This is a post-peer-review, pre-copyedit version of an article published in the International Journal of Computer Assisted Radiology and Surgery. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11548-015-1195-0.

Automatic phase prediction from low-level surgical activities

Germain Forestier · Laurent Riffaud · Pierre Jannin

Received: date / Accepted: date

Abstract
Purpose Analyzing surgical activities has received a growing interest in recent years. Several methods have been proposed to identify surgical activities and surgical phases from data acquired in operating rooms. These context-aware systems have multiple applications, including: supporting the surgical team during the intervention, improving the automatic monitoring, designing new teaching paradigms.
Methods In this paper, we use low-level recordings of the activities that are performed by a surgeon to automatically predict the current (high-level) phase of the surgery. We augment a decision tree algorithm with the ability to consider the local-context of the surgical activities and a hierarchical clustering algorithm.
Results Experiments were performed on 22 surgeries of lumbar disc herniation. We obtained an overall precision of 0.843 in detecting phases of 51,489 single activities. We also assess the robustness of the method with regard to noise.
Conclusion We show that using the local-context allows us to improve the results compared to methods only considering single activity. Experiments show that the use of the local context makes our method very robust to noise and that clustering the input data first improves the predictions.

Keywords Surgical Process · Temporal Analysis · Prediction · Surgery

G. Forestier
University of Haute-Alsace, MIPS
12, rue des freres Lumiere
68093 Mulhouse France
E-mail: germain.forestier@uha.fr

L. Riffaud
Department of Neurosurgery
Pontchaillou University Hospital
2 Av. du Pr Leon Bernard
35043 Rennes, France
E-mail: laurent.riffaud@chu-rennes.fr

P. Jannin
INSERM MediCIS, Unit U1099 LTSI
University of Rennes 1
2 Av. du Pr Leon Bernard
35043 Rennes, France
Rennes, France
E-mail: pierre.jannin@univ-rennes1.fr

1 Introduction

In recent years, Operating Rooms (ORs) have undergo tremendous changes with the increase of available technology to support and assist surgical teams. One of the targeted goals is the development of *context-aware* systems [4] that continuously monitor the activities performed in the ORs in order to provide an accurate and reliable support. The key challenge in developing these new methods is to process the data coming from sensors and real-time detection systems in order to provide useful information and support decision making. This task is challenging because of the complexity of the OR environment and the high variability of surgical interventions due to patient abnormalities, surgeon experience and OR specific constraints.

The field of Surgical Process Modeling (SPM) [11] targets the development of new methods that leverage from OR activities monitoring. In this field, several methods have already been proposed to automatically detect surgical activities. These methods rely either on manual annotations by an observer [5] or on sensors present in the OR (*e.g.*, camera) [8,12]. For example, the task performed by a surgeon can be automatically inferred by combining RFID (Radio Frequency Identi-

fication) chips on instruments (for identification) with accelerometers [14]. Note that phases and surgical activities are not the only interesting information to analyze. For example, Franke et al. [8] proposed a system to predict intervention time from low-level surgical activities.

The automatic recognition of the current phase during a surgery is of major interest for various applications in the OR. For example, peri-operative systems that support medical decision have to be aware of the current phase to understand the context upon which a specific activity is performed. Depending of the current phase, similar surgical activities do not have the same semantic and the same medical goal. The phase information can also be used to improve the coordination and communication among the surgical team or for general monitoring purposes.

A surgery can traditionally be modeled with different levels of granularity [11] (*e.g.*, procedure, steps, substeps, tasks, subtasks etc.). In this paper, we target the automatic prediction of high-level surgical phases from the low-level recordings of the surgical activities that are performed by surgeons. We model these activities as a triplet composed of action, anatomical structure and surgical instrument (*e.g.*, to *cut* the *skin* with a *scalpel*), in order to automatically infer the current phase of the surgery (*e.g.*, the opening phase).

In this paper, we propose a method based on a decision tree [17] to perform the phase prediction from low-level activities. We show that a good prediction accuracy can be obtained by only considering the surgical activities at a given time. We then further extend our method to use the local context of surgical activities to draw a more accurate prediction of the phases. In this extension, we do not only consider the prediction from the current activity but also the predictions made from the previous activities within a selected time window. We show that using the local context both improves the results of the prediction and increases the robustness with regard to noise in the data. We also show that a clustering of the input data allows us to improve the quality of phases identification, which suggests that information processing in CAI systems is critical.

Experiments were performed on a dataset of 22 lumbar disc surgeries which is the most commonly performed spinal surgery world-wide. Sixty-thousand such interventions are performed every year in France [2]. An interesting feature of this type of surgery is the variability of the number of phases needed during each intervention. Some phases are indeed optional (*e.g.*, hemostatis) and some phases have sometimes to be repeated (*e.g.*, disc removal/hemostatis). The number of phases can be difficult to predict from pre-operative information, as it depends upon the reaction of the patient to the surgery. The dataset contains 51,489 single surgical activities overall, classified into four phases.

The contributions of this paper are:

1. A system that can predict high-level surgical phases from low-level surgical activities and its extension to consider the local context of the activities.
2. An experimental study of the influence of noise in low-level activities on the prediction of high-level phases.
3. An evaluation of the influence of clustering the input data prior to phases detection on the improvement of the prediction accuracy.

2 Prediction of surgical phases

The prediction of our proposed method is based on decision trees [17]. This classification method has shown to be successful for the prediction of surgical phases. Bardram et al. [1] proposed a system using embedded and body-worn sensor data to train a decision tree in order to predict surgical phases. They studied sensor significance in order to identity the most important features for surgical phase prediction. Stauder et al. [21] used Random Forest (*i.e.*, a bag of decision trees) to predict surgical phases from sensors measurement. While these methods are using sensors to predict the phases, we target in this paper the prediction of the current phase from the current surgeon's activity. Our goal is to predict the surgical phase knowing what the surgeon is currently doing. Note that surgeons activities can themselves be derived from sensors data in specific contexts [10].

Other models like Hidden Markov Model (HMM) were also considered by Padoy et al. [15,16] for online recognition of surgical steps. In this work, surgical activities were extracted using image processing techniques on laparoscopic camera. Similarly, Bouarfa et al. [3] used HMM with a pre-processing on the input sensor data in order to improve the detection of high-level surgical taks. SVM classifier was also considered by Lalys et al. [12] to detect phases and low-level surgical tasks using cameras in pituitary surgery. Varadarajan et al. [22] used HMM to recognize and segment surgical gestures for surgical assessment and training. Learning the topology of an HMM is however still challenging and improving this step continues to be investigated [20].

In this paper, we chose decision trees for the readability of the produced model (*i.e.*, a tree). They can indeed be easily converted into decision rules that can then be discussed with medical experts. Thus, even by just analyzing manually the outputted tree, decisions

can be made on the organization of the OR without implementing complex systems. Furthermore, the learning step can be performed offline, and the online prediction is not computationally expensive as it is linear with the number of rules. This feature can be useful if the prediction is implemented in the OR through wearable device (*e.g.*, Google Glass), as their computational capabilities are limited. One drawback of using decision trees is the loss of the temporal aspect, as each prediction is only performed according to the current activity. However, this feature is quite useful when the number of phases is unknown and can be variable. In some applications, according to patient abnormalities, the number and the sequencing of the phases can be unknown. Thus, it is possible that new surgeries exhibit a phase sequencing that has never been used before. In that case, it can be difficult for an HMM to detect a phase sequencing that has not been present in the learning set. Another feature of decision trees is the possibility to stop and restart the prediction system. While HMMs generally use the past to predict the future, decision trees systems only use the present. HMM and decision trees system are thus complementary depending of the amount of available data, complexity of the temporal sequencing of the phases and computational resources available. Finally, using the local-context to draw the prediction, as proposed in this paper, allows to partially take into account the temporal aspect as it consist in using the *near* past.

3 Materials and Methods

3.1 Problem statement

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 representation and use its formalisation introduced in [6]. 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. Thus, an activity is a vector of seven nominal values corresponding to:

- the triplet for the right hand ;
- the triplet for the left hand ;
- a binary information on the use of the microscope.

An example of activity could be $\{$*(cut, scissors, muscle)*$_r$, *(hold, retractors, muscle)*$_l$, *false*$\}$. Each surgery is subdivided into several phases $P = \{p_1, ..., p_m\}$ (*e.g.*, closure phase) which corresponds to a high-level segmentation of the surgery. Each activity is performed during certain phase of the surgery. The goal of this paper is to predict the phase p_j given an activity a_i. The Figure 1 presents an example of a surgery composed of several activities. In this visualisation, each colour corresponds to a different activity. The phases we are targeting to predict are also displayed on the figure.

To create the prediction function $f : \mathcal{A} \rightarrow P$ that affects a phase to an activity, we used of a decision tree algorithm. In order to train this decision tree, a training set is composed of surgeries from which activities and their corresponding phases are known: $S =< (a_1, p_1), ..., (a_n, p_m) >$. Note that the use of Random Forest as presented in [21] is not relevant in our case as we handled a limited amount of features. Consequently, creating multiple trees with subsets of features is not likely to improve the results.

3.2 Considering the local-context information

In the problem statement, we only considered the activity performed at a single instant to draw the prediction. In order to take into account the local-context of the current activity, we propose to also consider the previous actions performed by the surgeon. This allows to take into account the local sequencing of the activities. Our intuition is that the current activity is related to the previously performed activities.

Let $a(t)$ be the activity performed at time t. We consider for the prediction the set of activities from $a(t)$ to $a(t-w)$, w being the size of a *time window* we are considering which is a parameter of the method.

We analyze the w previous activities and we combine the probability density functions (PDF) of the predictions related to these previous activities. The number of bins of each PDF is equal to the number of different possible phases in the surgery. The PDF of an activity a (pdf_a) is computed according to the class distribution (*i.e.*, phase distribution) of the instances (*i.e.*, surgical activities) present in the tree node used to perform the prediction. The value $pdf_a(p)$ corresponds to the fraction of instances of phase p in that node. Thus, the probability of each phase is computed by summing all the PDFs:

$$\hat{p}(p_j|A_w) = \sum_{a \in A_w} pdf_a(p_j) \quad (1)$$

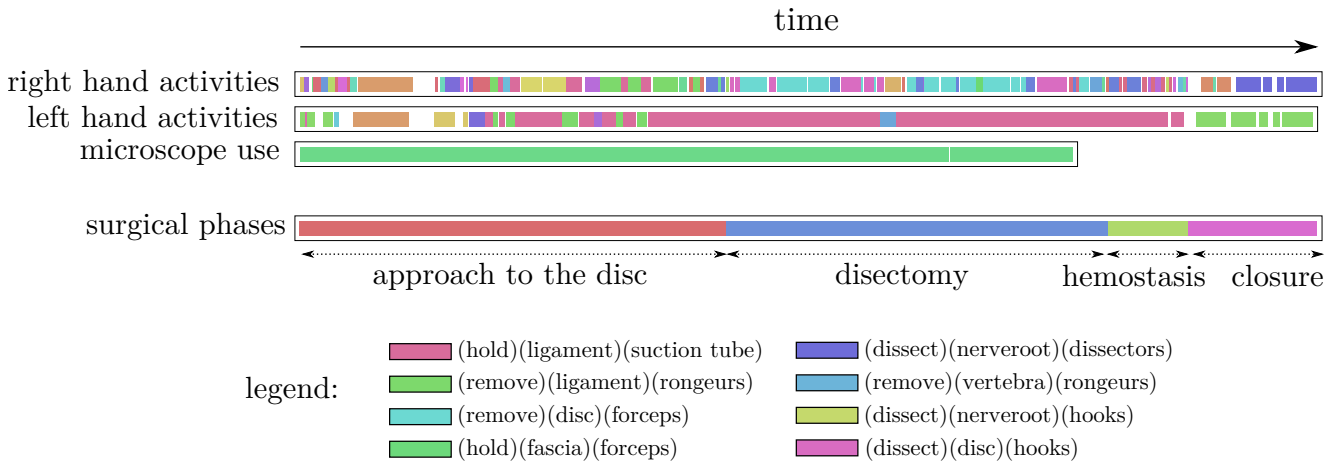

Fig. 1 Visualization of a surgery, each color corresponds to a different activity (see legend on the bottom). Our goal is to predict the phase (four phases in this example) by using the information of the activities.

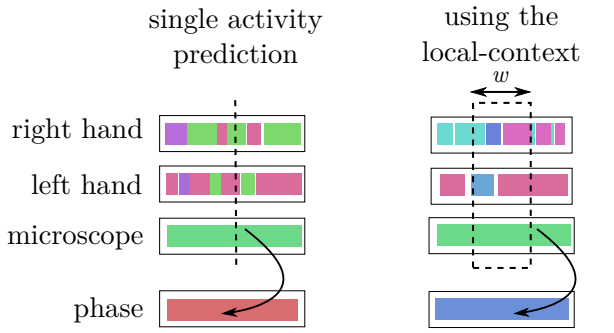

Fig. 2 Illustration of the activities used to draw the prediction: using a single activity (left) or the local-context (right) with the time widow w.

with $A_w = \{a(t-w), ..., a(t-1), a(t)\}$ the set of activities performed during the time window. The probability $\hat{p}(p_j|A_w)$ corresponds to the probability of predicting the phase p_j knowing the set of previous activities A_w. The final prediction is drawn by taking the mode, *i.e.*, the phase having the maximum probability:

$$\arg\max_{p_j \in P} \{\hat{p}(p_j|A_w)\} \quad (2)$$

The optimal size w of the time window can be experimentally found using cross-validation for a specific application. As presented later, 70 seconds gave the best results in our experiments. Note that the time window concerns the activities performed by the surgeon *before* the current activity that the system aim at predicting. Thus, there is no delay in the prediction, except for the first 70s of the surgery. The Figure 2 illustrates the two methods using a single activity and a window of local-context information.

Using the PDFs also allows to provide a confidence on the classification based on the highest probability provided by Eq. 2. For example, if the mode is of 90% for a specific phase, then the system is highly confident on the prediction. Ties can appears in Eq. 2 if multiple phases have same maximum probability. In that case, the prediction is made randomly from the set of phases having the maximum value. Note that this specific case never happened in the experiment.

4 Experiments and Results

4.1 Dataset

We evaluate our method using clinical data composed of 22 surgeries of lumbar disc herniation which is the most commonly performed spinal surgery [2]. The data were recorded at the Neurosurgery Department of Leipzig. The surgeries involved 9 male and 13 female patients, with a median age of 52 years. These were exclusively patients with newly diagnosed disc herniation, no patient had undergone previous lumbar spine surgery which might be supposed to increase surgical difficulties due to fibrosis. The herniated disc was approached via a posterior intermyolamar route.

This procedure is composed of 4 phases: (1) approach to the spine (from skin incision to the incision of the posterior longitudinal ligament or the removal of an excluded portion of the disc), (2) disc removal (from the end of the previous step to the beginning of hemostasis or closure), (3) hemostasis (this step may be not individualised if it was not performed), and (4) closure (from the end of disc removal or hemostasis to the end incision closure) [18].

Depending on patient specificities, a succession of disc removal / hemostasis phases is sometimes required.

Thus, the number of phases is unknown at the beginning of the surgery. This element motivates the need for a method without assumption on the number of phases. This variability on the number of phases is actually often present due to patient specificities. Among the 22 lumbar disc herniation surgeries of the dataset, 4 (18,2%) required only 3 phases (no need for hemostasis phase), 14 (63.6%) required 4 phases (one phase of hemostasis), 3 required 6 phases (13.7%) (two successions of disc removal / hemostasis) and 1 (4.5%) required 8 phases (three successions of disc removal / hemostasis). Each activity is labeled of the phase during which it is performed. We do not differentiate between disc removal / hemostasis that appear only once and disc removal / hemostasis that appears multiple times as their medical objective are identical.

For this surgery, the list of actions is: *cut, coagulate, hold, dissect, install, remove, irrigate, sew, swab* and *drill*. The list of anatomical structures is: *skin, fascia, muscle, vertebra, ligament, duramater, nerveroot* and *disc*. And the list of surgical instruments is: *scalpel, scissors, dissectors, rongeurs, hooks, high-speed drill, suction tube, needle-holders, saline solution, retractors* and *forceps*. Theoretically, 880 ($10 \times 8 \times 11$) different triplets could be created, which gives more that 1.5M of possible different activities performed by the surgeon at a single instant (considering right and left activities and microscope use). However, as all triplets are not present (some triplets of action, instrument, anatomical structure are irrelevant), our dataset contains only 108 different triplets, leading to 23 thousands possible different activities ($108 \times 108 \times 2$). The overall 22 surgeries contains 51,489 activities: 24,566 (48%) activities for the approach phase, 15,587 (30%) activities for the discectomy phase, 3,901 (8%) activities for the hemostasis phase and finally, 7,435 (14%) for the closure phase. Note that these numbers of possibilities do not consider the coordination between the hands of the surgeon, which would adds constraints on the possible combinations.

4.2 Phase prediction using single activity

In this experiment, we only considered the single activity the surgeon is performing at a given instant to draw the prediction of the current phase. We trained the decision tree on all but one of the available surgeries and tested the tree on the remaining one. This process was carried out for each surgery so that each one was used for learning and testing in a leave-one-out way. The evaluation were computed from the confusion matrix obtained from this process. We used as decision tree the algorithm C4.5 and its implementation in Weka (J48) [9] with default parameters. On average during the leave-one-out process, the obtained decision tree had 595 leaves and 662 nodes. However, by removing the leaves that do not contain instances, the tree is on average composed of only 150 rules. An example of decision tree and its corresponding rules set are provided on the companion web-page of the paper.

Table 1 Confusion matrix of the different phases obtained using a leave-one-out cross validation on the 22 surgeries.

classed as →	App.	Disc.	Hemos.	Clos.
Approach	21890.0	1858.0	652.0	166.0
Discectomy	3109.0	11215.0	1165.0	98.0
Hemostasis	1537.0	1148.0	1168.0	48.0
Closure	1172.0	88.0	43.0	6132.0

Table 2 Precision, recall and f-measure according to the four different phases using a single activity.

Phase	Precision	Recall	F-Measure
Approach	0.790	0.891	0.838
Discectomy	0.784	0.720	0.750
Hemostasis	0.386	0.299	0.337
Closure	0.952	0.825	0.884
Weighted mean	0.781	0.785	0.780

The Table 1 presents the confusion matrix for the four phases, rows are ground truth, columns are predictions. The Table 2 presents the precision, recall and f-measure (harmonic mean of precision and recall) for the four phases. The overall weighted (per class cardinality) precision is of 0.781, weighted recall is of 0.785 and weighted f-measure of 0.780. The Figure 3 illustrates the prediction performed by the system on the surgery presented in Figure 1 and the actual phases (*i.e.*, ground truth).

4.3 Phase prediction using local-context

In this experiment, we used the augmented version using the local-context. In order to use this evolution of the method, the w parameter corresponding to the size of the time window has to be fixed. To set-up this parameter on our dataset, we tested a range of time window sizes within $[0; 200]$ with a step of 5. We used an identical leave-out-approach as presented in the previous experiment. The Figure 4 illustrates the evolution of the f-measure according to the window size. We used the f-measure as it combines the information from the precision and the recall. From this experiment, we iden-

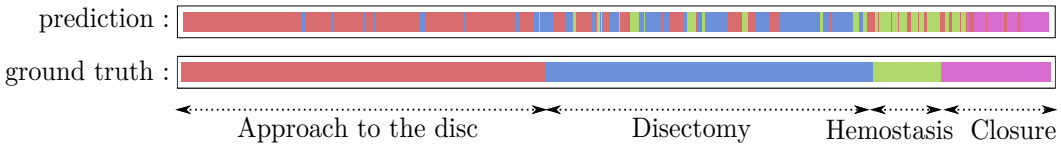

Fig. 3 Visualisation of the prediction of the phases for one surgery, the prediction (top) and the groundtruth phases (bellow).

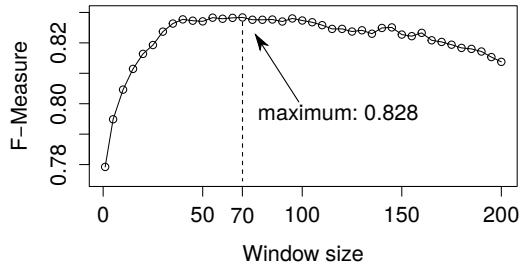

Fig. 4 Evolution of the Precision according to different size of Window time (t).

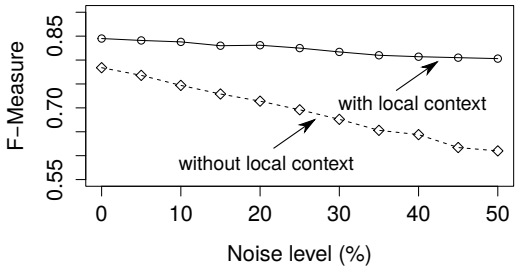

Fig. 5 Evolution of the F-Measure according to different level of noise in the data.

Table 3 Precision, recall and f-measure according to the four different phases using local-context.

Phase	Precision	Recall	F-Measure
Approach	0.834	0.967	0.896
Discectomy	0.817	0.802	0.810
Hemostasis	0.777	0.197	0.314
Closure	0.965	0.870	0.915
Weighted mean	0.843	0.845	0.828

tified that the optimal value of w for our application was 70. The Table 3 presents the results for the different phases with the parameter $w = 70$. In this experiment, the f-measure is reaching 0.828 (Table 3) compared to 0.780 (Table 2) by considering only a single activity. The precision is reaching 0.843 and the recall 0.845. It is interesting to note from Figure 4 that the context has to stay *local* as using many previous activities eventually reduces the quality of the prediction. Note that the method is not sensible to small variations of this parameter as the f-measure is barely stable in the range 40 to 100 (Figure 4).

4.4 Phase prediction under noisy data

Noise is an important parameter when processing surgical activity data. It is often present in activities inferred from sensors data [10] and even from data captured by an observer [6]. Noise can deeply influence the quality of phases prediction. In this experiment, we evaluate how our phase prediction system is influenced by noisy data. We artificially added noise to the available dataset by randomly switching the value of a given feature of an activity under a certain probability. This noise introduction simulated errors in manual labelling or detection errors of systems using sensors data. An example of noise introduction could be to switch the instrument of an activity from *scalpel* to *forceps*. A level of noise of $N\%$ means that each feature of each activity has $N\%$ of chance to have been modified randomly. With an increasing level of noise it is becoming more difficult to predict the phase as the tree has difficulties to identity valid decision rules. The Figure 5 presents the evolution of the f-measure according to different levels of noise (from 0% to 50%) for our prediction method with and without local-context information.

4.5 Phase prediction among clusters of surgeries

In the previous experiments, we used the entire dataset of 22 surgeries. In this experiment, we first perform a clustering of this dataset to create groups of similar surgeries. We then apply the phase prediction systems to the surgeries present in each cluster. The influence of reducing a training set using clustering techniques before learning a decision tree has already been investigated in the past [19,23]. As surgeries can exhibit important differences, we are expecting to improve the prediction results by creating clusters of highly similar surgeries. To create the clusters, we used the methodology proposed by Forestier et al. [5,6] which relies on Dynamic Time Warping (DTW) and ascendant hierarchical clustering. This methodology has proven its efficiency in creating clusters of similar surgeries [7]. The Figure 6 presents the dendrogram obtained from the hierarchical clustering process. From the analysis of this dendrogram, we identified two main clusters, named

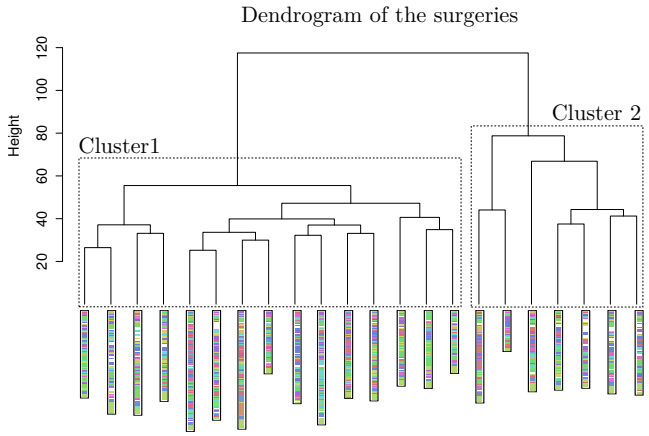

Fig. 6 Hierarchical clustering of the 22 surgeries used in the experiment. Two clusters are visually identifiable. On the bottom of the tree, the sequences of right hand activities.

Cluster 1 and Cluster 2 and containing respectively 15 and 7 surgeries. We then applied the system for phase prediction proposed in this paper individually to each cluster. The Table 4 presents the results of phases prediction within the two clusters for the methods with and without local-context usage. The results within Cluster 1 (15 surgeries) are quite interesting as they are better than the ones obtained on the entire dataset. Without the use of the local-context, we obtained a f-measure of 0.804 compared to 0.780 using the entire dataset. When using the local-context, the results further improved, reaching 0.845 compared to 0.828 when using the entire datasets. The results within Cluster 2 (7 surgeries) are however lower than using the entire dataset with a f-measure of 0.693 without the local-context and 0.720 while using it. The results of the application of the tree learned from the data of Cluster 1 to the entire dataset (Cluster 1 + Cluster 2) is presented in Table 6. In this Table, the f-measure is reaching 0.864 compared to 0.828 when using all the dataset to learn the tree.

5 Discussion

When using only the current activity of the surgeon (Table 1 and Table 2) the results are quite acceptable. Indeed, these results are interesting considering the very limited amount of used data (only seven features for each activity) representing what the surgeon is currently doing. The precision ranges from 0.952 for the *closure* phase to 0.386 for the *hemostasis* phase. The high precision rate for the *closure* phase can be explained as it is the most standardized phase of the surgery. Consequently, it is quite easy to learn a set of rules allowing to identify the activities of this phase. On the contrary, the *hemostasis* phase is more complex. First, it does not appear all the time, and second the duration of *hemostasis* phase is less important than the other phases of the surgery (*i.e.*, it only represents 8% of the activities of the dataset). The *approach* and *disectomy* phases have respectively a precision of 0.790 and 0.784 which is acceptable as they are the most present phases of the dataset (respectively 48% and 30% of whole the activities). The overall recall is of 0.785 which is also a good result.

When considering the local-context of the activities (Table 3), the overall f-measure increases from 0.780 to 0.828. Almost all the results increase, especially the precisions. The precision of the *hemostasis* phase increased from 0.386 to 0.777 but its recall fell from 0.299 to 0.197. This result means that the number of activity affected to the *hemostasis* phase reduced but the precision in the prediction increased. Thus, by considering what the surgeon did in his/her previous activities, the level of prediction errors can be reduced.

The added value of the local-context adjustment is even more visible when processing noisy data. In the Figure 5, one can see that the use of local-context allows the method to be less impacted by noisy data. The prediction system using only the current activity sees its performance decrease linearly with the increasing level of noise, while the method using the local-context is more robust. This is explained by the fact that the method relies on the predictions performed from previous activities. Thus, the influence of noise in one activity among the set of considered activities (A_w) is reduced.

The Figure 6 illustrates the clustering of the surgeries used in the experiment. We created two clusters out of the dendrogram obtained by hierarchical clustering. The results of the Cluster 1 (see Table 4) are better than the results on the entire dataset (Table 2 and 3). These results show that using an important number of surgeries does not necessarily improve the results. By reducing the number of surgeries (15 instead of 22), we substantially improved the results. This can be explained by the fact that the surgeries present in Cluster 1 are the most similar surgeries of the dataset. Consequently, it is easier for the decision tree algorithm to create decision rules that are valid for these 15 similar surgeries than a model for the entire dataset. However, the Cluster 2 shows lower results than when using the entire dataset. This can be explained by the higher dissimilarity within the clusters, which is visible from the height of the dendrogram links in Figure 6. Furthermore, Cluster 2 only contains 7 surgeries and with the leave-one-out cross-validation, only 6 surgeries were used at each fold to learn a predictive model. This lim-

Table 4 Precision, recall and f-measure according to the four different phases, with and without the use of local-context for the two clusters of surgeries.

Cluster 1	Without local-context			With local-context		
Phase	Prec.	Rec.	F-M	Prec.	Rec.	F-M
Approach	0.831	0.875	0.853	0.845	0.976	0.906
Discectomy	0.823	0.770	0.796	0.853	0.840	0.846
Hemostasis	0.396	0.248	0.305	0.783	0.182	0.296
Closure	0.858	0.971	0.911	0.966	0.881	0.921
Weighted mean	0.800	0.813	0.804	0.861	0.863	0.845

Cluster 2	Without local-context			With local-context		
Phase	Prec.	Rec.	F-M	Prec.	Rec.	F-M
Approach	0.680	0.802	0.736	0.722	0.888	0.796
Discectomy	0.702	0.635	0.667	0.796	0.674	0.730
Hemostasis	0.518	0.207	0.296	0.526	0.184	0.273
Closure	0.858	0.857	0.829	0.898	0.869	0.883
Weighted mean	0.698	0.704	0.693	0.755	0.760	0.745

Table 5 Precision, recall and f-measure according to the four different phases, with and without the use of local-context using the tree learned on Cluster 1 applied to Cluster 2.

Cluster 2	Without local-context			With local-context		
Phase	Prec.	Rec.	F-M	Prec.	Rec.	F-M
Approach	0.742	0.830	0.784	0.752	0.947	0.838
Discectomy	0.818	0.578	0.678	0.867	0.640	0.736
Hemostasis	0.470	0.555	0.509	0.692	0.447	0.543
Closure	0.763	0.934	0.840	0.902	0.875	0.888
Weighted mean	0.749	0.739	0.735	0.806	0.797	0.788

ited number of instances and the higher variability can explain these lower results.

Table 5 shows the result of the application of the tree learned from Cluster 1 to the data of Cluster 2 (Cluster 2*). In this configuration, the f-measure increased from 0.693 to 0.735 without the local-context and from 0.745 to 0.788 with the local-context. This result shows that using a model learned from one cluster can actually improves the results when applied to other clusters. Table 6 shows the result of the application of the tree learned on Cluster 1 to the entire dataset (Cluster 1 + Cluster 2). In that case, the f-measure is reaching 0.864 which is better than using the entire dataset (0.828). Thus, these results highlight that in building a CAI system for phase prediction, the data used for the learning step are really important and can substantially affect the results. We advice to create highly similar clusters of surgery specific to a group of surgeons before training a system. Furthermore, our experiments reveal that a model learned on a subset of the data can improve the overall prediction accuracy of the system.

Additional background knowledge could have been used to improve the results. For example, a natural ordering of some phases (*e.g.*, closure comes *always* after opening) could have been used to further improve the results. However, we wanted the method to stay as generic as possible to be easily reused in other configurations (*e.g.*, for other types of intervention). Not relying on a temporal model also allows the system to be paused and restarted very easily. This feature is interesting in multiple situations: to save battery, due to technical failure or problems in gesture recognition, etc. Moreover, due to patient abnormality or intervention specific context, the sequence of the phases can also be totally original. In this situation, not relying on a tem-

Table 6 Precision, recall and f-measure according to the four different phases, with and without the use of local-context with the tree learned on Cluster 1 applied to the entire dataset.

Clusters 1+2	Without local-context			With local-context		
Phase	Prec.	Rec.	F-M	Prec.	Rec.	F-M
Approach	0.835	0.876	0.855	0.843	0.971	0.902
Discectomy	0.863	0.716	0.783	0.892	0.789	0.837
Hemostasis	0.544	0.583	0.563	0.818	0.513	0.630
Closure	0.835	0.964	0.895	0.946	0.885	0.915
Weighted mean	0.821	0.818	0.817	0.871	0.869	0.864

poral model allows our system to be used anyway as no assumption is made on the phase sequencing. Thus, the system will work even if no data in the training set exhibits the phase sequencing of the surgery being processed. Furthermore, the system can be applied regardless of the number of phases in the surgery.

The low computational complexity of decision tree allows the system to be easily embedded on low powered devices present the OR (*e.g.*, Google Glass). Furthermore, decision tree can be visualized as a set of classification rules that could be used alone to take actions in the organization of the OR. Note that the source code of the methods proposed in this paper is available for download [1] so that they can easily be integrated and compared with other systems.

The number of phases, which is currently limited to four, only partially reveals the potential of the proposed method. However, as the method is totally generic, it can be directly used to detect and predict more phases. Furthermore, as surgical procedures have multiple levels of granularity, our method could also be applied to detect other levels than phases (*e.g.*, steps, substeps, etc.). The only limit is that the low-level surgical activities of these different levels have to exhibit specific characteristics that the tree could capture. Finally, a precise comparison against existing methods (*e.g.*, HMM based) would also be needed to highlight the performance of the proposed method. The fact that we released the source code of our application is a first step towards such comparison.

6 Conclusion

In this paper, we used decision trees to automatically predict high-level surgical phases from low-level activities. We proposed a method which uses the local context of an activity to draw a better prediction than a method using a single activity. Experiments highlighted that using the local-context improves the quality of the prediction. We also performed an evaluation to assess the robustness of the method towards noisy data. The use of the local context made the method more robust to noise in the data. Furthermore, we showed that creating clusters of similar surgeries could be used as a pre-processing step to improve the results of CAI phase prediction systems. Indeed, we showed that using a subset of 15 surgeries instead of the 22 of the entire dataset allowed to improve the results of the proposed methods. In future work, we are planing to take into account global information on the surgery and to combine global and local contexts in order to improve the quality of the prediction. Furthermore, we are currently investigating if training the decision tree with short sequences of activities would improve the prediction results.

[1] http://germain-forestier.info/src/ipcai2015/

Acknowledgement

The authors would like to thanks all the surgeons of the Neurosurgery Department of the Leipzig University Hospital, Germany involved in this work. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

Additional Material

The source code of the application is available for download (Java ARchive file) : http://germain-forestier.info/src/ipcai2015/.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki

declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

1. Bardram, J.E., Doryab, A., Jensen, R.M., Lange, P.M., Nielsen, K.L., Petersen, S.T.: Phase recognition during surgical procedures using embedded and body-worn sensors. In: IEEE International Conference on Pervasive Computing and Communications, pp. 45–53 (2011)
2. Blamoutier, A.: Surgical discectomy for lumbar disc herniation: Surgical techniques. Orthopaedics & Traumatology: Surgery & Research **99**(1), S187–S196 (2013)
3. Bouarfa, L., Jonker, P.P., Dankelman, J.: Discovery of high-level tasks in the operating room. Journal of Biomedical Informatics **44**(3), 455–462 (2011)
4. Bricon-Souf, N., Conchon, E.: Context awareness for medical applications. Medical Applications of Artificial Intelligence p. 355 (2013)
5. Forestier, G., Lalys, F., Riffaud, L., Collins, D.L., Meixensberger, J., Wassef, S.N., Neumuth, T., Goulet, B., Jannin, P.: Multi-site study of surgical practice in neurosurgery based on surgical process models. Journal of Biomedical Informatics **46**(5), 822 – 829 (2013)
6. Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., Jannin, P.: Classification of surgical processes using dynamic time warping. Journal of Biomedical Informatics **45**(2), 255–264 (2012)
7. Forestier, G., Petitjean, F., Riffaud, L., Jannin, P.: Nonlinear temporal scaling of surgical processes. Artificial Intelligence in Medicine **62**(3), 143–152 (2014)
8. Franke, S., Meixensberger, J., Neumuth, T.: Intervention time prediction from surgical low-level tasks. Journal of Biomedical Informatics **46**(1), 152 – 159 (2013)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)
10. 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 **8**(1), 39–49 (2013)
11. Lalys, F., Jannin, P.: Surgical process modelling: a review. International Journal of Computer Assisted Radiology and Surgery **8**(5), 1–17 (2013)
12. Lalys, F., Riffaud, L., Morandi, X., Jannin, P.: Automatic phases recognition in pituitary surgeries by microscope images classification. In: Information Processing in Computer-Assisted Interventions, vol. 6135, pp. 34–44. Springer (2010)
13. Mehta, N., Haluck, R., Frecker, M., Snyder, A.: Sequence and task analysis of instrument use in common laparoscopic procedures. Surgical endoscopy **16**(2), 280–285 (2002)
14. Meißner, C., Meixensberger, J., Pretschner, A., Neumuth, T.: Sensor-based surgical activity recognition in unconstrained environments. Minimally Invasive Therapy & Allied Technologies (2014)
15. Padoy, N., Blum, T., Feussner, H., Berger, M.O., Navab, N.: On-line recognition of surgical activity for monitoring in the operating room. In: AAAI, pp. 1718–1724 (2008)
16. Padoy, N., Mateus, D., Weinland, D., Berger, M.O., Navab, N.: Workflow monitoring based on 3d motion features. In: IEEE International Conference on Computer Vision Workshops, pp. 585–592 (2009)
17. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
18. 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)
19. Sebban, M., NockO, R., Chauchat, J., Rakotomalala, R.: Impact of learning set quality and size on decision tree performances. IJCSS **1**(1), 85 (2000)
20. Shi, Y., Bobick, A., Essa, I.: Learning temporal sequence model from partially labeled data. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 1631–1638. IEEE (2006)
21. Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H., Navab, N.: Random forests for phase detection in surgical workflow analysis. In: Information Processing in Computer-Assisted Interventions, pp. 148–157. Springer (2014)
22. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009, pp. 426–434. Springer (2009)
23. Ženko, B.: Learning predictive clustering rules. Informatica **32**, 95–96 (2008)