# A Boosted Segmentation Method for Surgical Workflow Analysis

N. Padoy[1,2], T. Blum[1], I. Essa[3], H. Feussner[4], M-O. Berger[2], and N. Navab[1]

[1] Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Germany
[2] LORIA-INRIA Lorraine, Nancy, France
[3] College of Computing, Georgia Institute of Technology, Atlanta, USA
[4] Chirurgische Klinik und Poliklinik, Klinikum Rechts der Isar, TU Munich, Germany
Nicolas.Padoy@cs.tum.edu

**Abstract.** As demands on hospital efficiency increase, there is a stronger need for automatic analysis, recovery, and modification of surgical workflows. Even though most of the previous work has dealt with higher level and hospital-wide workflow including issues like document management, workflow is also an important issue within the surgery room. Its study has a high potential, e.g., for building context-sensitive operating rooms, evaluating and training surgical staff, optimizing surgeries and generating automatic reports.

In this paper we propose an approach to segment the surgical workflow into phases based on temporal synchronization of multidimensional state vectors. Our method is evaluated on the example of laparoscopic cholecystectomy with state vectors representing tool usage during the surgeries. The discriminative power of each instrument in regard to each phase is estimated using AdaBoost. A boosted version of the Dynamic Time Warping (DTW) algorithm is used to create a surgical reference model and to segment a newly observed surgery. Full cross-validation on ten surgeries is performed and the method is compared to standard DTW and to Hidden Markov Models.

## 1 Introduction and Related Work

Workflow analysis related to business processes like document and record management, patient throughput and scheduling within hospitals, has been a well-established topic over the last decade[1].In recent years, workflow monitoring inside the Operating Room (OR) has gained more attention[2,3]. Automatic recovery and analysis of a surgical workflow will help designing future ORs, specialized for certain surgeries and capable of providing context-sensitive user interfaces as well as automatic report generation and monitoring. Furthermore, systems dedicated to the training and the evaluation of the surgical staff may also benefit from automatic workflow analysis.

High-level approaches deal with abstract representation of surgeries. Jannin et al. present in [4] a Unified Modeling Language (UML) diagram for multimodal neurosurgical procedures. They use it to improve multimodal information management as well as surgical planning. This model was further used in Raimbault

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part I, LNCS 4791, pp. 102–109, 2007.

et al.[5] to build a database of surgical cases, which can be queried to take benefit from past surgical experience. In [6], Neumuth et al. propose a system using business processes modeling to formalize and facilitate the abstract recording of a huge amount of surgeries by an operator in the surgical room. While such works pave the way for the statistical analysis of surgical workflow, they do not provide a direct representation in terms of surgical signals, as would be required for monitoring.

Other approaches focus on the analysis of dedicated surgical gestures. In [7], based on the torque/force signals provided by the Da Vinci robot, Lin et al. propose a method to recognize the elementary movements of a suturing task. Linear discriminants analysis is used in combination with a Bayes classifier to segment the motion. In Rosen et al.[8] the statistics of a surgical movement are analyzed for surgeon evaluation. The torque/force signals of the laparoscopic instruments recorded during a suturing task are learned with Hidden Markov Models (HMMs) in order to classify the skill level of the performing surgeon. To assess the quality of a surgical movement, Leong et al.[9] use the 3D trajectory of tracked laparoscopic instruments. The view invariant representations of the trajectories are evaluated with HMMs.

We present a complementary approach with an objective of automatically segmenting a *complete* surgery into phases using *live* signals from the OR. Our method is based on the temporal synchronization of multidimensional feature vectors to an average reference surgery. The algorithm is evaluated on the example of laparoscopic cholecystectomy, whose goal is to remove the gallbladder. This is a rather common but also complex surgery comprising many surgical phases. Even though the surgery depends in the details on the patient's anatomy, the surgeon follows a protocol consisting of 14 phases starting with the insertion of the trocars up till the suturing phase. These phases are illustrated in table 1.

While our algorithm is not limited to the use of a certain kind of features, we use binary vectors indicating instrument presence during the surgery. Many other signals would be available from the OR. However, we focus in this work on the usage of the surgical tools since it describes well the underlying workflow of a laparoscopic operation. The method is based on a modification of the Dynamic Time Warping (DTW) algorithm, which is applied with an adaptive distance measure. The measure is defined from the discriminative power of each instrument with respect to the current surgical phase, estimated by AdaBoost[10]. Widely used for feature selection[11], AdaBoost provides a natural way for feature weighting. This information is combined with temporal synchronizations to create an average model out of labeled training surgeries. Finally, the adaptive version of DTW is used to synchronize an unsegmented surgery to the model. Using this synchronization, labels from the average model can be carried over to an unsegmented surgery.

In an early work[12], DTW was used to synchronize several surgeries together in order to create an average model without any a-priori knowledge. The focus was set on surgical synchronization and results were evaluated in terms of simultaneous video visualization. For segmentation, we present in this work a

**Table 1.** The fourteen phases labeling each surgery

| | | | |
|---|---|---|---|
| 1 | CO2 inflation | 8 | Liver Bed Coagulation 1 |
| 2 | Trocar Insertion | 9 | Packaging of Gallbladder |
| 3 | Dissection Phase 1 | 10 | External Gallbladder Retraction |
| 4 | Clipping Cutting 1 | 11 | External Cleaning |
| 5 | Dissection Phase 2 | 12 | Liver Bed Coagulation 2 |
| 6 | Clipping Cutting 2 | 13 | Trocar Retraction |
| 7 | Gallbladder Detaching | 14 | Abdominal Suturing |

learning-based method. This new approach shows significant improvements and is evaluated with a complete cross-validation on a set of 10 surgeries. It is also compared to standard DTW without weights and to HMMs.

## 2 Methods

### 2.1 Overview

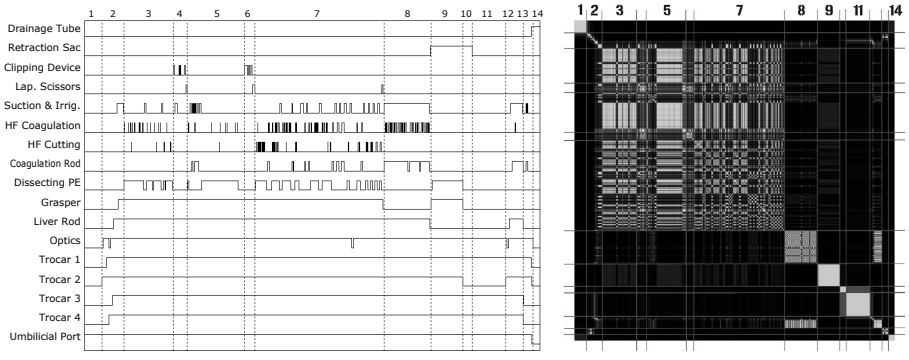
We first introduce the representation of the acquired signals in section 2.2. It is followed by the derivation of the weights per instrument and phase in section 2.3. The Adaptive Dynamic Time Warping (ADTW) algorithm is introduced in section 2.4. In the same section we describe its use for the segmentation of a new surgery. Finally, the computation of the average surgical model is presented in section 2.5.

### 2.2 Instrument Signals

In minimally-invasive surgeries the instruments strongly correlate with the underlying surgical workflow. To record the surgical actions during the procedure, instrument presence is acquired for $K = 17$ laparoscopic instruments and represented as a multivariate time series $\mathbb{I}$ where $\mathbb{I}_t \in \{0,1\}^K$ :

$$\mathbb{I}_{t,k} = 1 \quad \textbf{iff} \quad \text{instrument } k \text{ is used at time } t$$

The instrument signals for an exemplary operation are displayed in fig. 1(a). The vertical lines display the segmentation in phases. While several phases can be simply characterized by a few instruments, in the others the relation phase/instruments is more complicated (for instance for phases 3 to 7). This is well illustrated by fig. 1(b) where the self-similarity matrix of the temporal vector sequence is displayed. The similarity matrix $M$ is here defined by $m_{t_1,t_2} = exp^{-d(\mathbb{I}_{t_1},\mathbb{I}_{t_2})}$. The phases are marked by vertical and horizontal lines on the matrix. For phases involving very specific instruments, a distinctive block appears on the diagonal, while for phases involving the same instruments blocks are harder to identify. Note the distinctive blocks bottom-right off the diagonal for the liver bed coagulation phases, indicating their strong correlation as they use almost the same instruments (phases 8 and 12).




(a) Instrument presence (b) Self similarity matrix

**Fig. 1.** Temporal sequence of instrument vectors for one surgery, and its self similarity matrix, showing the amount of instrument usage similarity between all phases

### 2.3 Weighting Method

The instruments occurring within a phase vary and are generally not sufficient to characterize the phase, as the temporal sequence of actions often plays a decisive role. But the instruments can be weighted to reflect their ability to discriminate between neighboring phases. When synchronizing a surgery to the average reference model, using those weights, the ADTW algorithm will put a higher priority on the most significant instruments for each phase.

AdaBoost[10] builds a strong classifier out of a sum of weak classifiers. They are iteratively chosen to optimally classify weighted training data and are themselves weighted accordingly. For each phase $p$, a strong classifier trying to classify all the instrument vectors of the phase with respect to all the vectors of the neighboring phases is built. By choosing the pool of weak classifiers to be simply related to the instruments, weights for the instruments can be naturally derived from the strong classifier.

The weak classifiers are chosen to perform the classification based on the presence/absence of a single instrument: a simple weak learner $C_{n,x}$ classifies an instrument vector according to whether the state of the instrument $n$ within the vector is equal to $x$. AdaBoost selects at each step $i$ a classifier $C_{n_i,x_i}$ and a weight $\alpha_i$ to construct the strong classifier:

$$SC = \sum_i \alpha_i C_{n_i,x_i}$$

The variable $n_i$ and $x_i$ indicate the instrument and its state that were selected at step $i$. As the algorithm reweights the data that was hard to classify, the selected weak classifiers are the most important for the classification. The weights are obtained by looking at the influence of each instrument $k$ within the strong classifier:

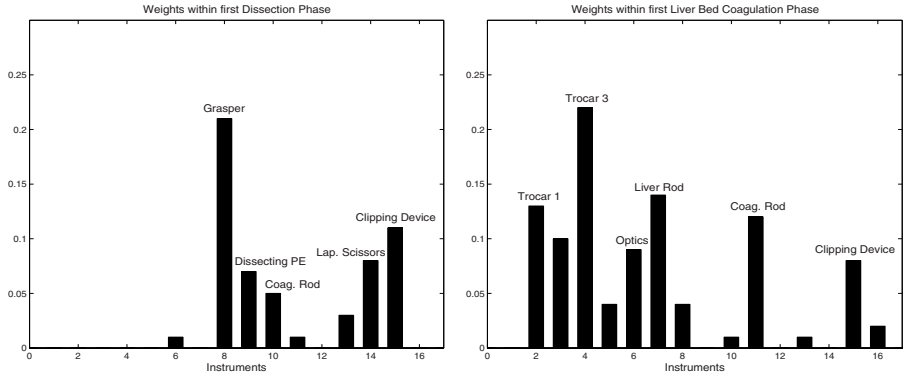



**Fig. 2.** Instrument weights computed for the first dissection phase (phase 3) and the first liver bed coagulation phase (phase 12)

$$w_k = \left| \sum_{n_i=k, x_i=1} \alpha_i - \sum_{n_i=k, x_i=0} \alpha_i \right|$$

They are then normalized to one. As they are computed for each phase, this leads to weights $w_k^{(p)}$, for all phase $p$ and instrument $k$. Depending on the phase, the convergence of AdaBoost requires a few to several dozens of steps. As some phases are very short, better results are obtained by classifying the phases with respect to the two previous and the two next phases. Fig. 2 displays the computed weights for two phases. In the first dissection phase, the three most significant instruments are found to be the grasper, which has to be present, and the clipping device and laparoscopic scissors, which have to be absent. In the first liver bed coagulation phase, they are the trocar 1 and 3 as well as the liver rod, which all have to be present.

### 2.4 Adaptive DTW

The Dynamic Time Warping algorithm[13] is both a time-invariant similarity measure and a method to synchronize two time series by finding a non-linear warping path. It has to warp each point in one time series onto at least one point in the other time series while respecting the temporal order. This is done in a way to minimize the sum of the distances between all points that are warped onto each other. It has been applied in various domains to synchronize series of application-dependent feature vectors[14,15]. As the length of a phase varies highly between different OPs, depending on the patient anatomy and the surgeon ability, the synchronization of a surgery to the model is also highly non-linear.

Traditional DTW computes the distance between the two series with a fixed distance function. As our reference time series is segmented in phases, we propose to use a distance function which is phase-dependent, so as to involve mainly instruments which are discriminative for a phase.

We define for each phase $p$ the weighted distance $d_p$ between instrument vectors $v_1$ and $v_2$:

$$d_p(v_1, v_2) = \sqrt{\sum_{k=1}^{k=K} w_k^{(p)} (v_{1,k} - v_{2,k})^2}$$

To compute the ADTW, within the dynamic time warping algorithm the distance function corresponding to the known phase of the reference series is used.

By warping an unsegmented surgery onto a segmented reference surgery, we can carry over the segmentation. As reference we use a model of an average surgery whose creation is described below.

### 2.5 Average Model Computation

Out of all training surgeries $\mathbb{O}_1 \dots \mathbb{O}_n$ an average surgery is computed. Let $\mathbb{P}_{ij}$ be the $i$th phase from surgery $\mathbb{O}_j$. The average phase $\overline{\mathbb{P}_i}$ is constructed as follows. Out of $\mathbb{P}_{i1} \dots \mathbb{P}_{in}$ the phase with length closest to the average length of this phase is chosen as initial average $\overline{\mathbb{P}_i}$. Next $\mathbb{P}_{i1} \dots \mathbb{P}_{in}$ are warped onto $\overline{\mathbb{P}_i}$ using DTW with the weighted distance $d_i$ of the current phase. Next, $\overline{\mathbb{P}_i}$ is updated by taking the average of the warped versions of $\mathbb{P}_{i1} \dots \mathbb{P}_{in}$.

These two steps are repeated iteratively until convergence of $\overline{\mathbb{P}_i}$. In the final step, the average surgery is built by simply concatenating all average phases $\overline{\mathbb{P}_1} \dots \overline{\mathbb{P}_{14}}$. While the training surgeries only consist of boolean values, stating whether an instrument is in use, the average can also contain non-boolean values. These can be interpreted as the probability of an instrument to be used at this moment.

## 3 Experiments and Results

For the experiments we use 10 surgeries of a cholecystectomy, labeled with 14 phases as described in the previous sections. One surgeon did 9 of the surgeries, where some parts have been performed by assistants. The 10th surgery has been done completely by another surgeon from the same school. A complete cross-validation has been performed, each time using 9 surgeries to compute weights with AdaBoost and construct the average surgery. The remaining surgery is then segmented using the three following algorithms for comparison: ADTW, standard DTW and HMMs. The standard DTW method is similar to ADTW, but all weights are set to be constant and equal. The labeled training information is thus only used in the creation of the reference model. For HMMs, the same amount of a-priori information is provided: left-right HMMs with fourteen states are used and trained on the labeled surgeries. The transition model is computed so that the expected state duration matches the average phase duration of the 9 surgeries in the training set. The observation model is computed from the usage frequency of each instrument within each phase, assuming instrument independence as this yields the best results. To evaluate the quality of the segmentation, we compute the following errors:

**Table 2.** Mean of the computed errors on all cross-validation tests

| | overall error | mean error per phase | max error per phase | skipped phases |
|---|---|---|---|---|
| HMM | 8.9% | 10.1% | 60.9% | 0.5 |
| standard DTW | 0.8% | 0.9% | 10.9% | 0 |
| ADTW | 0.3% | 0.3% | 4.5% | 0 |

- **overall error**: percentage of wrong segmentation labels in the complete surgery
- **mean error per phase**: percentage of wrong segmentation labels within a phase, mean on the 14 phases
- **max error per phase**: percentage of wrong segmentation labels within a phase, maximum on the 14 phases
- **skipped phases**: number of phases that have no overlap with their ground-truth.

The mean results on all cross-validation tests are displayed in table 2. The segmentation with HMMs provided the worst results. As a few phases are skipped in several surgeries, leading to a *max error per phase* of 100%, the resulting *max error per phase* in the table is very high. However, they still recognise 91.1% of the labels, with a time resolution of a second. None of standard DTW and ADTW provided skipped phases, but ADTW outperforms in mean DTW without adaptive weights by a factor greater than 2. It yields a very accurate segmentation for *all* phases, as the mean *max error per phase* is below 5%. The max *max error per phase* is 13.6% for ADTW, while it is 40.2% for standard DTW. Moreover, experiments with the 10 surgeries show the errors to decrease faster with ADTW than with standard DTW when the size of the training set grows. Finally, the surgery carried out by the second surgeon obtained also very good segmentation results.

## 4 Discussion and Conclusion

In this paper we presented a reliable way to automatically recognize the workflow of a laparoscopic operation using only little training data. We have shown that the laparoscopic instruments provide enough information to automatically segment fourteen procedural phases of laparoscopic cholecystectomies with a high success rate. To this end the laparoscopic instruments used in each phase are analyzed with AdaBoost and weighted according to their discriminative power. An adaptive dynamic time warping algorithm using those weights synchronizes the workflow to a reference model, yielding the segmentation.

While the segmentation is automatic after acquisition of the input information, up-to-now not all input signals are obtained automatically for practical reasons. Automatic signal acquisition is for example currently possible for instruments like the coagulation/cutting device or the optics. With the use of sensors, it would also be possible to get it for the others.

Experiments were carried out on 10 cholecystectomies and cross-validation proved the algorithm to outperform both standard DTW and HMMs. These

results are very promising and we believe they can apply to other kinds of laparoscopic surgeries. Examples of valuable by-products of this research are the automatic reporting of a surgical operation and/or the evaluation and comparison of trainees. Future work will focus on selecting appropriate input information that can be obtained automatically, so as to provide a fully automatic system and pave the way for surgical monitoring.

**Acknowledgments.** This research is partially funded by Siemens Medical Solutions.

## References

1. Dazzi, L., Fassino, C., Saracco, R., Quaglini, S., Stefanelli, M.: A patient workflow management system built on guidelines. In: AMIA 1997, pp. 146–150 (1997)
2. Herfarth, C.: 'lean' surgery through changes in surgical workflow. British Journal of Surgery 90(5), 513–514 (2003)
3. Cleary, K., Chung, H.Y., Mun, S.K.: Or 2020: The operating room of the future. Laparoendoscopic and Advanced Surgical Techniques 15(5), 495–500 (2005)
4. Jannin, P., Raimbault, M., Morandi, X., Gibaud, B.: Modeling surgical procedures for multimodal image-guided neurosurgery. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 565–572. Springer, Heidelberg (2001)
5. Raimbault, M., Morandi, X., Jannin, P.: Towards models of surgical procedures: analyzing a database of neurosurgical cases. In: Med. Imaging, SPIE, pp. 97–104 (2005)
6. Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H.U., Burgert, O.: Acquisition of process descriptions from surgical interventions. In: Bressan, S., Küng, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 602–611. Springer, Heidelberg (2006)
7. Lin, H.C., Shafran, I., Murphy, T.E., Okamura, A.M., Gregory, D., Hager, D.D.Y.: Automatic detection and segmentation of robot-assisted surgical motions. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 802–810. Springer, Heidelberg (2005)
8. Rosen, J., Solazzo, M., Hannaford, B., Sinanan, M.: Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. Comput Aided Surg. 7(1), 49–61 (2002)
9. Leong, J., Nicolaou, M., Atallah, L., Mylonas, G., Darzi, A., Yang, G.Z.: HMM Assessment of Quality of Movement Trajectory in Laparoscopic Surgery. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 752–759. Springer, Heidelberg (2006)
10. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
11. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV 57(2), 137–154 (2004)
12. Ahmadi, S.A., Sielhorst, T., Stauder, R., Horn, M., Feussner, H., Navab, N.: Recovery of surgical workflow without explicit models. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 420–428. Springer, Heidelberg (2006)
13. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. 26(1), 43–49 (1978)
14. Darrell, T., Essa, I.A., Pentland, A.: Task-specific gesture analysis in real-time using interpolated views. IEEE Trans. PAMI 18(12), 1236–1242 (1996)
15. Kassidas, A., MacGregor, J.F., Taylor, P.A.: Synchronization of batch trajectories using dynamic time warping. AIChE Journal 44(4), 864–875 (1998)