgery can be affected by intra-surgeon factor like stress level and skill, and extra-surgeon level like the complexity of the case, the level of experience and skill of supporting staff, and availability of resources during surgery. Furthermore, as stated in [44] :*"While fast behavior in experienced individuals is afforded by skill, fast behavior in novices is likely instigated by high stress levels, at the expense of accuracy. Humans avoid adjusting speed to skill and rather grow their skill to a predetermined speed level, likely defined by neurophysiological latency"*

These results confirm that using only duration of the surgery is not sufficient to analyse and identify surgical behaviours, and stress on the importance of identifying activity sequencing and pattern analysis.

4.2. Evaluation of behaviours across site and expertise

The approach used in this study allows the classification of SPMs based both on the sites where surgery was performed and on the surgeon's expertise. Such methods may be advantageous for the two applications that have been considered and that have been previously introduced: comparison of surgical tools/approaches/systems and objective evaluation of surgical skills.

Comparisons of tool used, surgical approaches or systems using SPM methodology, allow for quantitative validation and assessment of their impact on a surgical procedure. Current studies conducted within the OR still need new tools for robust, efficient and objective assessment of SPs. At a first level of our study, surgical behaviours could be classified according to different site locations. This could help the integration of new computer-assisted-surgical systems.

Then, the objective surgical skills evaluation could also be considered. At a second level of our study the surgeon's expertise could be recognized, opening perspectives for the automatic assessment of surgeons. As these tasks remain very time-consuming and, to some extent, subjective, the idea of using this approach for skills evaluation would be to automate data acquisition process using different sensors, and then automatically process the SPMs, for example by comparing the current SPM with a training data-set of SPMs. New approaches have been proposed in the literature for automatic recognition of low-level tasks (*i.e.* activities) from videos that can be combined with this work for automating both the acquisition and the analysis processes [5, 10, 45, 46]. As stressed in [47] the use of human examiners in the evaluation process, as for example for the OSATS (Objective Structured Assessment of Technical Skills) can introduce an important bias in the evaluation. Recording the activities of the surgeon and relying only on this information for relative comparison between behaviours is one of the keys of objective surgical skill evaluation.

4.3. Study limitations

The proposed study suffers from some limitations. First, it relies strongly on the quality of the data acquisition step. Indeed, acquiring the data is currently a tedious process as it involves that an operator has to be present in the OR during the surgery. Relying on human acquisition is currently the only way to dispose of a high-level description of the surgery. This manual acquisition can introduce errors in the data. However, experience showed that the amount of error was limited. One way to cope with this problem would be to use sensors or videos to capture the activities of the surgeon. However, automatic identification of the activities is currently limited due to the complexity of the information to analyze.

Second, the proposed method assesses to what extend two surgeries are similar but tools explaining more precisely these differences are currently missing. New tools have to be developed in order to identify and describe the differences to eventually understand and explain them.

Finally, a finer analysis could be conducted with the introduction of other criteria. Only two criteria were mainly considered in this analysis, *i.e.* the surgical site and the surgeons' expertise, but a multitude of parameters from the patient or from the surgical intervention could also be correlated. From the patient, the outcome could be considered, as well as age or specific information about the disease. From the intervention, the complexity of surgery could also be considered, or adverse-events could be taken into account. In the end, a large set of parameters could be introduced in the analysis, showing the various possibilities of this type of SPM-based study.

5. Conclusion

We presented in this paper a SPM-based multi-site study. The approach used for comparing surgeries enabled to focus on the sequentiality of the activities performed during the surgeries by disregarding time differences. Experiments conducted on 41 surgeries of ACD performed in three different clinical sites showed that our approach was able to identify different surgical behaviours according to the location where surgery was performed, and also according to the level of expertise of the surgeon. This work is a milestone in identifying and understanding surgical behaviours. It opens new perspectives for SPM-based study, for the assessment of surgical approaches, tools, systems but also for surgical skills evaluation. Toward the creation of the new generation of CAS systems, the use of SPM may therefore prove its efficiency for facilitating surgical decision-making process as well as improving pre-operative human-computer interface and medical safety.

Acknowledgement

References

[1] Mackenzie, C., Ibbotson, J., Cao, C., Lomax, A.. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minimally Invasive Therapy and Allied Technologies 2001;10(3):121--127.

[2] Jannin, P., Raimbault, M., Morandi, X., Seigneuret, E., Gibaud, B.. Design of a neurosurgical procedure model for multimodal image-guided surgery. In: Proceedings of Computer-Assisted Radiology and Surgery; vol. 1230. Elsevier; 2001, p. 102--106.

[3] 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--611.

[4] Jannin, P., Morandi, X.. Surgical models for computer-assisted neurosurgery. NeuroImage 2007;37(3):783--791.

[5] Lalys, F., Bouget, D., Riffaud, L., Jannin, P.. Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. International Journal of Computer Assisted Radiology and Surgery 2012;:1--11.

[6] Lalys, F., Riffaud, L., Bouget, D., Jannin, P.. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. Biomedical Engineering, IEEE Transactions on 2012;(99):1--1.

[7] Padoy, N., Blum, T., Ahmadi, A., Feussner, H., Berger, M., Navab, N.. Statistical modeling and recognition of surgical workflow. Medical Image Analysis 2012;16(3):632--641.

[8] Bhatia, B., Oates, T., Xiao, Y., Hu, P.. Real-time identification of operating room state from video. In: National Conference on Artificial Intelligence; vol. 22. 2007, p. 1761.

[9] Blum, T., Feussner, H., Navab, N.. Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2010, p. 400--407.

[10] Bouarfa, L., Jonker, P., Dankelman, J.. Discovery of high-level tasks in the operating room. Journal of biomedical informatics 2011;44(3):455--462.

[11] James, A., Vieira, D., Lo, B., Darzi, A., Yang, G.. Eye-gaze driven surgical workflow segmentation. International conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2007;:110--117.

[12] Nara, A., Izumi, K., Iseki, H., Suzuki, T., Nambu, K., Sakurai, Y.. Surgical workflow monitoring based on trajectory data mining. New Frontiers in Artificial Intelligence 2011;:283--291.

[13] Neumuth, T., Jannin, P., Schlomberg, J., Meixensberger, J., Wiedemann, P., Burgert, O.. Analysis of surgical intervention populations using generic surgical process models. International Journal of Computer Assisted Radiology and Surgery 2011;6:59--71.

[14] McKenna, J.. The case for motion and time study in surgery. Frank and Lillian Gilbreth: Critical Evaluations in Business and Management 2003;2:279.

[15] Gilbreth, F.B.. Motion study in surgery. 1916.

[16] van Oostveen, C.J., Vermeulen, H., Gouma, D.J., Bakker, P.J., Ubbink, D.T.. Explaining the amount of care needed by hospitalised surgical patients: a prospective time and motion study. BMC health services research 2013;13:42.

[17] Den Boer, K., de Wit, L., Davids, P., Dankelman, J., Gouma, D.. Analysis of the quality and efficiency in learning laparoscopic skills. Surgical Endoscopy 2001;15(5):497--503.

[18] Darzi, A., Mackay, S.. Skills assessment of surgeons. Surgery 2002;131(2):121--124.

[19] 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--285.

[20] Malik, R., White, P., Macewen, C.. Using human reliability analysis to detect surgical error in endoscopic DCR surgery. Clinical Otolaryngology & Allied Sciences 2003;28(5):456--460.

[21] Cao, C., MacKenzie, C., Payandeh, S.. Task and motion analyses in endoscopic surgery. In: Proceedings ASME Dynamic Systems and Control Division. Citeseer; 1996, p. 583--590.

[22] Claus, G., Sjoerdsma, W., Jansen, A., Grimbergen, C.. Quantitative standardised analysis of advanced laparoscopic surgical procedures. Endoscopic surgery and allied technologies 1995;3(4):210.

[23] Ibbotson, J., MacKenzie, C., Cao, C., Lomax, A.. Gaze patterns in laparoscopic surgery. Studies in Health Technology and Informatics 1999;:154--160.

[24] 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--332.

[25] Neumuth, T., Loebe, F., Jannin, P.. Similarity metrics for surgical process models. Artificial Intelligence in Medicine 2011;54:15--27.

[26] Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., Jannin, P.. Classification of surgical processes using dynamic time warping. Journal of Biomedical Informatics 2012;.

[27] White, S.. Introduction to BPMN. IBM Corporation 2004;31.

[28] Zur Muehlen, M.. Organizational management in workflow applications--issues and perspectives. Information Technology and Management 2004;5(3):271--291.

[29] Neumuth, T., Jannin, P., Schlomberg, J., Meixensberger, J., Wiedemann, P., Burgert, O.. Analysis of surgical intervention populations using generic surgical process models. International Journal of Computer Assisted Radiology and Surgery 2010;6:59--71.

[30] Hiroaki, S., Chiba, S.. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 1978;26:43--49.

[31] Neumuth, T., Jannin, P., Strauss, G., Meixensberger, J., Burgert, O.. Validation of knowledge acquisition for surgical process models. Journal of the American Medical Informatics Association 2009;16(1):72--80.

[32] Gruber, T., et al. Toward principles for the design of ontologies used for knowledge sharing. International journal of human computer studies 1995;43(5):907--928.

[33] Golbreich, C., Zhang, S., Bodenreider, O.. The foundational model of anatomy in owl: Experience and perspectives. Web Semantics: Science, Services and Agents on the World Wide Web 2006;4(3):181--195.

[34] Spackman, K., Campbell, K., CÃ, R., et al. Snomed rt: a reference terminology for health care. In: Proceedings of the AMIA annual fall symposium. American Medical Informatics Association; 1997, p. 640.

[35] Bodenreider, O.. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 2004;32(suppl 1):D267--D270.

[36] Navigli, R.. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 2009;41(2):10.

[37] Mori, A., Gangemi, A., Steve, G., Consorti, F., Galeazzi, E.. An ontological analysis of surgical deeds. Artificial Intelligence in Medicine 1997;:361--372.

[38] Bentley, T., Brown, P.. Exploring the ontology of surgical procedures in the read thesaurus. Meth Inform Med 1998;37:420--5.

[39] Jain, A.K.. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 2010;31(8):651--666.

[40] Manning, C., Schutze, H., MITCogNet, . Foundations of statistical natural language processing; vol. 59. MIT Press; 1999.

[41] Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H., Burgert, O.. Acquisition of process descriptions from surgical interventions. In: Database and expert systems applications. Springer; 2006, p. 602--611.

[42] Kruskal, J.. Nonmetric multidimensional scaling: a numerical method. Psychometrika 1964;29(2):115--129.

[43] Kruskal, J.. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1--27.

[44] Pavlidis, I., Tsiamyrtzis, P., Shastri, D., Wesley, A., Zhou, Y., Lindner, P., et al. Fast by nature-how stress patterns define human experience and performance in dexterous tasks. Scientific Reports 2012;2.

[45] Bouarfa, L., Akman, O., Schneider, A., Jonker, P.P., Dankelman, J.. In-vivo real-time tracking of surgical instruments in endoscopic video. Minimally Invasive Therapy & Allied Technologies 2012;21(3):129--134.

[46] Neumuth, T., Meißner, C.. Online recognition of surgical instruments by information fusion. International journal of computer assisted radiology and surgery 2012;7(2):297--304.

[47] Schijven, M., Reznick, R., ten Cate, O., Grantcharov, T., Regehr, G., Satterthwaite, L., et al. Transatlantic comparison of the competence of surgeons at the start of their professional career. British Journal of Surgery 2010;97(3):443--449.