Assessment of surgical skills using Surgical Processes and Dynamic Time Warping

Germain Forestier, Florent Lalys, Laurent Riffaud, Brivael Trelhu, and Pierre Jannin

INSERM / INRIA / CNRS / Univ. Rennes 1, VISAGES U746, Rennes, France

Abstract. Toward the creation of new computer-assisted intervention systems, Surgical Process Models (SPMs) is an emerging concept used for analyzing and assessing surgical interventions. SPMs represent Surgical Process (SPs) which are formalized as symbolic structured descriptions of surgical interventions, using a pre-defined level of granularity and a dedicated terminology. In this context, an important challenge is the creation of new metrics for the comparison and the evaluation of SPs. Thus, correlations between these metrics and pre-operative data allow to classify surgeries and highlight specific information on the surgery itself and on the surgeon, such as his/her level of expertise. In this paper, we explore the automatic classification of a set of SPs based on the Dynamic Time Warping (DTW) algorithm. DTW allows to compute a distance between two SPs that focuses on the different types of activities performed during surgery and their sequencing, by minimizing time differences. Indeed, it turns out to be a complementary approach to classical methods focusing on the time and the number of activities differences only. Experiments were carried out on 24 lumbar disc herniation surgeries to discriminate the level of expertise of surgeons according to prior classification of SPs. Unsupervised classification experiments have shown that this approach was able to automatically identify groups of surgeons according to their level of expertise (senior and junior), and opens many perspectives for the creation of new metrics for surgeries comparison and evaluation.

1 Introduction

In the domain of medical engineering, the analysis and the modeling of surgical procedures has recently emerged. Surgical procedures can be decomposed into four main levels of granularity [10]: the phases, steps, tasks and motions. A surgical intervention can then be described using a pre-defined level of granularity to create a Surgical Process (SP). Then, Surgical Process Models (SPMs), which are progression models of surgical interventions, are used to study, evaluate and analyze surgical activities in the Operating Room (OR). In this field, a recent and important challenge is the design of new methods to compare and group similar SPs in order to identify relevant patterns, that can be correlated with other pre-operative data in order to highlight specific information on the surgery. The main issue of such analysis is the definition of similarity metrics between

This is the author's version of an article published at MICCAI Workshop M2CAI 2011.

SPs that reveal objective and quantitative differences at every granularity level of the surgical procedure. Indeed, SPs from the same intervention type can have a high variability, which can be caused by many parameters such as the different operating techniques, the intrinsic difficulty of the surgical procedure or the surgeon's expertise. Consequently, similarity measures have to be designed to accurately assess the similarity between SPs according to their content (*i.e.* the different activities performed by the surgeon) and their sequencing (*i.e.* the order in which the activities are performed).

In a recent work, Riffaud et al. [21] computed similarity metrics and performed statistical analysis for comparing groups of senior and junior surgeons (*i.e.* experimented and not experimented surgeons). The metrics used were (i) General parameters of the procedure: the operating time for the whole procedure and for each step, (ii) General parameters of the surgeon's activity: the number of activities performed with either the right or the left hand and the number of changes in microscope position, (iii) Specific parameters of the surgeon's activity: all the gestures performed by the surgeon, the instruments used and the anatomical structure treated. Some of these metrics were found statistically significantly different when comparing the junior and the senior groups. Following this work, we introduced in this paper a new approach by exploring the use of the Dynamic Time Warping (DTW) [7] algorithm to evaluate the similarity between SPs. DTW allows to measure the similarity between two sequences which may vary in time or speed. As SPs have been acquired in different environments, they can easily vary in time which makes DTW particularly suitable for comparing SPs. Using DTW to compare SPs allows us to focus on the sequencing of the activities composing the SPs. Indeed, DTW allows to reduce the importance of time variations in the comparison and focuses on the number of activities and their organization in the timeline of the surgery.

Using this similarity metric, we addressed the problem of the automatic clustering of SPs. We focused our evaluation on the correlation between the clustering of SPs and the dedicated level of expertise of surgeons. We present experiments using 24 SPs of lumbar disc herniation surgery whose half were performed by senior surgeons, and half by junior surgeons. Evaluation studies shown that our approach was able to automatically identify these two clusters of surgeons based on the comparison of the SPs using DTW. Furthermore, our approach was also able to go further by identifying sub-clusters of surgeons.

2 Related work

Following the current technological advent within the Operating Room (OR), an important need emerged for tools allowing to assess and evaluate the impact of these new technologies. Within this field, the development of new methods for objective surgical skill evaluation is an important issue. Surgical skills can be assessed based on five factors: knowledge, decision making, technical skills, communication skills and leadership skills. From these five factors, many researches

have been conducted for developing objective methods for technical skill evaluation. A comprehensive review can be found in [19].

One approach for surgical evaluation [1] is to consider the patient's outcome for assessing surgeons. Unfortunately, this metric is highly variable and dependent of the patient's specific characteristics. Additionally, patient outcome is usually a multiple factors criterion requiring long term follow-up. Even if outcome-based metrics are straightforward to use, they are not objective enough and they do not study in detail the differences within the surgical procedure. Another approach uses human grading techniques. The underlying idea is to ask to a senior surgeon to provide an evaluation rating scale using dedicated check-lists during the observation of an intervention. Several scores have been proposed: Objective Structured Assessment of Technical Skills (OSATS) [20], Objective Structured Clinical Examinations (OSCE) [12], or Global Rating Scale (GRS) [4] and have shown good results. However, this method turned out to be very time-consuming and also very observer-dependant. Motion has then been investigated to analyze dexterity, by tracking the surgeon's hand [3], arm [5], or instruments [22], using various and complementary tracking systems [2] or in the context of robotic assistance [6]. These works have focused on motion patterns analysis, using for instance time series analysis of the different motions. The main drawback of such approaches is its low level of granularity which do not give insight into the surgical scenario followed.

The on-line and off-line recordings of surgeries have been of growing interest for analysing procedures and assessing surgeons. Recordings can be performed using sensor devices or directly by an observer. This data extraction process can be supported by fixed protocol created by experts. In this case, the first step consist in building up its own vocabulary. A new terminology is employed and permits a knowledge representation that is proper to the own surgeon's experience and to the specific surgical environment. In this context, SP recordings can be driven by specific Surgical Process Models (SPMs), including complex dedicated ontologies.

The need of model-based systems for assisting and monitoring Computer-Assisted Surgical (CAS) has been discussed by Jannin et al. [9]. A model was proposed in the context of neurosurgical intervention, based on an UML class diagram and a textual description for decomposing the procedure. [15] focused on the description of concepts and technologies for the acquisition of surgical workflows by monitoring surgical interventions. They introduced an universal and adaptable recording scheme describing the subdivision of the surgical interventions into manual work steps detailed records. A new software was implemented (*i.e.* a surgical workflow editor: the ICCAS editor system) that permitted to record process during the intervention. They also introduced methods for computing generic SPMs that could serve for the generation and the comparison of surgical procedures [14,16].

Finally, Padoy et al. [17] proposed a system for the recognition of surgical workflow. They used Hidden Markov Models and Dynamic Time Warping to analyze and process a set of SPs and identify the different phases of the surgeries.

3 Methods

3.1 Surgical Process (SP) as sequence of activities

A Surgical Process (SP) can be seen in the real world as a sequence of flow objects [23]. According to the Workflow Management Coalition (WFMC) terminology [24], we name flow objects representing surgical work steps 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 i^{th} activity). Each activity in a SP 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 $\mathbf{ac}_i$. A full example of one 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. An ontology can be used to describe the vocabulary for a specific type of surgery [14].

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 $start(\mathbf{ac}_i) < stop(\mathbf{ac}_i)$, inducing 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 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 the surgery ($\mathcal{AC}_k$) :

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

3.2 Comparing SPs using Dynamic Time Warping (DTW)

When dealing with SPs, an important challenge is the design of metrics to evaluate the similarity of SPs. Indeed, defining a similarity measure is often the first step to identify patterns among a set of objects. As a SP can be seen as sequence of activities, we propose to use the Dynamic Time Warping (DTW) algorithm [7] to compare them. DTW is based on the Levenshtein distance (or edit distance), and was originatively 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 has to be defined to evaluate the similarity between the different elements composing the sequence. In our case, it means defining a distance between

two activities. Thus, we defined this distance as a binary function which returns 0 if all the three components (Eq. 1) of the two activities are equals, 1 else :

$$d(\mathbf{ac}_i, \mathbf{ac}_j) = \begin{cases} 0 & \text{if } \mathbf{ac}_i(\mathbf{a}) \stackrel{*}{=} \mathbf{ac}_j(\mathbf{a}) \text{ and} \\ & \quad \mathbf{ac}_i(\mathbf{s}) \stackrel{*}{=} \mathbf{ac}_j(\mathbf{s}) \text{ and} \\ & \quad \mathbf{ac}_i(\mathbf{i}) \stackrel{*}{=} \mathbf{ac}_j(\mathbf{i}) \\ 1 & \text{else} \end{cases} \tag{3}$$

with $\stackrel{*}{=}$ a boolean operator performing the comparison between the action, the anatomical structure or used the instrument(s) (*e.g.* $d($ <cut, skin, scalpel>,<cut, skin, scalpel>$)=0$).

To compare two SPs using DTW, the sequence of activities is first stretched by considering the starting and stopping of each activity. This step is needed to have the two SPs on the same timeline and to be able to compare the activities performed in each SP at given time point t in the timeline. The activity performed at the time point t will be $\mathbf{ac_i}(t)$ iff $t \in [start(\mathbf{ac_i}); stop(\mathbf{ac_i})]$. Note that it is not necessary that the two SPs last the same amount of time, the only assumption we make is that for both SPs, the first activity started at the same moment ($t = 0$).

Considering two SPs, $\mathbf{sp}_k =< \mathbf{ac}_1^{(k)}, \mathbf{ac}_2^{(k)}, \ldots, \mathbf{ac}_{n_k}^{(k)} >$ and $\mathbf{sp}_l =< \mathbf{ac}_1^{(l)}, \mathbf{ac}_2^{(l)}, \ldots, \mathbf{ac}_{n_l}^{(l)} >$ the cost of the optimal alignment can be recursively computed with :

$$\begin{aligned} &d(\mathbf{sp}_k(t), \mathbf{sp}_l(t)) = \\ &d(\mathbf{ac}_i^{(k)}(t), \mathbf{ac}_j^{(l)}(t)) + min \begin{cases} d(\mathbf{sp}_k(t-1), \mathbf{sp}_l(t-1)) \\ d(\mathbf{sp}_k(t), \mathbf{sp}_l(t-1)) \\ d(\mathbf{sp}_k(t-1), \mathbf{sp}_l(t)) \end{cases} \end{aligned} \tag{4}$$

where $\mathbf{sp}_k(t)$ is the subsequence $< \mathbf{ac}_1^{(k)}, \ldots, \mathbf{ac}_i^{(k)}(t) >$. Direct implementation of this recursive definition has an exponential cost. Fortunately, by decomposing it in subproblems, the complexity can be narrowed down to $N_k \times N_l$, N being the number of time points in a SP, which is equivalent to the stopping value of the last activity of the SP.

DTW has already been successfully used as similarity measure for clustering [8], for example in [18] where the authors used DTW to perform KMeans clustering on sequential data to identify patterns in remote sensing images. Thus, we propose to use the similarity measure defined using DTW to automatically identify clusters of similar surgeries. We used an ascendent hierarchical clustering approach with the average-link method [11], which consists in evaluating the similarity of two clusters according to the average distance between all the couple of objects in the two clusters.

3.3 Data presentation

Twenty-four procedures (10 men, 14 women, median age of 52 years) of lumbar disc herniation surgery were recorded in the Neurosurgery Department of the

Leipzig University Hospital, Germany. The procedure can be divided into three major steps: the approach of the herniated disc via a posterior intermyolamar route, the discectomy including the dissection and removal of the disc, and the closure step. Additionally, a step of hemostasis might be necessary before the closure. Five senior surgeons and five junior ones have participated to the study. Senior surgeons had already performed more than 100 removals of lumbar disc herniation, whereas junior ones had performed more than 2 years of their residency program. Among the 24 recorded procedures, 12 were performed by one senior surgeon with the aid of one junior surgeon, and in the 12 remaining cases, surgery was performed by one junior surgeon with the aid of one senior surgeon. During all junior recordings, the only step that was performed by junior surgeons without the help of senior ones was the closure step. Thus, in this paper, we focused on the analysis of this last step for a better discrimination of junior and senior performance.

The data were acquired using the Surgical Workflow Editor [13]. SPs were recorded on-line by an observer, a senior neurosurgeon, with the help of a touch-screen laptop to facilitate the recording task.

4 Results

The figure 1 presents the dendrogram of the AHC for the closure step, which is the only step performed by junior and senior surgeons alone. Along with the clustering, the right-hand activities of each surgery are shown below as index-plots. The idea of index-plot is to display the sequence by representing an activity as a rectangle of a specific color for each activity, and a width proportional to its duration Three clusters visually emerged from the analysis. When cutting the dendrogram for creating 2 clusters, 12 surgeries in each cluster can be extracted. The first cluster (left part, in green and blue) contains 10 surgeries performed by senior and 2 by juniors. For the second cluster (right part, in red), 10 surgeries operated by juniors and 2 by seniors are present. Additionally, within the first cluster, a sub-classification can be found (between red and blue parts), where each sub-cluster contains five surgeries performed by seniors and 1 by junior. By keeping the two main clusters, an accuracy of 83.33% is found, considering that 20 surgeries over 24 are classified into the right cluster.

5 Discussion

5.1 Classification of surgeon's experience

According to figure 1, two main clusters are clearly identifiable. It turns out that these two clusters are strongly correlated with surgeon's experience, which is not surprising. Indeed, an actual tendency shows that senior surgeons perform less gestures than juniors. Particularly during a mechanical step (the closure step), experienced surgeons are more economical with their movements than the inexperienced ones, which explains that the DTW similarity can be able to capture

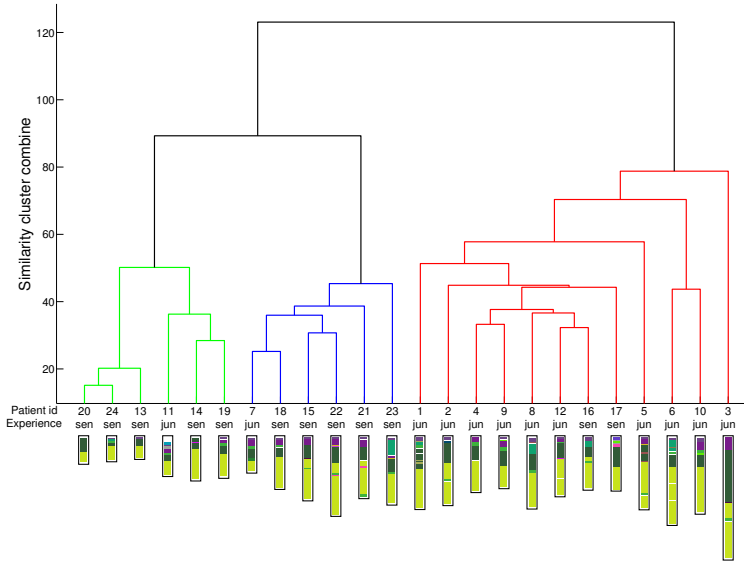

Fig. 1. Dendrogram of the clustering of the 24 SPs

the differences between both groups. Moreover, the two junior surgeries classified as being senior ones were actually both operated by the same junior surgeon. During the preliminary phase of experience's classification, our reference neurosurgeon wondered whether this surgeon had to be considered as being junior or senior regarding his intermediate experience (*i.e.* seventh year of resident training), which could explain this classification error. On the contrary, the two senior surgeries classified as being junior ones were not particularly complicated surgeries, as the total time of both surgeries were both quite low. This classification error could be explained by a lower gesture economy of corresponding senior surgeons, or even a lower manual dexterity.

According to figure 1, two sub-clusters of senior surgeons can be extracted. Similarly to the correlation with surgeon's experiences, explications with other pre-operative data were explored, without success. However, both sub-clusters contain surgeries performed by same senior surgeons. It reveals that seniors can have different operating techniques and preferably sequence of activities that differ from one senior to the other and that can explain this distinct separation.

5.2 A new metric for surgery comparison

The DTW approach for surgeries comparison allows to focus only on the sequentiality of surgeries by minimizing the time differences. Indeed, the DTW algorithm has been first used to synchronize two time series, for instance in the context of speech recognition. Using this method for synchronizing surgeries

permit to take into account the difference of activities sequences, without time constraint. Assuming that for skill-evaluation the time is not a major parameter, the number of activities associated with their sequentiality is more relevant and surgery dissimilarities can be objectively quantified using the DTW distance. This metric is therefore an interesting and innovative way for comparing surgeries, and turns out to be a complementary approach of standard time/number of occurrences comparison approaches [21].

5.3 Specific applications: training and assessment

Training and assessment of surgeons are now considered as crucial issues for patient safety. Training of junior surgeons is a very time-consuming, interactive and subjective task. As all juniors are currently learning with the teaching help of seniors, there has been a new demand for simulation devices. Moreover, some surgeons are clearly superior in performing tasks than other, resulting in a growing pressure to demonstrate their competences. These two challenges have motivated the creation of automatic systems for objective assessment of surgical skills. With automatic techniques recently proposed using sensor devices, systems are able to precisely recognize activities through different level of granularities, from the simple gesture to the global phases of the surgery, which is a powerful tool to automate surgical assessment and surgical training without being biased by humans. For assessment, surgical activities can be scored for precision, dexterity or global performance. For training, it would allow surgeons to benefit from constructive feedback and to learn from their mistakes. Similar methods can also be employed for other type of surgeries, or even other members of the surgical team.

6 Conclusion

An important challenge is the creation of new metrics for the comparison and the evaluation of SPs. In this paper, we proposed a new surgery metric based on the DTW algorithm that permit to focus the analysis on the different types of activity performed during the surgery and their sequencing, and not on the time differences. Results on the classification of the level of expertise of surgeons were shown.

One possibility for improving the analysis would be to introduce semantic in the surgeries similarity metrics. In this work, at each time step, a binary comparison of two surgical activities is performed. The idea would be to introduce semantic matrix in order to link each activity using different distance values of a predefined similarity scale for more complex analysis of SPs.

References

1. Bridgewater, B., Grayson, A., Jackson, M., Brooks, N., Grotte, G., Keenan, D., Millner, R., Fabri, B., Mark, J.: Surgeon specific mortality in adult cardiac

surgery: comparison between crude and risk stratified data. British Medical Journal 327(7405), 13 (2003)
2. Chmarra, M., Grimbergen, C., Dankelman, J.: Systems for tracking minimally invasive surgical instruments. Minimally Invasive Therapy & Allied Technologies 16(6), 328–340 (2007)
3. Datta, V., Mackay, S., Mandalia, M., Darzi, A.: The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. Journal of the American College of Surgeons 193(5), 479–485 (2001)
4. Doyle, J., Webber, E., Sidhu, R.: A universal global rating scale for the evaluation of technical skills in the operating room. The American Journal of Surgery 193(5), 551–555 (2007)
5. Francis, N., Hanna, G., Cuschieri, A.: The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester: contrast validity study. Archives of Surgery 137(7), 841 (2002)
6. Guthart, G., Salisbury Jr, J.: The IntuitiveTM telesurgery system: overview and application. In: IEEE International Conference on Robotics and Automation. vol. 1, pp. 618–621 (2000)
7. Hiroaki, S., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26, 43–49 (1978)
8. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8), 651–666 (2010)
9. Jannin, P., Raimbault, M., Morandi, X., Riffaud, L., Gibaud, B.: Model of surgical procedures for multimodal image-guided neurosurgery. Computer Aided Surgery 8(2), 98–106 (2003)
10. 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 10(3), 121–127 (2001)
11. Manning, C., Schutze, H., MITCogNet: Foundations of statistical natural language processing, vol. 59. MIT Press (1999)
12. Moorthy, K., Munz, Y., Sarker, S., Darzi, A.: Objective assessment of technical skills in surgery. British Medical Journal 327(7422), 1032 (2003)
13. Neumuth, T., Durstewitz, N., Fischer, M., Strauß, G., Dietz, A., Meixensberger, J., Jannin, P., Cleary, K., Lemke, H., Burgert, O.: Structured recording of intraoperative surgical workflows. In: SPIE Medical Imaging. vol. 6145, p. 61450A (2006)
14. 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 6, 59–71 (2010)
15. Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H., Burgert, O.: Acquisition of process descriptions from surgical interventions. In: Database and expert systems applications. pp. 602–611 (2006)
16. Neumuth, T., Jannin, P., Strauss, G., Meixensberger, J., Burgert, O.: Validation of knowledge acquisition for surgical process models. J Am Med Inform Assoc 16(1), 72 – 80 (2009)
17. Padoy, N., Blum, T., Ahmadi, A., Feussner, H., Berger, M., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis (2010)
18. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition 44, 678–693 (2011)

19. Reiley, C., Lin, H., Yuh, D., Hager, G.: Review of methods for objective surgical skill evaluation. Surgical Endoscopy pp. 1–11 (2011)
20. Reznick, R., Regehr, G., MacRae, H., Martin, J., McCulloch, W.: Testing technical skill via an innovative "bench station" examination. The American Journal of Surgery 173(3), 226–230 (1997)
21. Riffaud, L., Neumuth, T., Morandi, X., Trantakis, C., Meixensberger, J., Burgert, O., Trelhu, B., Jannin, P.: Recording of surgical processes: a study comparing senior and junior neurosurgeons during lumbar disc herniation surgery. Neurosurgery 67, 325–332 (2010)
22. Rosen, J., Hannaford, B., Richards, C., Sinanan, M.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Transactions on Biomedical Engineering 48(5), 579–591 (2001)
23. White, S.: Introduction to BPMN. IBM Corporation 31 (2004)
24. Zur Muehlen, M.: Organizational management in workflow applications–issues and perspectives. Information Technology and Management 5(3), 271–291 (2004)