Optimal Sub-Sequence Matching for the Automatic Prediction of Surgical Tasks

Germain Forestier[1], Francois Petitjean[2], Laurent Riffaud[3,4], and Pierre Jannin[3]

[2] Faculty of Information Technology, Monash University, Melbourne, Australia
[3] INSERM MediCIS, Unit U1099 LTSI, University of Rennes 1, Rennes, France
[4] Department of Neurosurgery, Pontchaillou University Hospital, Rennes, France

Abstract. Surgery is one of the riskiest and most important medical acts that is performed today. The desires to improve patient outcomes, surgeon training, and also to reduce the costs of surgery, have motivated surgeons to equip their Operating Rooms with sensors that describe the surgical intervention. The richness and complexity of the data that is collected calls for new machine learning methods to support pre-, peri- and post-surgery (before, during and after).

This paper introduces a new method for the prediction of the next task that the surgeon is going to perform during the surgery (*peri*). Our method bases its prediction on the optimal matching of the current surgery to a set of pre-recorded surgeries.

We assess our method on a set of neurosurgeries (lumbar disc herniation removal) and show that our method outperforms the state of the art by providing a prediction (of the next task that is going to be performed by the surgeon) more than 85% of the time with a 95% accuracy.

1 Introduction

More than half a million surgeries are performed every day worldwide [1], which makes surgery one of the most important component of global health care.

This has motivated the growing interest in Computer Assisted Surgery (CAS) tools. More and more Operating Rooms (ORs) are getting equipped systems with sensing devices that can capture the surgeon's activities and environment. For example, using cameras in pituitary surgery, both the phases of the surgery [2] and the low-level surgical tasks [3] can be detected and recorded automatically. The task performed by the surgeon can also be automatically inferred by combining RFID chips on instruments (for identification) with accelerometers [4]. The collected information is very precise and rich, because it corresponds to the low-level actions and tools that are performed and used by the surgeon. Because it is so precise, the data is however extremely challenging to analyze. For example, two surgeons performing the same surgery on the same patient might exhibit a very different course of specific actions, while being surgically very similar: they might use the same technique, have the same patient outcome, etc. However, from the low-level point of view (the sequence of low-level tasks like *cut*, *suture*, etc.), these surgeries will look very different from each other.

Extracting useful high-level knowledge from this low-level data has been one of the research themes targeted by the field of Surgical Process Modeling (SPM) [5,6], which aims at understanding surgeries to improve the quality of care. The above-mentioned sensors capture the surgical tasks performed in real-time, which opens the door to using artificial intelligence methods to provide real-time information to the surgical team.

This paper seeks the prediction of the surgeon's subsequent actions, using low-level information only. Such a prediction system is critical for OR management: it will provide useful real-time information to the surgical team (nurses, anesthetist, junior surgeon), while allowing the surgeon to focus on more demanding tasks. Because predicting the next surgical task is central, such a prediction system will also be a keystone to the development of many other systems.

Learning to predict the future from past observations is one of the key components that make it possible to bring value to the massive data stores that have been collected in medicine. For example, the system proposed by Liu et al. [7] has already proven its usefulness to predict patient information (e.g. blood count panel) from patient electronic health records. This system use a hierarchical dynamical system and two modeling approaches: linear dynamical systems and gaussian processes to support predictive inferences. In the context of surgery, Franke et al. [8] proposed a system for intervention time prediction from surgical low-level tasks. A surgical process model optimized for time prediction was designed together with a prediction algorithm. The predictions were used to support intervention scheduling and resource management. In all these applications, the performed predictions are used by medical teams to improve the quality and the efficiency of healthcare.

In this paper, we focus on predicting the next surgical actions from the low-level information that can be captured during the surgery (*e.g.*, [3,9,10]). We use the series of surgical activities performed by the surgeon to represent the course of the surgery. We capture the activity of both hands for three different elements: *used instrument*, *performed action* and *targeted anatomical structure* [11]. Learning to predict the next activity of the surgeon from such low-level information is extremely challenging, because the next surgical action depends upon high-level information (phase of the surgery, technique used, patient-specific information, so-far reaction of the patient to the surgery, etc.), while a surgery is represented by a series of actions like "*cut,scalpel,skin*".

Intuitively, our approach matches the on-going surgery to every surgery of a reference set of surgeries, and use the next actions that have been performed in the reference set of surgeries to draw a prediction about the next action that will be performed in the current surgery. Our approach includes the three following features:

1. **Optimal registration of a partial surgery:** We propose a method to optimally register the on-going surgery (partial surgery) to any complete pre-recorded surgery. Our approach is based on the Dynamic Time Warping similarity measure [12], which is consistent with surgical processes [13].

2. **Voting for high-confidence prediction:** Using the optimally registered reference set of surgeries, we use voting to draw a high-confidence prediction about the next action that going to be performed by the surgeon.

3. **Detecting when to predict with high-confidence:** Using the agreement rate among multiple predictors, we are able to detect when to perform a prediction and when it is not possible to draw an accurate prediction.

Our framework is assessed using clinical data of lumbar disc herniation surgeries. Dr. L. Riffaud recorded 24 surgeries performed by multiple surgeons as part of a stay at the Neurosurgery Department of the Leipzig University Hospital, Germany. We show that our method outperforms the state of the art by providing a prediction more than 85% of the times with a 95% accuracy.

This paper is organized as follows. In Section 2, we present our solution for high-confidence prediction of the next surgical activity that is going to be performed. In Section 3, we conduct experiments that demonstrate the quality and performance of our approach compared to the state of the art. Finally, we conclude this work and describe future research in Section 4.

2   High-confidence prediction of the next surgical activity

We present our proposed approach in this section. We start by presenting our method for optimal sub-sequence matching in Section 2.1. We then present how we use this method to draw high-confidence predictions about the next surgical action in Section 2.2.

2.1   Optimal sub-sequence matching

Let $\mathbb{S} = \{S_1, \cdots, S_N\}$ be the reference set of N sequences (surgeries), $S = \langle s_1, \cdots, s_l \rangle$ be one sequence of this set (a complete surgery), and $S^\star = \langle s_1^\star, \cdots, s_k^\star \rangle$ be a partial sequence (the ongoing surgery). Let us denote $S_{1,l'}$ a sub sequence $\langle s_1, \cdots, s_{l'} \rangle$ of S. Our objective is to find the sub-sequence $S' = S_{1,l'}$ so that the cost of optimally registering the partial sequence $S^\star$ onto the subsequence S' of the reference sequence S is minimal.

Finding the cost of an optimal registration of one sequence onto another has been studied by the literature. The Dynamic Time Warping (DTW) similarity measure [12] makes it possible to find the optimal alignment of two sequences (and thus register them) in $\Theta(l_1 \cdot l_2)$ operations (with l_1 and l_2 the respective lengths of the realigned sequences). The consistency of this measure has been demonstrated for surgical processes in [13,14].

In this section, we 1) introduce a new objective function for finding the sub-sequence S' that best matches $S^\star$, and 2) introduce a new algorithm, based on DTW, that can find S' in $\Theta(k \cdot l)$ operations only.

Objective function Our goal is to find the matching point l' in S that minimizes the optimal alignment between $S^\star$ and the sub-sequence $S_{1,l'}$:

$$match(S^\star, S) = \underset{1 \leqslant l' \leqslant l}{\arg\min} \ \ \mathrm{DTW}(S^\star, S_{1,l'}) \tag{1}$$

Figure 1 presents the intuition about our objective function, compared to DTW's one. Figure 2 presents the trend of this objective function versus the value taken by l' on an example.
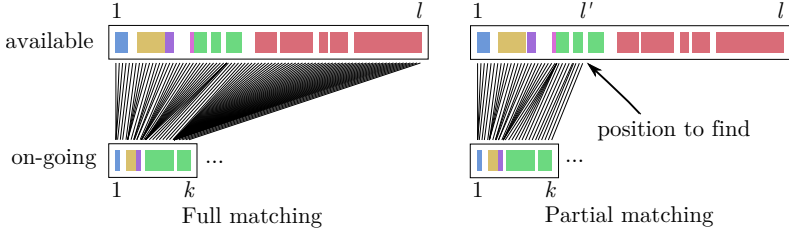

Fig. 1. Illustration of the difference between a full (left) and partial (right) matching.
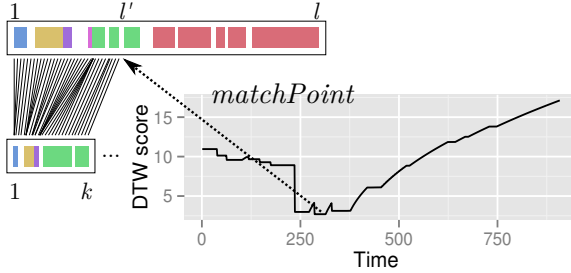

Fig. 2. Illustration of the *matchPoint* resulting of the partial matching.

Efficient algorithm An exhaustive search among all the possible matching points for l' will take $\Theta(\frac{l \cdot (l+1)}{2} \cdot k) = \Theta(l^2 \cdot k)$ operations. Such a cubic complexity with the length of the matched sequences is incompatible with real-time matching, because a typical surgery will often have more than 10,000 elements long.

We now show how to modify the Dynamic Time Warping (DTW) algorithm to obtain an exact solution in $\Theta(l \cdot k)$ operations without sacrificing the soundness of the process. Note that with 10,000 elements, the difference in the complexity corresponds to an algorithm running 4 orders of magnitude faster than the naive solution. Our solution is presented in Algorithm 1, where we adapted the original DTW algorithm to identify the optimal *matchPoint* (l') during sequences registration. In this algorithm, we kept the core of DTW and we added a condition (*i.e.*, the *if* statement) allowing to store the optimal *matchPoint* during the computation of the matrix storing the partial costs.

Note that although this algorithm can be further optimized depending on δ (*i.e.*, the distance function between elements of the sequences), we chose here

---

Algorithm 1 Optimal sub-sequence matching

Require: $S^\star = \langle s_1^\star, \cdots, s_k^\star \rangle$
Require: $S = \langle s_1, \cdots, s_l \rangle$
Let δ be a similarity between the elements of the sequences
Let $m[k, l]$ be a matrix storing partial costs
Let $l' \leftarrow 1$ be the matching point to find
  $m[1, 1] \leftarrow \delta(s_1^\star, s_1)$
  for $i \leftarrow 2$ to k **do** $\{m[i, 1] \leftarrow m[i - 1, 1] + \delta(s_i^\star, s_1)\}$
  for $j \leftarrow 2$ to l **do** $\{m[1, j] \leftarrow m[1, j - 1] + \delta(s_1^\star, s_j)\}$
  for $j \leftarrow 2$ to l **do**
    for $i \leftarrow 2$ to k **do** $\{m[i, j] \leftarrow \delta(s_i^\star, s_j) + \min(m[i - 1, j], m[i, j - 1], m[i - 1, j - 1])\}$

    if $m[k, j] < m[k, l']$ **then** $l' \leftarrow j$
  end for
  return l'

---

to give the algorithm for the general case. Furthermore, this adaptation of the algorithm did no alter the properties of optimality of DTW.

2.2   A voting approach to draw high-confidence predictions

Our method uses the proposed optimal sub-sequence matching to draw predictions about the next surgical activity that will be performed. We will use the optimal sub-sequence matching from the on-going surgery $S^\star$ to every sequence S_i of $\mathbb{S}$. We can then use this information to draw a probability distribution $\hat{p}_{\text{next}}$ over the next possible state of the current surgery. More formally, the maximum likelihood estimate $\hat{p}_{\text{next}}$ for the next activity to be s given the previous activities $S^\star$ is:

$$\hat{p}_{\text{next}}(s|S^\star) = \frac{|\{S(match(S^\star, S) + 1) = s\}_{S \in \mathbb{S}}|}{|\mathbb{S}|} \tag{2}$$

Finally, we draw a prediction from the maximum a posteriori estimate of $\hat{p}_{\text{next}}$ using a majority vote [15], *i.e.*, select s for which $\hat{p}_{\text{next}}(s|S^\star) \geqslant 0.5$. In order to ensure and confer high-confidence to the system, we do not draw a prediction if no s obtains a majority or in case of ties. Note that $\hat{p}_{\text{next}}(s|S^\star)$ can be seen as an agreement rate on the prediction: a high value indicates an important agreement amongst the recorded surgeries about the next action that is going to be performed and conversely. The threshold on the agreement rate (0.5 in this paper) can be tuned according to need for a system performing very accurate predictions but in limited number or a large number of predictions with an increased probability of errors. We have developed a web application [1] to allow the reader to try this prediction system easily. An open-source standalone implementation of the method is accessible at the same URL.

---

[1] http://germain-forestier.info/src/aime2015/ (Accessed: 30 March 2015)

3   Experiments

3.1   Clinical data

Figure 3 presents the dataset. The framework is evaluated using clinical data composed of 24 lumbar disc herniation surgeries recorded at the Neurosurgery Department of the Leipzig University Hospital, Germany. Surgeries contain on average 680 actions. The surgeries involved 10 male and 14 female patients, with a median age of 52 years. These lumbar disc surgeries are divided into three main steps: (1) approach of the disc, (2) discectomy and (3) closure. The herniated disc is approached via a posterior intermyolamar route. The patients were operated on by five junior and five senior surgeons. Senior surgeons have performed at least a hundred removals of lumbar disc herniation. All the junior surgeons have passed more than two years of their residency program but have only performed a few removals of lumbar disc herniation. In this paper, we focused on the closure phase, because it allows us to ensure that the main surgeon is the one operating (for a junior surgery, his or her senior sometimes takes over the surgery).
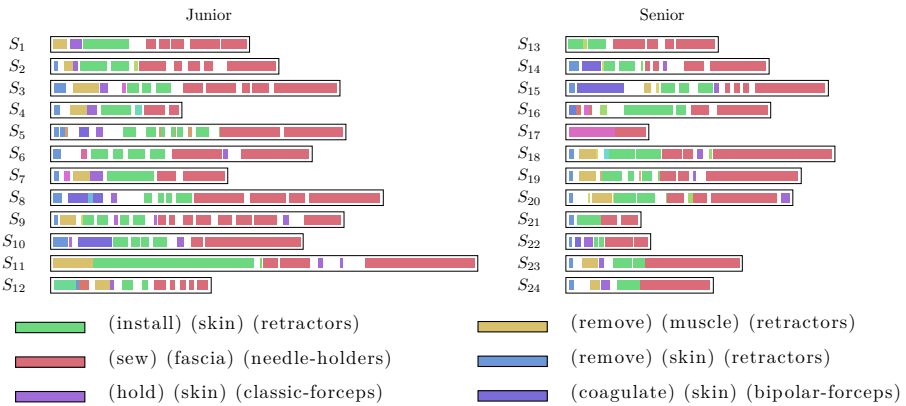

Fig. 3. The dataset of 24 surgeries used for the experiments and the legend for the six most frequent actions.

3.2   Methodology

We compare three configurations: using only the senior surgeries, using only the junior surgeries and using all surgeries. Our aim is to observe the influence of the available surgeries (training data) on the quality of the predictions that are drawn. A leave-one-out cross-validation approach was used for each configuration: we select one surgery out of the set of surgeries, and use it as the on-going intervention (this surgery is then removed from the set of reference surgeries). The left-out surgery is used to test our predictions, as if it was progressively discovered. Predictions are made every 5% of the progression of the intervention.

We can then compare every prediction with the *actual* activity of the surgery. Every surgery is in turn considered as the on-going intervention.

We evaluate our system using the precision $\mathcal{P}$ (*i.e.*, number of good predictions / total number of predictions), the recall $\mathcal{R}$ (*i.e.*, number of predictions / total number of expected predictions). We also use the F-measure $\mathcal{F}$ (harmonic mean between prediction and recall) to provide an overall evaluation. We compare the results of our method to the one of the Euclidean state-of-the-art method. We use the exact same process, but replace the optimal sub-sequence matching with uniform scaling [16]. Uniform scaling performs a linear transformation that increases or decreases sequences by a scale factor so that they have the same length.

3.3   Results

Figure 4 presents the general results of the F-measure comparing the two methods and the three configurations. We can see that our approach outperforms the
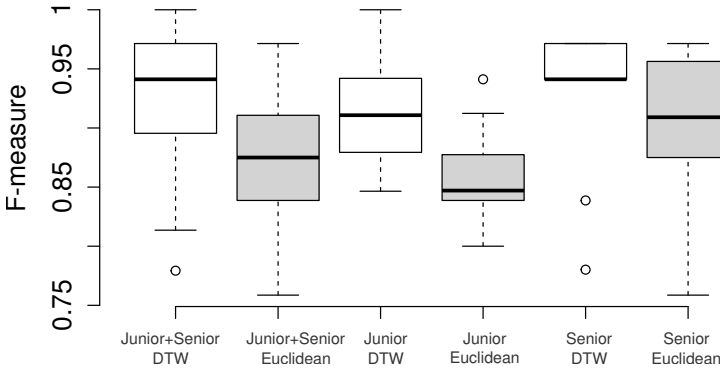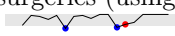

Fig. 4. Results on the three configurations (Junior+Senior, Junior, Senior) for the two methods (DTW in white, Euclidean in gray).

state-of-the-art Euclidean approach, regardless of the considered configuration. The compact dispersion of the results for the senior case (compared the junior case) suggests that seniors have a more homogeneous behavior than junior surgeons, which is consistent with previous studies comparing junior and senior practices [13]. This result also illustrates the influence of the set of available recordings in the quality of the prediction. Even though mixing all the surgeries together provides very good results, the best results are obtained for senior surgeons, whose surgical practice is usually more standardized and homogeneous. This supports our intuition that the more dedicated the training data is to the operating surgeon, the more accurate the predictions will be.

Table 1 details the prediction results for every one of the 24 surgeries (using the 23 remaining surgeries as the training set). A sparkline (*e.g.*,) presents, for each sequence, the evolution of the agreement rate among the predictions over the course of the surgery. The gray rectangle represents the interval

(0.5, 1] for which a majority is obtained. The blue dots represent the cases where our system did not provide a prediction (because no majority was obtained), while the red dots represent the inaccurate predictions. The precision of our

Table 1. Detailed results for every surgery; results with $\mathcal{F} \geqslant .9$ are shown in boldface.

	Junior					Senior			
Surg.	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Agreement	Surg.	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Agreement
S_1	1.00	0.94	**0.97**		S_{13}	1.00	0.89	**0.94**	
S_2	0.94	0.89	**0.91**		S_{14}	0.94	0.94	**0.94**	
S_3	0.85	0.72	0.78		S_{15}	1.00	0.83	**0.91**	
S_4	1.00	0.94	**0.97**		S_{16}	1.00	0.72	0.84	
S_5	0.94	0.94	**0.94**		S_{17}	1.00	0.83	**0.91**	
S_6	1.00	0.94	**0.97**		S_{18}	1.00	0.89	**0.94**	
S_7	0.88	0.94	**0.91**		S_{19}	1.00	0.94	**0.97**	
S_8	0.94	0.94	**0.94**		S_{20}	0.93	0.83	0.88	
S_9	0.88	0.89	0.88		S_{21}	1.00	0.94	**0.97**	
S_{10}	1.00	0.83	**0.91**		S_{22}	0.94	0.94	**0.94**	
S_{11}	1.00	1.00	**1.00**		S_{23}	1.00	1.00	**1.00**	
S_{12}	0.93	0.78	0.85		S_{24}	0.75	0.89	0.81	

system is very high: no mistakes are committed for more than half. Overall, our systems exhibits an average precision of 95%: our predictions do not eventuate only 5% of the times.

Moreover, our system provides a prediction 89% of the times (recall). This means that for the vast majority of cases, an agreement can be reached and a decision made. Furthermore, the consistency of our voting procedure is confirmed: for all the cases where the MAP (maximum a posteriori) estimate was below the majority threshold, and for which we thus did not provide a prediction (*i.e.*, blue dots in Table 1 – *Agreement* column), the MAP estimate was actually wrong. This confirms the relevance of our approach, by showing that we actually do not provide a prediction when no reliable choice can be made from the training set. This corresponds to the case where not enough similarity can be found between the on-going surgery and the reference set, which can be the case if specific activities are required during the surgery. The highest number of errors are committed in S_{24} with a sequence of four wrong predictions in a row. This corresponds to the green activity in Figure 3, where the surgeon installs the retractors on the skin without stopping, while all the other surgeries exhibit several pauses. Finally, every prediction is made in less than 200 ms, which is compatible with real-time prediction in the OR.

Note that the current system is dependent of the heterogeneity of the sequences inside the reference database. If the reference sequences are highly heterogeneous, the system could have difficulties to perform the partial matching. As the size of the available reference set is limited, we are currently matching all the sequences of the reference database with the target sequence. However, a

threshold on DTW score could be used to select only the most similar sequences to perform the prediction.

4   Conclusion

We have presented a method for the prediction of the next surgical task that is going to be performed during the surgery. Our contributions include:

1. definition of the objective function for the registration of a partial sequence to a complete reference sequence.
2. an efficient algorithm, based on DTW, to optimally minimize the above-mentioned objective function.
3. a prediction system that combines our optimal sub-sequence matching with MAP estimation and filtering.

The experiments have shown that our method outperforms the state of the art and provides a prediction more than 85% of the times with a 95% accuracy.

Because the prediction of surgical tasks is central to computer assisted surgery, this work naturally opens up a number of clinical applications. We have mentioned in the introduction how this information can help ensuring a smooth running of the surgery. Another application concerns the training of junior surgeons, where our system could be integrated in a simulation environment in order to provide help and feedback to the junior surgeon [17]. Our system could, on demand, provide a warning to the surgeon about his or her deviation from the standard practice of his or her colleagues. The agreement rate would then inform about the importance of the deviation. In future work, we want to validate this method on a more important dataset (> 300 surgeries) and use our recent work on Dynamic Time Warping [18] to improve the predictions.

Supplementary materials

Prediction package: Java package containing the source code for the proposed method. (Java ARchive file) – `http://germain-forestier.info/src/aime2015/source-code-aime-2015.jar` (Accessed: 30 March 2015)

Acknowledgments

References

1. Haynes, A.B., Weiser, T.G., Berry, W.R., Lipsitz, S.R., Breizat, A.H.S., Dellinger, E.P., Herbosa, T., Joseph, S., Kibatala, P.L., Lapitan, M.C.M., et al.: A surgical safety checklist to reduce morbidity and mortality in a global population. New England Journal of Medicine **360**(5) (2009) 491–499

2. 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. Volume 6135 of LNCS. Springer (2010) 34–44

3. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: An application-dependent framework for the recognition of high-level surgical tasks in the OR. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer (2011) 331–338

4. Meißner, C., Meixensberger, J., Pretschner, A., Neumuth, T.: Sensor-based surgical activity recognition in unconstrained environments. Minimally Invasive Therapy & Allied Technologies (0) (2014) 1–8

5. Lalys, F., Jannin, P.: Surgical process modelling: a review. International journal of computer assisted radiology and surgery **8**(5) (2013) 1–17

6. Forestier, G., Petitjean, F., Riffaud, L., Jannin, P.: Non-linear temporal scaling of surgical processes. Artificial Intelligence in Medicine **62**(3) (2014) 143 – 152

7. Liu, Z., Hauskrecht, M.: Clinical time series prediction with a hierarchical dynamical system. In: Conference on Artificial Intelligence in Medicine. Springer (2013) 227–237

8. Franke, S., Meixensberger, J., Neumuth, T.: Intervention time prediction from surgical low-level tasks. Journal of Biomedical Informatics **46**(1) (2013) 152 – 159

9. Ruda, K., Beekman, D., White, L.W., Lendvay, T.S., Kowalewski, T.M.: Surgtrak – a universal platform for quantitative surgical data capture. Journal of Medical Devices **7**(3) (2013) 030923

10. Ahmidi, N., Gao, Y., Béjar, B., Vedula, S.S., Khudanpur, S., Vidal, R., Hager, G.D.: String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI. Springer (2013) 26–33

11. Mehta, N., Haluck, R., Frecker, M., Snyder, A.: Sequence and task analysis of instrument use in common laparoscopic procedures. Surgical endoscopy **16**(2) (2002) 280–285

12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **26**(1) (1978) 43–49

13. 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) (2013) 822 – 829

14. Forestier, G., Lalys, F., Riffaud, L., Trelhu, B., Jannin, P.: Classification of surgical processes using Dynamic Time Warping. Journal of Biomedical Informatics **45**(2) (2012) 255–264

15. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3) (1998) 226–239

16. Yankov, D., Keogh, E., Medina, J., Chiu, B., Zordan, V.: Detecting time series motifs under uniform scaling. In: International Conference on Knowledge Discovery and Data mining, ACM (2007) 844–853

17. Zhou, Y., Bailey, J., Ioannou, I., Wijewickrema, S., O'Leary, S., Kennedy, G.: Pattern-based real-time feedback for a temporal bone simulator. In: Symposium on Virtual Reality Software and Technology, ACM (2013) 7–16

18. Petitjean, F., Forestier, G., Webb, G., Nicholson, A., Chen, Y., Keogh, E.: Dynamic Time Warping averaging of time series allows faster and more accurate classification. In: IEEE International Conference on Data Mining. (2014)