Statistical Modeling and Recognition of Surgical Workflow

Nicolas Padoy[*,a,1], Tobias Blum[b], Seyed-Ahmad Ahmadi[b], Hubertus Feussner[c], Marie-Odile Berger[d], Nassir Navab[b]

[a] *Engineering Research Center for Computer-Integrated Surgical Systems and Technology, Johns Hopkins University, Baltimore, Maryland, USA*
[b] *Computer Aided Medical Procedures (CAMP), Technische Universität München, Germany*
[c] *Department of Surgery, Klinikum Rechts der Isar, Technische Universität München, Germany*
[d] *LORIA-INRIA Lorraine, Nancy, France*

Abstract

In this paper, we contribute to the development of context-aware operating rooms by introducing a novel approach to modeling and monitoring the workflow of surgical interventions. We first propose a new representation of interventions in terms of multi-dimensional time-series formed by synchronized signals acquired over time. We then introduce methods based on Dynamic Time Warping and Hidden Markov Models to analyze and process this data. This results in workflow models combining low-level signals with high-level information such as predefined phases, which can be used to detect actions and trigger an event. Two methods are presented to train these models, using either fully or partially labeled training surgeries. Results are given based on tool usage recordings from sixteen laparoscopic cholecystectomies performed by several surgeons.

Key words: surgical workflow, context aware operating room, surgical assistance system, Hidden Markov Model, cholecystectomy

1. Introduction

The operating room (OR) needs to be constantly adapted to the introduction of new technologies and surgical procedures. A key element within this process is the analysis of the workflow inside the OR[7, 16]. It has impact on patient safety, working conditions of surgical staff and even overall throughput of the hospital. While new technologies add much complexity to the daily routine of OR staff, they also facilitate the design of assistance systems that can relieve the surgical staff from performing simple but time-consuming tasks and assist them in the tedious ones.

In this paper, we focus on the design of a context-aware system that is able to recognize the surgical phase performed by the surgeon at each moment of the surgery.

[*]Corresponding author. Email address: *padoy@jhu.edu*
[1]This work was carried out when Nicolas Padoy was at the institutions *b* and *d*.

We believe that robust real-time action and workflow recognition systems will be a core component of the Operating Room of the Future. They will enable various applications, ranging from automation of simple tasks to detecting failures, suggesting modifications, documenting procedures and producing final reports.

Our contribution is threefold: In the introduction we identify the generic need for an automatic recognition system and introduce a signal based modeling of the surgical actions to achieve such automation. We then propose two statistical models constructed from generic signals from the OR: the annotated average surgery and the annotated Hidden Markov Model, for off-line and on-line recognition of the surgical phases in a standard endoscopic surgery. The models are built based on a set of training surgeries where the phases have been labeled. We also propose a method only requiring partially labeled data. We finally demonstrate and evaluate the methods on the example of laparoscopic cholecystectomy using binary instrument usage information and illustrate their use in some potential applications: automatic generation of report sketches, automatic triggering of reminders for the surgical staff and automatic prediction of the remaining time of the surgery.

1.1. Motivation

In the last years, many experts have tried to predict which changes need to be made to the operating room in order to meet future requirements in terms of appearance, ergonomics and operability [33, 10, 3]. Aiming always at better patient treatment and higher hospital efficiency, issues and solutions have been further discussed in international workshops gathering clinicians, researchers and medical companies [7, 16]. The focus has been on surgical workflow optimization, system integration and standardization, in particular in new clinical fields like image guided surgery. Parallel to these studies, different testbeds have been established to experiment and report on the development of advanced surgery rooms [32, 1, 20].

A general consensus is that all systems present in the surgery room of the future should and will be fully integrated into a final system and network. For specific procedures, companies like BrainLab[2] and Medtronic[3] already provide fully integrated OR suites. In [22], a surgery room is presented where various contextual informations are presented on a unified display. Such ORs, where a multitude of signals and information are available within a unique computer system, offer great opportunities for the design of powerful context-aware systems which will be reactive to the environment.

Signals are already provided by anesthesia systems, medical and video images and digital patient files. In addition, more signals from advanced electronic surgical tools and navigation systems that are tracking patients, clinical staff, surgeon and equipments will provide an extensive and rich dataset in the future. Note that the use of RFID tags is being widely investigated for tracking of material and persons inside the hospital and the OR [8, 29]. Another environment providing rich sensor information is robotic surgery. The availability of these signals on a daily basis will highly facilitate the analysis and recognition of all actions performed in the OR.

[2] `http://www.brainlab.com`
[3] `http://www.medtronic.com`

Context-awareness is not only profitable for assistance inside the OR. In many cases, OR delays come merely from poor synchronization between the workflows inside and outside the OR [11]. Incorporating context-aware ORs inside a global hospital awareness system would therefore greatly improve the overall efficiency.

In this work, we present several methods to recognize the surgical phases of a surgery which has a well-defined workflow. They can either be applied on-line during the surgery for context-awareness, or off-line after the surgery for instance for documentation and/or report generation.

1.2. Related Work

Close work comes from the robotic community, where awareness is required for robot control. [23, 38] target the development of a robotic scrub nurse that automatically provides the correct instrument to the surgeon. The approach is based on time-automata combined with a vision front-end. The conception of the model is however time-consuming as it is done by hand. A further limitation is that the prototype only works with a tiny set of instruments provided in a predefined order.
In [15], a task model of cholecystectomy is designed for guidance of a laparoscopic robot which is controlling the camera pose. A viewing mode is assigned to each surgical stage and transition rules between the stages are manually defined based on the currently used tool that is detected using color markers. A surgical stage cannot be always uniquely recognized from the current surgical tool. Therefore some ambiguities cannot be distinguished with this deterministic approach.

Also demonstrated on cholecystectomy, we have proposed in previous work approaches based on Dynamic Time Warping for segmenting the surgical phases of the surgery using laparoscopic tool usage [2, 26]. In [27], we proposed an on-line approach using Hidden Markov Models constructed from data containing visual cues computed from the endoscopic video. In [5], we addressed the automatic generation of human-understandable models of surgical phases. We propose in this paper a unified framework, address the case of data where the phases have only been labeled partially and present in more details the motivations and applications.

A statistical model, signals and a detection approach that relates the signals to the model are needed to recognize the surgical phase. Above references strive to provide a complete solution. Other existing work addresses one of these three aspects. In [13], a model based on Unified Markup Language (UML) is proposed in order to understand and optimize the usage of imaging modalities during a neurosurgical procedure. [24] presents ontologies and tools to describe and record surgeries in a formal manner. The actions and interactions occurring in surgeries can be recorded manually by assistants using a software that helps generating standardized descriptions, which can be in turn used for an in-depth analysis of the workflow. The method has been validated on a large dataset of simulated data [25]. This manual approach is complementary to ours in that it permits in depth formal description and understanding of the workflow despite the fact that only few sensors are currently available in the OR, as required for monitoring.

Interesting signals for the analysis of surgical gestures are the positions of tools or the forces applied to them. They can be obtained indirectly using a tracking system or directly when a robot is used and thus provides the positioning information. Such signals have been mainly used for evaluating and comparing surgeons performing on

a simulator [30, 21, 18, 17]. For recognition of several actions in minimally invasive surgery, [19, 34] use the endoscopic video. In [12], the surgeon's attention is tracked with an eye-gaze system to detect the clipping of the cystic duct during a pig cholecystectomy.

Beyond surgeries, different work is also addressing context awareness in clinical environment. For instance [37, 4] use either vital signs available from the anesthesia systems or an external camera to automatically detect when patients are entering or leaving the surgery room.

Dynamic Time Warping (DTW) [31] and Hidden Markov Model (HMM) [28] algorithms have emerged from the speech recognition community and have since been extensively used for classification in many domains. In the following, we do not use them for classification, but for constructing a statistical model of a surgical workflow, in which we intend to recognize the phases. This is permitted by the notion of an annotated model that we introduce.

1.3. Organization

The motivation for signal based workflow recognition was presented in section 1.1, related work in section 1.2. The remainder of this paper is as follows: the two core models are introduced in section 2. Their use in segmentation and recognition is given in section 3. Section 4 presents the medical application and an experimental comparison of their performance. Different applications of the off-line segmentation and on-line recognition as well as a discussion of the results are provided in section 5. Finally, conclusions are given in section 6.

2. Model construction

In this section, we explain how to construct two statistical models whose usage for phase recognition is described in section 3: the annotated average surgery and the annotated Hidden Markov Model.

2.1. Assumptions and Notations for Signals and Phases

We assume the signals from the operating room to be normalized and represented by a multidimensional time-series $\mathbb{O}$ where $\mathbb{O}_t \in [0,1]^K$. In the example of laparoscopic cholecystectomy which is used in our experiments and detailed later in section 4, the signals represent the instrument usage at a temporal resolution of one second and

$$\mathbb{O}_{t,k} = 1 \;\; \textbf{if and only if} \;\; \text{instrument } k \text{ is used at time } t \;\; . \tag{1}$$

In this case, $K = 17$. We assume the usage to be sufficient for inferring the actions performed by the surgeon as most surgical actions are performed by using an instrument. Any other additional signal, binary or not, can however be included into the modeling with only little further modification. A surgery room during an endoscopic operation and the view of the endoscopic camera are shown in fig. 1. Signals acquired during a surgical procedure are shown in fig. 2.

We define a phase as a subpart of the surgery that can be identified uniquely in all examples of the surgery and that is validated by a medical expert. We assume

Figure 1: Picture of a surgery room (left) during an endoscopic operation. View of the endoscopic video (right).

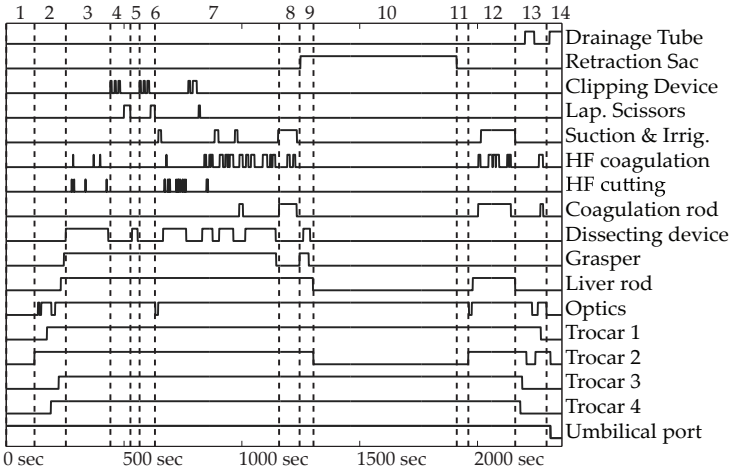

Figure 2: Signals recorded from a surgery.

the phases to be performed consecutively in time and in the same order, as this is the case in standard surgeries. In the current stage of our research, we are not explicitly handling exceptions in the workflow. This will be discussed in section 5.2. Phases contain actions, which themselves can occur repetitively in all phases. In case of binary instrument usage, a change in a signal of the time-series represents a transition to a new action. At the beginning, we assume the phases to be contiguous in time, meaning that each time step corresponds to a phase. This assumption can be relaxed when only subparts of the surgery need to be detected, as explained in section 2.5.

In the following, P is the number of phases and L the size of the training set. The surgeries in the training set are denoted by $\mathbb{O}^l$, where $l \in 1 \dots L$. A new surgery that has to be monitored will be denoted by $\mathbb{O}^{test}$. The phase at time t for a surgery $\mathbb{O}$ is referred to by $\mathcal{T}(\mathbb{O}, t)$.

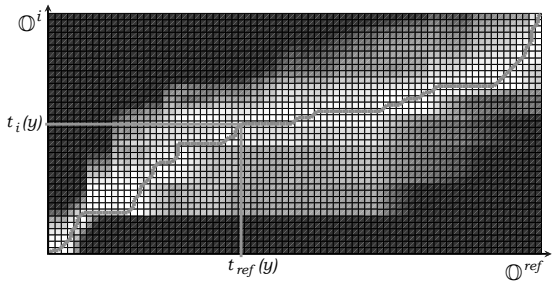

Figure 3: DTW distance matrix. The path drawn through the valley displays the optimal synchronization between surgeries $\mathbb{O}^i$ and $\mathbb{O}^{ref}$.

2.2. DTW averaging

A statistical model of the surgery, called the average surgery, is created based on a set of surgeries, the training set. The surgeries in the training set are all synchronized using the Dynamic Time Warping algorithm to generate an average surgery on an average timeline, preserving the average length of the surgeries, the phases and the actions. This average surgery can be used for off-line segmentation of new surgeries or for an efficient generation of the HMM models as will be explained later.

The averaging approach is based on [36]. In the following, we adapt the presentation to the time discrete case that is considered in this work. Let the surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^L$ be of length $T^1, \ldots, T^L$ and $\mathbb{O}^{ref}$ be a surgery taken as reference with length T^{ref}.

The Dynamic Time Warping algorithm [31] uses dynamic programming to perform a nonlinear synchronization between two time series so as to minimize their warped distance. The reference $\mathbb{O}^{ref}$ is first synchronized to each training surgery $\mathbb{O}^i$ yielding for $i \in \{1 \ldots L\}$ the synchronization functions

$$\text{sync}_{ref \to i}(y) = (t_{ref}(y), t_i(y)) \quad . \tag{2}$$

These functions give discrete correspondences between the timelines of the reference (t_{ref}) and of the surgeries (t_i). Due to the properties of the DTW, after a step on the discrete timeline from y to $y + 1$ either $t_{ref}(y)$ or $t_i(y)$ is incremented by one. An example of a synchronization function is represented graphically in fig. 3. The variable y represents a discretization of the optimal path through the valley.

The average timeline is computed from the reference as the function $\text{avg}_{time}(t)$:

$$\begin{aligned} \{1, \ldots, T^{ref}\} &\to [1, T^{mean}] \\ t &\to \frac{1}{L} \sum_{i=1}^{L} \frac{1}{\#\{y:\ t_{ref}(y)=t\}} \sum_{\{y:\ t_{ref}(y)=t\}} t_i(y) \quad , \end{aligned} \tag{3}$$

where $\#\cdot$ denotes the cardinality operator. The function avg_{time} takes real values, is monotonically increasing between 0 and $T^{mean} = \frac{1}{L} \sum_i T^i$ and can therefore be inverted. It is used to compute the average surgery $\mathbb{O}^{avg}$ on the discrete timeline $\{1, \ldots, \lfloor T^{mean} \rfloor\}$

using linear interpolation and averaging over all surgeries. Let t_a and t_b be the closest values in the range of avg_{time} around the integer t with $t \in [t_a, t_b]$. Denoting $\mathcal{Y}_a = \{y : \ t_{ref}(y) = \text{avg}_{time}^{-1}(t_a)\}$ and $\mathcal{Y}_b = \{y : \ t_{ref}(y) = \text{avg}_{time}^{-1}(t_b)\}$, we then define

$$\mathbb{O}_t^{avg} = \frac{1}{l} \sum_{i=1}^{l} \frac{(t_b - t)}{(t_b - t_a)} \frac{1}{\#\mathcal{Y}_a} \sum_{y \in \mathcal{Y}_a} \mathbb{O}^i_{t_i(y)} + \frac{(t - t_a)}{(t_b - t_a)} \frac{1}{\#\mathcal{Y}_b} \sum_{y \in \mathcal{Y}_b} \mathbb{O}^i_{t_i(y)} \ . \tag{4}$$

Similarly to [36], we use three steps for our average surgery computation:

1. Compute initial reference
2. Compute first average surgery
3. Iterate average surgery computation using previous average surgery as reference

There are various ways to choose the initial reference. It could simply be one of the training surgeries. However, when some actions have not been performed in this specific surgery, but in one or several of the other surgeries from the training set, it is likely that these actions are also not represented correctly in the average surgery. Experiments have shown that the following approach avoids this problem. First, the training surgeries are averaged pairwise using DTW. The resulting average surgeries are then again iteratively merged pairwise and the last one is taken as the initial reference.
An average surgery that has been computed using the method described above is shown in figure 4.

A straightforward extension to the average surgery is the notion of surgical similarity, which can be given for each observation vector in an average surgery. Let us define

$$SIM_t = \frac{1}{K} \sum_{k=1}^{K} max(\mathbb{O}_{t,k}^{avg}, 1 - \mathbb{O}_{t,k}^{avg}) \ . \tag{5}$$

When SIM_t is close to 1, it means that a reliable synchronization point between all surgeries was found. More intuitively, it means that the surgical activity for this time point was unambiguous across all training surgeries. In contrary, a value close to 0.5 implies ambiguity. This will be used later to efficiently construct HMMs models.

2.3. HMM modeling

A discrete Hidden Markov Model is a quintuplet $\lambda = (N, M, A, B, \pi)$ where N is the number of states $\{x_i : 1 \leq i \leq N\}$ in the model, M the size of the observation alphabet, A the transition probability matrix between the states, B the observation probability matrix and π a probability distribution over the initial states. HMMs offer a much more general statistical description of a process than the average surgery, especially by permitting non-sequential modeling and by explicitly including probabilities, as needed for on-line recognition. The relationship between HMM and DTW is explicated in [14].

In this work, observations are the vectors of the time-series representing the surgeries. We aim at constructing HMMs that represent one kind of surgery and also allow us to recognize on-line the phase carried out by the surgeon. Traditionally, e.g. in speech recognition [28], HMMs are used to model the stochastic properties of a training set of time series. The likelihood $P(\mathbb{T}|\lambda)$ that the time series $\mathbb{T}$ has been generated

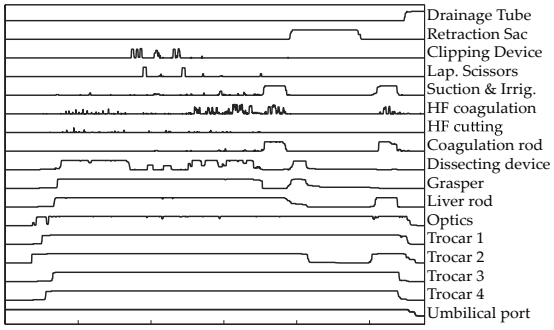

Figure 4: Signals of an average surgery.

by the HMM λ can for instance be used for classification if different HMMs are used to represent the classes. Given a training set of time-series $\{\mathbb{T}^i\}$, there exist many methods to initialize and train an HMM λ such that the log-likelihood $\sum_i logP(\mathbb{T}^i|\lambda)$ is maximized. The model construction described in section 2.4 will be based for comparison on the three following initialization methods:

- Fully connected HMMs *(HMM-full)*: the number of states is fixed, the parameters are randomly initialized and the model is trained by expectation-maximization (EM).
- Sequential HMMs *(HMM-seq)*: the number of sequentially connected states is inferred from the amount of training data, training sequences are split and then assigned to the states for initialization of the probabilities.
- Model merging *(HMM-merged)*: the method is described in [35]. An exhaustive topology is built out of all training data. Pairs of states are then chosen and merged together until the decrease of the log-likelihood exceeds a given threshold.

For the sequential HMMs construction, the number of states is chosen to be $\sqrt{(\frac{\mathcal{L}}{2})}$ where $\mathcal{L}$ is the average length of the training sequences. The different resulting topologies of these models are illustrated in fig. 5.

Using the assumptions on the phases described in section 2.1, we aim at automatically building an *annotated HMM* that describes the stochastics of the surgical signals and permits to infer the current phase using labeled training data. This construction is described in the next section.

2.4. Annotated statistical models

We define an *annotated average surgery* as a pair

$$\alpha DTW = (\mathbb{O}^{avg}, \mathcal{B}) \tag{6}$$

where

$$\mathcal{B}_t : \{1, \ldots, P\} \rightarrow [0, 1] \tag{7}$$

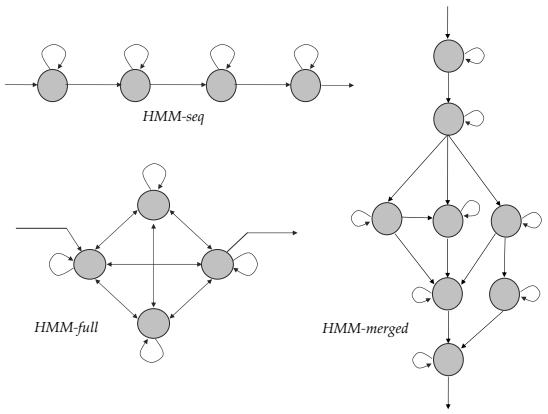

Figure 5: Different HMM topologies resulting from the construction methods mentioned in sec. 2.3.

and $\mathcal{B}_t(p)$ is the probability of being in phase p at timestamp t from $\mathbb{O}^{avg}$.

We define an *annotated HMM* as a pair

$$\alpha HMM = (\lambda, \mathcal{A}) \tag{8}$$

where

$$\mathcal{A}_x : \{1, \ldots, P\} \rightarrow [0, 1] \tag{9}$$

and $\mathcal{A}_x(p)$ is the probability of being in phase p while being in state x of λ.

The results of segmentation and phase recognition using the methods described above largely depend on how the annotated models are constructed. The methods presented in this section can be applied to both the construction of the annotated average surgery and the construction of the annotated HMM model. Since they are similar in both cases, we only describe them for the construction of the annotated HMM model. The extension to the annotated surgical average is straightforward. The two presented methods differ from the required amount of labeled data and the step where the labels are used. While the first method depends on fully labeled training data, the second method only requires a partially labeled training set.

2.4.1. A-priori phase-wise construction

This is a fully supervised framework. It supposes all sequences of the training set to be labeled. In that case, a HMM λ_p for each phase p is constructed and all these models are concatenated to form an overall model λ as presented in figure 6. Each sub-model is initialized from the training data of the corresponding phase using one of the three methods presented in section 2.3. For each state x of the overall model λ, the annotation assigns the probability 1 to the phase p, given that the state x originally stemmed from the sub-model λ_p. Otherwise, the probability 0 is assigned. If the EM algorithm is used after concatenation, the annotation $\mathcal{A}$ is updated similarly to the other HMM parameters, leading to an annotation function that does not take only binary values anymore.

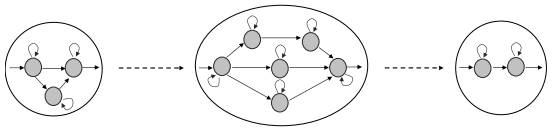

Figure 6: A-priori phase-wise model construction: Sub-HMMs are constructed for each phase or cluster of data and appended as displayed in this image.

2.4.2. A-posteriori automatic model annotation

In this case, the model is first directly constructed from all data without using any label information. Afterwards, it is automatically annotated using the training sequences that are labeled. To do this, the labeled time-series $\{\mathbb{O}^l\}$ are synchronized to the model using the Viterbi algorithm, giving paths

$$\text{path}_l : \{1, \ldots, T^l\} \rightarrow \{(x_i)_{1 \leq i \leq N}\} \quad . \tag{10}$$

The labels from the surgeries are then carried over to the model to update the probabilities in the states that were visited:

$$\mathcal{A}_x(p) = \frac{\#\{l, t : \ \text{path}_l(t) = x \ \ and \ \ p = \mathcal{T}(\mathbb{O}^l, t)\}}{\#\{l, t : \ \text{path}_l(t) = x\}} \quad . \tag{11}$$

Intuitively, the formula counts the number of visits that a label makes to each state. When several labeled surgeries are used, the annotation is therefore averaged.

Initializing the model from all the training data is computationally very expensive. The surgical similarity is used to speed up the construction, as explained below.

2.5. Discussion and Construction Speed-up

The phase-wise construction requires all phases of all training surgeries to be labeled. In contrast, the a-posteriori annotation provides the advantage that not all of the phases or surgeries from the training data need to be labeled, while the other model parameters are still initialized from all the available data. Additionally, when some phase detection is not very reliable, the user can be asked to add a few training surgeries where e.g. only these phases and their neighbors are labeled. When for an application only some subparts of the surgeries are interesting to be detected, it is also straightforward to construct the model using only data where these subparts are annotated.

Training an HMM model from full data or from data of a long phase can be computationally expensive, especially when using the model merging method which requires iterative EM computation. To alleviate this burden, we split the training data and concatenate the trained sub-models in an unsupervised approach, similarly to what is done in section 2.4.1. In section 2.4.1, the transitions between phases are used as synchronization points. We use the average surgery to find such consistent splitting point and to construct the full model from subparts. Using the surgical similarity SIM_t, we split at points which are sufficiently spaced from each other and are also local similarity maxima, where the maximum allowed size of a split is used as a single parameter. This

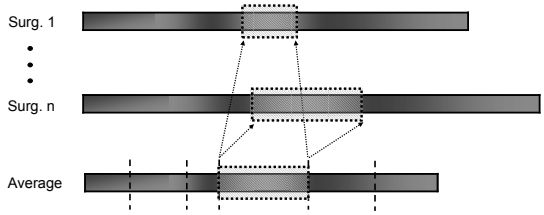

Figure 7: Training data obtained from surgical similarity splits. Each subpart of the average surgery corresponds to synchronized subparts of the training surgeries, which are then used to construct the sub-HMM.

provides clusters of training data by back-projecting from the average surgery to the training time-series, as described in fig. 7. HMM sub-models are trained from these clusters of training data and finally concatenated.

3. Off-line Segmentation and On-line Recognition

In this section, we explain how to use the annotated models for phase-detection in an unknown surgery, both in the off-line and on-line case.

3.1. Off-line segmentation

Off-line segmentation is the process of segmenting a new surgery after the acquisition of all signals $\mathbb{O}^{test}_{t,k}$. The objective is to compute the phase $\mathcal{T}(\mathbb{O}^{test}, t)$ at each time step t while knowing the *complete* signals $\mathbb{O}^{test}_1 \dots \mathbb{O}^{test}_{T^{test}}$, where T^{test} denotes the end of the surgery.

3.1.1. Using an annotated average surgery

If an annotated average surgery

$$\alpha DTW = (\mathbb{O}^{avg}, \mathcal{B}) \tag{12}$$

is available from training data, the process is the following: the time-series $\mathbb{O}^{test}$ is synchronized to the average surgery $\mathbb{O}^{avg}$ using DTW. This gives a synchronization function $\text{sync}_{test \to avg}(y) = (t_{test}(y), t_{avg}(y))$. The labels from the average surgery are then carried over to the new surgery for each time t. Since it is possible that the time t is synchronized to different consecutive times of the average surgery annotated with different most likely phases, the overall most likely annotation is used:

$$\mathcal{T}(\mathbb{O}^{test}, t) = \underset{p}{\text{argmax}} \sum_{\{y:\ t_{test}(y)=t\}} \mathcal{B}_{t_{avg}(y)}(p) \quad . \tag{13}$$

3.1.2. Using an annotated HMM model

If an annotated HMM model

$$\alpha HMM = (\lambda, \mathcal{A}) \tag{14}$$

is available from training data, the Viterbi algorithm is used to find the most likely path through the topology of λ that would generate the time-series $\mathbb{O}^{test}$. This synchronizes the time-series to the model, giving for each time step the corresponding state:

$$\text{path} : \{1, \ldots, T^{test}\} \rightarrow \{(x_i)_{1 \leq i \leq N}\} \ . \tag{15}$$

The labels from the model are then carried over to the new surgery:

$$\mathcal{T}(\mathbb{O}^{test}, t) = \underset{p}{\text{argmax}}\, \mathcal{A}_{\text{path}(t)}(p) \ . \tag{16}$$

3.2. On-line phase recognition

On-line recognition is the computation of the most probable phase when only *partial* signals $\mathbb{O}_1^{test} \ldots \mathbb{O}_t^{test}$ up to the actual time t are known. While the DTW synchronization performs very well for off-line segmentation, when the complete time series is known, on-line DTW synchronization is very unreliable. This stems from the absence of precise modeling of unlikely observations through the topology or observation probabilities, as available in HMMs. For this reason, only HMMs are used. In the annotated HMM model

$$\alpha HMM = (\lambda, \mathcal{A}) \tag{17}$$

the so-called forward probabilities [28] permit to compute $P(X_t = x_i | \mathbb{O}_1^{test} \ldots \mathbb{O}_t^{test})$, the probability of being in state x_i at time t knowing the partial observations, using dynamic programming. This provides a convenient way to obtain the most probable phase. We are looking for

$$\begin{aligned} \mathcal{T}_{test}(t) &= \mathcal{T}(\mathbb{O}^{test}, t) \\ &= \underset{u}{\text{argmax}}\, P(\text{phase} = u | \mathbb{O}_1^{test} \ldots \mathbb{O}_t^{test}) \ . \end{aligned} \tag{18}$$

From the annotation of the model, we know the probability $P(\text{phase} = u | X = x) =_{def} \mathcal{A}_x(u)$ of being in phase u while being in the HMM state x, thus

$$\begin{aligned} \mathcal{T}_{test}(t) = \underset{u}{\text{argmax}} \sum_{x_i} & P(phase = u \quad | \quad X_t = x_i) \times \\ & P(X_t = x_i \quad | \quad \mathbb{O}_1^{test} \ldots \mathbb{O}_t^{test}) \end{aligned}$$

$$= \underset{u}{\text{argmax}} \sum_{x_i} \mathcal{A}_{x_i}(u) P(X_t = x_i \quad | \quad \mathbb{O}_1^{test} \ldots \mathbb{O}_t^{test}) \ . \tag{19}$$

4. Experiments

4.1. Medical Application

The methods described above can be applied on any standard endoscopic surgery, but also on other kinds of surgery as long as informative signals about the operation can be made available. Endoscopic surgeries belong to minimally invasive surgeries and are performed through natural or small incisions. They offer less inconvenience and faster recovery times to the patient. Additionally, the limited number of laparoscopic tools used during surgery, along with the camera that observes the operating field, promise the development of automatic tool detection systems in the near future.

For the experiments, we decided to focus on laparoscopic cholecystectomy, which is the removal of the gallbladder. This is a common but complex surgery performed laparoscopically in 95% of the cases [9], with a low conversion rate to open surgery. It is therefore convenient for the recording of the signals as well as for the demonstration of the method. It starts with the positioning of trocars on the patient, for insertion of the instruments inside the body, and finishes with their removal and the suturing of the induced holes. Even though some details of the surgery depend on the patients anatomy, the surgeon follows a consistent protocol that we, in congruence with our medical partners, defined to consist of 14 workflow phases. The most important intermediate phases are the dissection, clipping and cutting of the bile duct and of the cystic artery. The gallbladder is then separated from the liver and removed using a retraction sac. For it to pass through the endoscopic hole, the gallstones that caused the operation are removed one by one beforehand. This is followed by a final control phase of the abdominal area and the removal of all instruments. A view of the OR during such a surgery is shown in figure 1. All the phases from the insertion of the trocars till the suturing are displayed in table 1.

4.2. Data Acquisition

To obtain the data, we recorded 16 laparoscopic cholecystectomies at our partner medical institution. The surgeries were performed by four surgeons of varying skill level, but from the same medical school. The recorded data comprised synchronized videos from the endoscopic view and from two external views. A convenient, custom-designed video viewing and labeling software has allowed us to manually indicate which tools were present at each time step. Note that obtaining these signals automatically is technically feasible. It is however not our primary focus in this paper, since it is a long process to introduce such a system in a real surgery room. With our medical partner we are currently working on introducing an instrument detection system based on barcodes, using a camera observing the instruments inserted inside each trocar. The system is currently being tested on pig experiments.

4.3. Results

For each method, the current phase was estimated once per second and compared to the ground truth. The ground truth was specified manually using the phases defined by our medical partners. The three following evaluation measures have been computed: accuracy, average recall and average precision. *Accuracy* provides the percentage of

1	CO2 Inflation	8	Liver Bed Coagulation 1
2	Trocar Insertion	9	Gallbladder Packaging
3	Dissection Phase 1	10	External 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

Table 1: The fourteen phases of a cholecystectomy used in the detection.

correct phase detections in the complete surgery. The two following measures are defined *per phase*: *Recall* is defined as the number of correct detections inside the ground truth phase divided by its length. *Precision* is the sum of correct detections divided by the number of correct and incorrect detections. This is complementary to recall by indicating whether parts of other phases are detected incorrectly as the considered phase. To present summarized results, we will use *accuracy* together with *average recall* and *average precision*, corresponding to recall and precision averaged over all phases. Since the phase lengths can vary largely, incorrect detections inside short phases tend to be hidden within the accuracy, but are revealed within precision and recall. For the same reason, relative values are more indicative than absolute values in minutes.

We have evaluated all presented methods on data of 16 cholecystectomies. Results for off-line segmentation are shown in table 2, results for on-line recognition in table 3. The three different names αHMM-full, αHMM-seq and αHMM-merged refer to the three different initialization methods for HMMs presented in section 2.3. We performed a leave-one-out cross-validation, where for each of the 16 surgeries the model was build from the remaining 15 ones. The results display the mean evaluation measures over all surgeries.

Obviously, since all information is available, results for off-line segmentation are higher than for on-line recognition when the same methods are compared. For off-line segmentation, methods based on DTW have accuracy above 95%. The results for the αHMM-seq method are only slightly lower. HMM topologies are designed general enough for on-line recognition and adaptation to rare observations.

As can be expected, results with a-posteriori annotation are lower than with a-priori annotation, as less knowledge is used. They are however reasonable, as an accuracy above 90% can be achieved for HMMs. αHMM-full is based on random initialization. Multiple initializations followed by expectation-maximization training are performed and the model performing best on the training data is kept. This method is less reliable in presence of limited data than both other methods, which are directly initialized deterministically from the training data. For DTW, the results are decreased by less then 3 points. This shows the interest of a-posteriori annotation, since the results even remain similar when only 50% of the training data is labeled.

For online-recognition, the best accuracy is above 90% and obtained with αHMM-seq. A-posteriori annotation slightly decreases the results, but the decrease is always

	Accuracy (%)	Avg. Recall (%)	Avg. Precision (%)
αDTW (pre)	97.3 (±6.6)	97.6 (±5.6)	97.0 (±5.7)
αDTW (post)	95.1 (±6.6)	95.5 (±6.0)	94.0 (±6.2)
αHMM-full (pre)	90.1 (±7.7)	90.2 (±8.6)	89.5 (±7.9)
αHMM-full (post)	85.4 (±12.4)	83.5 (±12.2)	80.1 (±14.2)
αHMM-seq (pre)	96.0 (±6.3)	96.5 (±5.6)	95.9 (±5.5)
αHMM-seq (post)	94.9(±5.1)	94.7(±4.9)	93.6(±5.9)
αHMM-merged (pre)	93.9 (±6.9)	93.9 (±7.9)	94.6 (±7.5)
αHMM-merged (post)	88.2 (±10.1)	88.8 (±9.0)	85.1 (± 11.4)

Table 2: Off-line results. Mean and std over all surgeries. *(pre)* indicates a-priori phase-wise construction, *(post)* construction with a-posteriori annotation.

below 5 percentage points. This is again very interesting because when only half of the data is labeled, the results remain similar for all methods, as shown in fig. 8.

αHMM-merged performed lower than the other methods. This method has however the advantage that the topology of the resulting HMM has semantic meaning [6].

All phases are usually recognized. The errors are mainly caused by a delay of some seconds when detecting a new phase, which is acceptable for most applications. The few phases that are incorrectly detected are phases that are very short and that are not recognized due to the delay. For instance, the second dissection phase has an average length of 1 minute and 49 seconds. In a few surgeries, its duration is below 15 seconds. In Fig. 9 the average length of each phase and the corresponding mean detection error for each phase are indicated in minutes. This error indicates the number of *incorrect* detections inside the ground truth phase divided by its length (corresponding to 100% - recall). The errors were obtained for αHMM-seq with a-priori phase-wise construction. The maximum mean error occurs for phase 3 and is of 1 minute and 14 seconds. For comparison, αHMM-seq with a-posteriori annotation yield a maximum mean error per phase of 1 minute and 11 seconds, occurring in phase 5. The average length among all surgeries is 48 minutes. Note that the errors are slightly higher in the phases 3 to 5. This comes from the fact that high variations in duration occur within these rather short phases while moreover phases 3 and 5 are similar. They indeed both consist in dissection.

The fact that these results were obtained using only few training data of four different surgeons with varying skill levels is especially encouraging, since this demonstrates the robustness of our statistical methods towards inter-person variability.

5. Applications and Discussion

5.1. Applications

Using above described methods, several applications could be realized with a direct and potentially large impact on patient welfare, peri-operative hospital organization and surgeon's control over the surgical processes.

	Accuracy (%)	Avg. Recall (%)	Avg. Precision (%)
αHMM-full (pre)	88.7 (±9.1)	89.4 (±8.4)	85.8 (±9.6)
αHMM-full (post)	85.4 (±11.8)	84.2 (±9.6)	82.4 (±12.6)
αHMM-seq (pre)	91.3 (±8.7)	91.6 (±7.6)	89.9 (± 8.6)
αHMM-seq (post)	91.6 (±7.1)	89.7 (±7.3)	88.5 (±9.7)
αHMM-merged (pre)	88.1 (±12.0)	88.8 (±11.3)	87.8 (±11.9)
αHMM-merged (post)	84.8 (±14.1)	84.9 (±12.0)	85.1 (± 14.7)

Table 3: On-line results. Mean and std over all surgeries. *(pre)* indicates a-priori phase-wise construction, *(post)* construction with a-posteriori annotation.

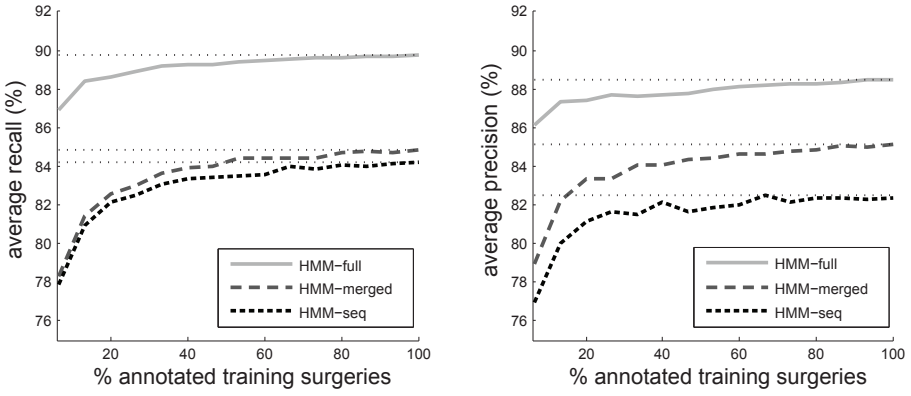

Figure 8: Mean recall and precision for αHMM-seq, αHMM-merged and αHMM-full using a-posteriori annotation. Influence of number of annotated surgeries. The horizontal lines refer to the best result for each method, obtained when all training surgeries are labeled.

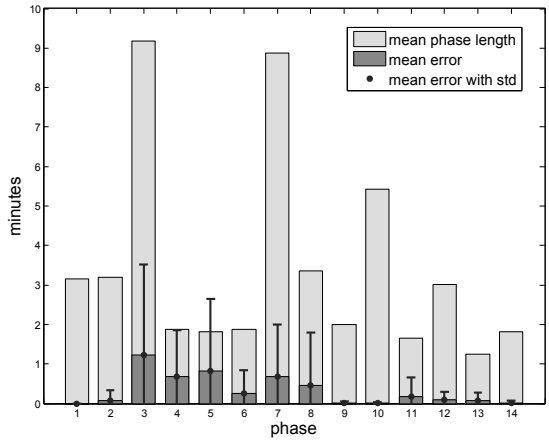

Figure 9: For each phase, average length over all surgeries overlayed with mean detection error, in minutes. Errors are computed for αHMM-seq with a-priori phase-wise construction.

5.1.1. Event triggering

Recognition of the phases can mainly serve in triggering events, like calling automatically the next patient, notifying the cleaning personnel, informing the next surgeon or giving reminders to the surgical staff. This can also be used to control a user-interface providing context-aware information. Calling the next patient is actually an important issue, since if done too soon, the next patient might stay anesthetized for an unnecessarily long time. If done too late, the operating room will remain unused during some time, which reduces the hospital efficiency. In case of cholecystectomy, this is usually done in phase 7. This could therefore be done automatically and reliably, relieving the OR staff from this task. This phase was always detected during the experiments and its mean detection error is 5.9% in case of αHMM-seq with a-priori annotation. This corresponds to an absolute error of 40 seconds, as can be seen in fig. 9.

5.1.2. Remaining time prediction

Another interesting application is the prediction of the remaining time. Making this information automatically available to hospital personnel outside the OR could significantly improve the planning of schedules. After detecting the current HMM state during a running surgery, the average remaining time can be easily inferred from the HMM. It starts to be accurate with the beginning of phase 10, with a mean prediction error below 5 minutes. A more accurate prediction should use information about the patient (e.g. size, weight, state of inflammation) and about the surgeons (e.g. skills for the kind of surgery). We unfortunately lack this data, but hope to be able to obtain it in future recordings.

5.1.3. Reporting and Training

The accurate phase identification in off-line segmentation could be used to sketch a precise and objective report, containing each phase and their starting times. This can

potentially shorten the report writing by providing a partially filled report template.

The presented methods are also very interesting for training purposes. Indeed, they permit to provide quantitative results about the performance of a training surgeon throughout the surgery. Additionally, synchronization to the average surgery can be used for evaluation and comparison through video-replay, e.g. by synchronously visualizing two surgeries of the same kind, performed by a trainee and an expert surgeon. When using on-line synchronization, this could be used on demand to show, e.g. to the trainee, how an expert has performed the current phase on another patient.

5.2. Discussion

In our experiments, we have used binary signals obtained manually from video data to prove the concept of our recognition approach. We would like to emphasize that this is not a limitation. Indeed, the automatic acquisition of such information is technically feasible, e.g. using RFID technology. Such systems are however are not present in the OR yet. Additionally, non-binary signals can also be used with the two presented methods in a straightforward manner. The annotated average surgery model can deal with such signals by default and the annotated Hidden Markov Model solely requires a proper observation model, e.g. a mixture of gaussians. Therefore, signals from all surgical devices and sensors, like the ones mentioned in the related work section, could be used as input information.

We have considered the cholecystectomy as application for the convenience of the recordings. It has a linear workflow, but it should be mentioned that the annotated Hidden Markov Model can also cope with non-linear workflows, ie. workflows containing alternatives. The annotated average surgery cannot cope with alternatives between the phases, but is still useful e.g. for synchronization inside similar phases, whose importance has been mentioned for training. These two models can deal with variations in the signals: exceptions having a short duration compared to the length of the phases will not corrupt the detection afterwards. We have experimented this effect with two recordings of cholecystectomy which included a biopsy right at the beginning. Specification in the model or extension of the modeling is however required to recognize such variations as being exceptions.

The linearity of the cholecystectomy workflow permits to obtain good results with a limited amount of training data. A workflow containing multiple alternatives will require larger training datasets to obtain similar results, since the different surgical performances need to be observed. Ideally, the definition of such complex workflows should incorporate the formal knowledge obtained by approaches providing formal workflow descriptions, like [24].

Additionally, with a large enough database of cases, it would be possible to generate surgeon-specific models of the surgery. This can be expected to further increase recognition accuracy.

6. Conclusion

In this work, we have proposed to model interventions in terms of synchronized generic signals acquired over time and present methods for off-line segmentation and

on-line recognition of the phases of a complete surgery, using either fully or partially annotated data. Using the a-priori phase-wise construction, the annotated average surgery has proven to provide reliable off-line results with an accuracy above 97%. This can for instance be used for automatic report generation after the surgery or training through synchronous video replay of surgeries. The annotated HMM permits the on-line recognition of the phases with an accuracy above 90%. Such an approach can be used for triggering events inside the operating room or for improved scheduling of the operating suites.

In both cases, the a-posteriori annotation method yields slightly lower results. But this method has the noticeable advantage that not all training surgeries need to be labeled for the model construction. When only half of the training surgeries are labeled for a-posteriori annotation, in the off-line and on-line cases the accuracy decreases by only a few percentages.

In future, we will address more complex workflows, e.g. containing alternative phases. We will also work on introducing the system into the operating room, as we believe that the analysis and processing of such signals is of growing importance. On one side, trends towards integrated surgical environments and growing numbers of sensors will lead to a vast amount of information that can be obtained from future ORs. On the other side, increasing complexity of technical systems and need for hospital efficiency demand for intelligent computer systems that can optimally assist and unburden surgical staff.

References

[1] Sheetal Agarwal, Anupam Joshi, Tim Finin, Yelena Yesha, and Tim Ganous. A pervasive computing system for the operating room of the future. *Mob. Netw. Appl.*, 12(2-3):215–228, 2007.

[2] Seyed-Ahmad Ahmadi, Tobias Sielhorst, Ralf Stauder, Martin Horn, Hubertus Feussner, and Nassir Navab. Recovery of surgical workflow without explicit models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 420–428, 2006.

[3] G. Berci, E. H. Phillips, and F. Fujita. The operating room of the future: what, when and why? *Surgical Endoscopy*, 18(1):1–5, 2004.

[4] Beenish Bhatia, Tim Oates, Yan Xiao, and Peter Hu. Real-time identification of operating room state from video. In *Proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 1761–1766, 2007.

[5] Tobias Blum, Nicolas Padoy, Hubertus Feussner, and Nassir Navab. Modeling and online recognition of surgical phases using hidden markov models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 627–635, New York, USA, September 2008.

[6] Tobias Blum, Nicolas Padoy, Hubertus Feussner, and Nassir Navab. Workflow mining for visualization and analysis of surgeries. *International Journal of Computer Assisted Radiology and Surgery*, 3(5):379–386, 2008.

[7] Kevin Cleary, Ho Young Chung, and Seong K. Mun. Or 2020: The operating room of the future. *Laparoendoscopic and Advanced Surgical Techniques*, 15(5):495–500, 2005.

[8] Marie T. Egan and Warren S. Sandberg. Auto identification technology and its impact on patient safety in the operating room of the future. *Surg Innov*, 14(1):41–50, 2007.

[9] H. Feussner, A. Ungeheuer, L. Lehr, and J.R. Siewert. Technik der laparoskopischen cholezystektomie. *Arch Chir*, 376(6):367–374, 1991.

[10] Hubertus Feussner. The operating room of the future: A view from europe. *Surg Innov*, 10(3):149–156, 2003.

[11] C Herfarth. 'lean' surgery through changes in surgical workflow. *British Journal of Surgery*, 90(5):513–514, 2003.

[12] Adam James, D. Vieira, Benny P. L. Lo, Ara Darzi, and Guang-Zhong Yang. Eye-gaze driven surgical workflow segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 110–117, 2007.

[13] P. Jannin and X. Morandi. Surgical models for computer-assisted neurosurgery. *Neuroimage*, 37(3):783–91, 2007.

[14] B.-H. Juang. On the hidden markov model and dynamic time warping for speech recognition – a unified view. *AT and T Technical Journal*, 63(7):1213–1243, 1984.

[15] Seong-Young Ko, Jonathan Kim, Woo-Jung Lee, and Dong-Soo Kwon. Surgery task model for intelligent interaction between surgeon and laparoscopic assistant robot. *International Journal of Assitive Robotics and Mechatronics*, 8(1):38–46, 2007.

[16] H. U. Lemke, O. M. Ratib, and S. C. Horii. Workflow in the operating room: review of Arrowhead 2004 seminar on imaging and informatics (Invited Paper). In O. M. Ratib and S. C. Horii, editors, *Medical Imaging 2005: PACS and Imaging Informatics (SPIE)*, volume 5748, pages 83–96, April 2005.

[17] Julian Leong, Marios Nicolaou, Louis Atallah, George Mylonas, Ara Darzi, and Guang-Zhong Yang. HMM Assessment of Quality of Movement Trajectory in Laparoscopic Surgery. *Computer Aided Surgery*, 12(6):335–346, 2007.

[18] Henry C. Lin, Izhak Shafran, David Yuh, and Gregory D. Hager. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery*, 11(5):220–230, 2006.

[19] Benny P. L. Lo, Ara Darzi, and Guang-Zhong Yang. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 230–237, 2003.

[20] Ronald Mårvik, Thomas Langø, and Yunus Yavuz. An experimental operating room project for advanced laparoscopic surgery. *Surg Innov*, 11(3):211–216, 2004.

[21] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and Evaluation of Surgical Performance Using Hidden Markov Models. *Biomedical Engineering, IEEE Transactions on*, 53(10):1911–1919, 2006.

[22] Mark A. Meyer, Wilton C. Levine, Marie T. Egan, Brett J. Cohen, Gabriel Spitz, Patricia Garcia, Henry Chueh, and Warren S. Sandberg. A computerized perioperative data integration and display system. *International Journal of Computer Assisted Radiology and Surgery*, 2(3-4):191–202, 2007.

[23] F. Miyawaki, K. Masamune, S. Suzuki, K. Yoshimitsu, and J. Vain. Scrub nurse robot system - intraoperative motion analysis of a scrub nurse and timed-automata-based model for surgery. *IEEE Transactions on Industrial Electronics*, 52(5):1227–1235, 2005.

[24] T. Neumuth, N. Durstewitz, M. Fischer, G. Strau, A. Dietz, J. Meixensberger, P. Jannin, K. Cleary, H. U. Lemke, and O. Burgert. Structured recording of intraoperative surgical workflows. In *SPIE Medical Imaging 2006 - PACS and Imaging Informatics, Progress in Biomedical Optics and Imaging*, volume 6145, 2006.

[25] T. Neumuth, P. Jannin, G. Strauss, J. Meixensberger, and O. Burgert. Validation of knowledge acquisition for surgical process models. *J Am Med Inform Assoc*, 16(1):72–80, 2009.

[26] Nicolas Padoy, Tobias Blum, Irfan Essa, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. A boosted segmentation method for surgical workflow analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 102–109, 2007.

[27] Nicolas Padoy, Tobias Blum, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. On-line recognition of surgical activity for monitoring in the operating room. In *Proceedings of the 20th Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 1718–1724, July 2008.

[28] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] A. Rogers, E. Jones, and D. Oleynikov. Radio frequency identification (rfid) applied to surgical sponges. *Surg Endosc*, 21(1237):1235, 2007.

[30] J. Rosen, J.D. Brown, L. Chang, M.N. Sinanan, and B. Hannaford. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. *IEEE Trans. on Biomedical Engineering*, 53(3):399–413, 2006.

[31] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.*, 26(1):43–49, 1978.

[32] Warren S. Sandberg, Bethany Daily, Marie Egan, James E. Stahl, Julian M. Goldman, Richard A. Wiklund, and David Rattner. Deliberate perioperative systems design improves operating room throughput. *Anesthesiology*, 103(2):406–418, 2005.

[33] Richard M. Satava. The operating room of the future: observations and commentary. *Semin Laparosc Surg.*, 10(3):99–105, 2003.

[34] Stefanie Speidel, Gunther Sudra, Julien Senemaud, Maximilian Drentschew, Beat Peter Mller-Stich, Carsten Gutt, and Rdiger Dillmann. Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling. In *Med. Imaging*. SPIE, 2008.

[35] Andreas Stolcke and Stephen M. Omohundro. Best-first model merging for hidden markov model induction. Technical Report Technical report TR-94-403, ICSI,Berkeley, CA, 1994.

[36] Kongming Wang and Theo Gasser. Alignment of curves by dynamic time warping. *Annals of Statistics*, 25(3):1251–1276, 1997.

[37] Yan Xiao, Peter Hu, Hao Hu, Danny Ho, Franklin Dexter, Colin F. Mackenzie, F. Jacob Seagull, and Richard P. Dutton. An algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time. *Anesth Analg*, 101:823–829, 2005.

[38] K. Yoshimitsu, F. Miyawaki, T. Sadahiro, K. Ohnuma, Y. Fukui, D. Hashimoto, and K. Masamune. Development and evaluation of the second version of scrub nurse robot (SNR) for endoscopic and laparoscopic surgery. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 2288–2294, 2007.