Workflow and Activity Modeling for Monitoring Surgical Procedures

Nicolas Padoy

Nancy-Université

Workflow and Activity Modeling for Monitoring Surgical Procedures

Nicolas Padoy

Dissertation

Technische Universität München, Fakultät für Informatik
Chair for Computer-Aided Medical Procedures & Augmented Reality
&
Université Henri Poincaré, Nancy 1, UFR STMIA
Ecole doctorale IAEM Lorraine, Département de formation doctorale en informatique
Laboratoire Lorrain de Recherche en Informatique et ses Applications

Workflow and Activity Modeling for Monitoring Surgical Procedures

THÈSE

déposée le 24 août 2009 à l'Université Henri Poincaré et
présentée et soutenue publiquement à Munich le 14 Avril 2010

par

Nicolas Padoy

pour l'obtention conjointe par cotutelle du grade de docteur de la Technische Universität München et de l'Université Henri Poincaré, dans la spécialité informatique.

Président :	Darius Burschka	Professeur, Technische Universität München
Rapporteurs :	Gregory Hager	Professeur, Johns Hopkins University
	Jocelyne Troccaz	Directeur de recherche, CNRS, Grenoble
Examinateurs :	Nassir Navab	Professeur, Technische Universität München
	Marie-Odile Berger	Chargé de recherche, INRIA, Nancy
	Heinz U. Lemke	Professeur, Technische Universität Berlin
	Adam Cichon	Professeur, Université Henri Poincaré

Technische Universität München, Fakultät für Informatik
Chair for Computer-Aided Medical Procedures & Augmented Reality
&
Université Henri Poincaré, Nancy 1, UFR STMIA
Ecole doctorale IAEM Lorraine, Département de formation doctorale en informatique
Laboratoire Lorrain de Recherche en Informatique et ses Applications

Workflow and Activity Modeling for Monitoring Surgical Procedures

Nicolas Padoy

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München und von der Ecole doctorale IAEM Lorraine, Département de formation doctorale en informatique de l'Université Henri Poincaré zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender :	Prof. Dr. Darius Burschka	Technische Universität München
Prüfer :	Prof. Gregory Hager, PhD	Johns Hopkins University
	Dr. Jocelyne Troccaz	CNRS, Grenoble
Examinator :	Prof. Dr. Nassir Navab	Technische Universität München
	Dr. Marie-Odile Berger	INRIA, Nancy
	Prof. Dr. Heinz U. Lemke	Technische Universität Berlin
	Prof. Dr. Adam Cichon	Université Henri Poincaré

Die Dissertation wurde am 24. August 2009 bei der Université Henri Poincaré eingereicht und am 14. April 2010 in München öffentlich verteidigt.

Abstract

The department of surgery is the core unit of the patient care system within a hospital. Due to continuous technical and medical developments, such departments are equipped with increasingly high-tech surgery rooms. This provides higher benefits for patient treatment, but also increases the complexity of the procedures' workflow. This also induces the presence of multiple electronic systems providing rich and various information about the surgical processes.

The focus of this work is the development of statistical methods that permit the modeling and monitoring of surgical processes, based on signals available in the surgery room. These methods combine low-level signals with high-level information, and can be used to detect events and trigger pre-defined actions. A main application is the development of context-aware surgery rooms, providing adaptive user interfaces, better synchronization within the surgery department and automatic documentation.

We introduce and formalize the problem of recognizing phases within a workflow, using a representation of interventions in terms of multidimensional time-series formed by synchronized signals acquired over time. We then propose methods for the modeling, off-line segmentation and on-line recognition of surgical phases. The main method, a variant of hidden Markov models augmented by phase probability variables, is demonstrated on two medical applications. The first one is the monitoring of endoscopic interventions, using cholecystectomy as illustrative surgery. Phases are recognized using signals indicating tool usage and recorded from real procedures. The second application is the monitoring of a generic surgery room workflow. In this case, phase recognition is performed by using 4D information from surgeries performed in a mock-up operating room in presence of a multi-view reconstruction system.

Keywords:
Surgical Workflow, Surgical Activity Analysis, Context Aware Operating Rooms, Hidden Markov Models, Cholecystectomy, Recognition from Multi-view Reconstruction

Résumé

Le bloc opératoire est au coeur des soins délivrés dans l'hôpital. Suite à de nombreux développements techniques et médicaux, il devient équipé de salles opératoires hautement technologiques. Bien que ces changements soient bénéfiques pour le traitement des patients, ils accroissent la complexité du déroulement des opérations. Ils impliquent également la présence de nombreux systèmes électroniques fournissant de l'information riche et variée sur les processus chirurgicaux.

Ce travail s'intéresse au développement de méthodes statistiques permettant de modéliser le déroulement des processus chirurgicaux et d'en reconnaitre les étapes, en utilisant des signaux présents dans le bloc opératoire. Ces méthodes combinent des signaux de bas niveau avec de l'information de haut niveau et permettent à la fois de détecter des événements et de déclencher des actions pré-définies. L'une des applications principales est la conception de salles opératoires sensibles au contexte, fournissant des interfaces utilisateurs réactives, permettant une meilleure synchronisation au sein du bloc opératoire et produisant une documentation automatisée.

Nous introduisons et formalisons le problème consistant à reconnaitre les phases réalisées au sein d'un processus chirurgical, en utilisant une représentation des chirurgies par une suite temporelle et multi-dimensionnelle de signaux synchronisés. Nous proposons ensuite des méthodes pour la modélisation, la segmentation hors-ligne et la reconnaissance en-ligne des phases chirurgicales. La méthode principale, une variante de modèle de Markov caché étendue par des variables de probabilités de phases, est démontrée sur deux applications médicales. La première concerne les interventions endoscopiques, la cholécystectomie étant prise en exemple. Les phases endoscopiques sont reconnues en utilisant des signaux indiquant l'utilisation des instruments et enregistrés lors de chirurgies réelles. La deuxième application concerne la reconnaissance des activités génériques d'une salle opératoire. Dans ce cas, la reconnaissance utilise de l'information 4D provenant de chirurgies réalisées dans une maquette de salle opératoire et observée par un système de reconstruction multi-vues.

Mots Clés:
Déroulement des Processus Chirurgicaux, Analyse des Activités Chirurgicales, Salles d'Opération Réactives au Contexte, Modèles de Markov Cachés, Cholécystectomie, Reconnaissance à partir de Reconstruction Multi-vues

À Raymond, Marie-Antoinette
et Marie-Thérèse.

Acknowledgments

First of all, I would like to thank Gregory Hager and Jocelyne Troccaz for taking the time to review my manuscript. I would also like to thank Darius Burschka, Adam Cichon and Heinz Lemke for accepting to be members of my thesis committee.
I am very much indebted to my adviser Nassir Navab, who established an extremely fun and motivating research environment. I have very much appreciated his support, his numerous advices and the opportunities he gave me to discover the diverse aspects of the life of a researcher. I am also grateful to my second adviser, Marie-Odile Berger, for her continuous supervision and care during this joint PhD thesis. Additionally, I owe many thanks to Erwan Kerrien for his precious advices throughout the thesis and his careful proof-reading.

This work would not have been possible without the collaborative interest of our medical partners. I would like to acknowledge the support of Hubertus Feussner, from Klinikum rechts der Isar, and the support of Sandro Heining, from Klinikum Innenstadt. Many thanks also to Armin Schneider, to nurse Franceska and to the anonymous medical staff who helped us in various ways during the data acquisitions.

I have had many exciting discussions related to the topics of this thesis with two of my former master students and colleagues: Tobias Blum and Ahmad Ahmadi. Many thanks to them for the nice 'workflow moments' shared together, for the support in the OR and for the proof-reading. I additionally thank Diana Mateus and Daniel Weinland for the long and fruitful discussions concerning my research with the 4D reconstruction data. I also would like to acknowledge Alexander Ladikos, who introduced me to the multi-camera reconstruction system. For further interesting discussions on my research directions, I am thankful to Pierre Jannin. During my thesis, I have had the chance to supervise several students: Tobias, Ahmad, Michael, Nitesh, Uli, Daniel, Kaveh, Yury, Tolga and Kateryna. I enjoyed working with them. Many thanks also to Isabelle and to Martina, for having taken such good care of the administrative issues both in Magrit and in CAMP, and to Martin Horn, for being much more than a system administrator.

Without my friends and colleagues from Munich, Nancy and elsewhere, my PhD time would surely not have been so rich and exciting! I owe a lot of gratitude to them for organizing and joining many interesting and fun activities which distracted me from the captivating dissertation work. I will keep unforgettable memories of the shared evenings, of the relaxing Biergärten, of the improtheater sessions, of the soccer and volley-ball games im Englischen Garten, of the hiking and skiing trips in the Bavarian/Austrian Alps and of the joint vacations. This balancing of the work as well as the close friendships have been invaluable.

I especially want to thank Pierre Georgel for great fun, many coffee break discussions, his music discoveries and so much more. The couch, books and music of Michael Aron have been of great help during my stays in Nancy. Merci Michael. I am indebted to Troels

and Natacha Frimor: they have constantly supported me in diverse ways, especially with delicious cakes! I also warmly thank Olivier Pauly for great moments spent outside the university and the interesting research that we were able to accomplish together in the last year. For motivating discussions and many nice events he has organized, sometimes in unexpected places, I am thankful to Jörg Traub.
Also, for very nice moments, I would like to thank: Adrien, Ajitha, Andreas H., Andreas K., Anne-Caro, Anne-Claire, Axel, Ben, Céline, Chris, Darko, Elena, Gilles, Hauke, Helmuth, Irène, Jérémie, Jose, Latifa, Macarena, M'baki, Mahssa, Marco, Martin G., Mathieu, Nicolas N., Nicolas P., Patience, Sarah, Selen, Selim, Stefanie D., Stefanie L., Sonia I., Sébastien H., Sébastien G., Stanka, Tarik, Thomas, Tobias S., Virginie, Wolfgang, Yannick and all the members of Magrit and CAMP.

Finally, I thank all the members of my family for their care, patience and encouragements during my thesis.

Comment vivre sans inconnu devant soi ?

René Char

Contents

List of Figures

List of Tables

Part I

Introduction, Motivation and Related Work

In Chapter 1, we introduce the topic of the dissertation, namely the problem of recognizing activities within a surgical workflow. We motivate this objective by describing how modern operating rooms can benefit from context-aware support and how new technologies enable it. Related work in activity recognition is then presented in Chapter 2, with a particular emphasis on the medical field. Finally, chapter 3 describes in details the two applications addressed in this thesis: the cholecystectomy application and the daily OR workflow.

CHAPTER 1

Introduction

The department of surgery is the core unit of the patient care system in a hospital. This is the place where therapeutic treatments are performed, in close collaboration with several other medical disciplines such as anesthesia. Surgery is continuously subject to technological and medical innovations, illustrated by the accelerated development and introduction of new imaging technologies, advanced surgical tools, navigation and patient monitoring systems. The purpose of these advances is to improve patient treatment. But at the same time, they also transform and complexify the pre-, intra- and post-operative daily routine.

The increasing complexity of processes occurring in surgery rooms, as well as the growing amount of available information, raise high interest for contextual support to the Operating Room (OR) staff. Contextual support could display information in the most suitable way at each timestep of the surgery, relieve the personnel from performing simple but time-consuming tasks and also assist them in the tedious ones. Context awareness in the OR is on the verge of becoming possible with the surgery rooms being revolutionized by taking fully advantage of Information Technologies.

Context aware support requires that events and activities occurring in the OR can be automatically recognized, based on signals available from the different OR tools and systems. The current difficulty to gather signals from the OR is a limiting factor to research in this area. However, efforts towards the introduction of common standards should permit to collect signals from many systems using a unique central interface in the future. Clinicians and researchers have indeed recently stressed the crucial need for a new Operating Room, called Operating Room of the Future, in which assistance systems should in particular be fully integrated to deliver increased benefits for patients, surgical staff, hospitals and the healthcare system in general [Cleary et al., 2005, Lemke et al., 2005].

Context-awareness and recognition within the OR are at their very early stages: it is the purpose of this thesis to perform some groundwork in these directions. This is accomplished in several steps. We first propose to model interventions in terms of mul-

tidimensional time-series formed by synchronized signals acquired over time. We then present machine learning approaches based on this representation to model, synchronize and recognize phases within a surgical workflow.

We address two complementary scenarios to demonstrate the concepts: endoscopic surgery workflow and daily operating room workflow. These two scenarios were not only chosen because of their complementary nature: for both of them, there exist among our medical partners interest in the application and also sustained effort to acquire the required signals during real interventions.

1.1 The Operating Room of the Future

The Operating Room of the Future (ORF) is a concept for better operating rooms and interventional suites. The improvements are defined in terms of patient treatment, information handling, system integration and ease-of-use, system and person communications, workflow, patient throughput and cost efficiency [Satava, 2003, Feussner, 2003, Berci et al., 2004, Cleary et al., 2005, Lemke et al., 2005, Sandberg et al., 2005]. In the following, we briefly present several major directions for the ORF. They show how the OR is getting increasingly high-tech and digital, up to a point where almost every activity and action can be "observed" by some system signals. This will permit to stress how the involved technologies make context-aware systems possible and also how the staff and the surgery department could benefit from contextual support.

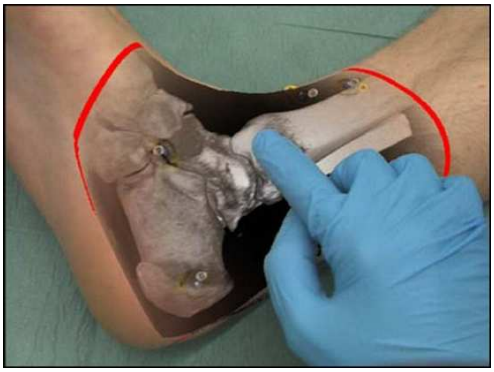
1.1.1 Image-guided Surgery

Less invasiveness has been a main objective in the development of image-guided procedures [Peters, 2006]. Contrary to open surgery, in which significant lesions are necessarily caused to soft tissues to obtain a direct access to the anatomical region of interest, minimally invasive procedures only require small incisions and are characterized by less trauma for the patient and by a faster recovery time. This is made possible by the use of specific systems which allow indirect access to and visualization of the operation field.

A widespread technique for minimally invasive surgery is endoscopy, in which the region of interest is imaged in real-time by an endoscope, a generic name for a tool consisting of a light source, an optic unit and a camera at its proximal end. It is introduced trough a small incision or a natural opening and can be rigid (e.g. laparoscope or arthroscope) or flexible (e.g. bronchioscope or colonoscope). Endoscopes can be used both for diagnosis or surgery. To control his movements, the surgeon looks indirectly at the anatomy and at the inserted tools on a video screen (see Figure 1.1(a)).

Endoscopes cannot however be used in all situations. Many pathologies remain for instance invisible to the surgeon on simple video images. Alternative solutions targeted in image-guided surgery involve the use of other imaging techniques for intra-operative guidance. Currently available imaging modalities include X-rays, fluoroscopy, computed tomography, magnetic resonance, positron emission tomography, nuclear probes and ultrasound. Even though they can sometimes be used regularly during the surgery, they usually provide only indirect visualization to the surgeon. Imaging solutions suitable for

(a) Endoscopic surgery. The surgeon looks at the monitor for hand-eyes coordination.



(b) Visualization of Joerg Traub's foot with augmented reality (courtesy of Oliver Kutter and Joerg Traub).

Figure 1.1: Example of minimally invasive techniques: navigation with endoscopy and augmented reality.

intra-operative usage are x-rays, fluoroscopy, ultrasound and sometimes magnetic resonance. Modalities which can only be acquired pre-operatively or sparsely during the surgery still provide valuable information about the structures and tissues to design a surgical roadmap.

Several issues need however to be solved in order to optimally use the information provided by these modalities during the surgery. A key difficulty is to relate with precision the positions of the surgical tools in the real world (i.e. within the patient) to the different coordinate systems of the images. This involves challenges in different areas, such as segmentation, registration, tool tracking, compensation of patient movements and tissue deformations [Peters, 2006, Yaniv and Cleary, 2006]. Presenting the resulting information to the surgeon in the right way is also a challenge in itself. A promising direction is Augmented Reality (AR) (see [Sielhorst et al., 2008] for a review of medical AR), in which guidance information is directly incorporated into the surgeon's field of view (as shown in Figure 1.1(b)). The AR technology is particularly exciting for the development of new ways to provide contextual information and interaction mechanisms to the surgeon [Navab et al., 2007].

1.1.2 Surgical Robotics

The OR2020 report [Cleary et al., 2005] evokes the idea of a surgery room completely robotized, where patients would go through similarly to cars on industry assembly lines. This is far from being possible with the actual state-of-the-art in robotics, as the capacities of decision and anticipation of a surgeon are way beyond the current possibilities of machines. However, machines are predictable and their motion can be modeled. This is not the case for humans, whose behaviours are hard to quantify and are influenced by many factors such as e.g. stress, fatigue and lack of attention. Even though human motion can be very precise, it may also be poorly reproducible and quantified.

Robotics could therefore be complementary to human abilities. The key idea is not

Figure 1.2: Operating room in Paris using the da Vinci surgical robot. The robot is controlled through a remote console, located on the left in the picture. Courtesy of Intuitive Surgical, Inc.

to replace the surgeon, but to provide him/her with tools that enhance his capabilities, under his/her control. Robotic assistance is especially indicated in fine operations or minimally invasive surgery scenarios, in which the requirements create harder working conditions, smaller areas of operation, and diminished visual and haptic perceptions. Benefits include precision, miniaturization of access and ergonomy. The developments and applications of robotic surgery are numerous [Davies, 2000, Taylor, 2008]. Most robots are designed for specific surgeries. We briefly mention two major application areas of surgical robots: the execution of planned tasks and the extension of the surgeon's hand. In the execution of planned tasks, the robot is used to perform an autonomous task under surveillance of the medical team. Its planning has been decided beforehand by taking into account pre-operative patient information. This has advantages in precision and accuracy over a human. Example of application is the insertion of a needle based on a target indicated in a pre-operative image. In the case of extending the surgeon's hand, robotic tools are controlled by the human through a human machine interface. This permits enhancing features such as filtering of high-frequency motion in the human hand, preventing movements from deriving away from a pre-planned route, or performing non-natural gestures.

Extending the surgeon's hands offers promising surgical possibilities, such as telesurgery [Satava, 2005]. Telesurgery has been demonstrated on real operations performed by two commercial robots: the da Vinci robot [Guthart and Jr., 2000] (see Figure 1.2) and the Zeus robot [Ghodoussi et al., 2002]. In fully robot assisted surgery room, telesurgery could even permit to conduct remote surgeries in unmanned OR. This is for example targeted by the Traumapod project [Friedman et al., 2007], which aims to provide an unmanned OR on the battlefield controlled remotely by a surgeon.

In robotic environments, rich signals for analyzing and monitoring the surgical processes are directly available from the robots [Lin et al., 2006].

1.1.3 Digital Operating Room

Most information used and exchanged during a surgery tend to be digital. Patient information, operation schedules, surgical reports and image modalities are handled by computers. Patient monitoring systems, image guided assistance systems, electronic tools, intraoperative imaging devices and robots could have their signals digitally available. Currently however, most devices present in the OR are working independently and do not share information on a common platform. This slows down data exchanges between staff, systems, departments and hospitals [Cleary et al., 2005, Lemke et al., 2005]. This also prevents the design of user interfaces providing a complete overview of the OR status [Meyer et al., 2007]. Lack of standardized interfaces and concurrent developments are at the origin of this issue.

Propositions for new standards to design an integrated and digital OR are emerging [Lemke, 2007, Lemke and Vannier, 2006]. The driving idea is to extend the well-established DICOM standard designed for picture archiving and communication systems (PACS) used in radiology to incorporate all components intervening in the surgical processes in addition to imaging modalities. The concept proposed by Lemke et al. [Lemke and Berliner, 2007] is called Therapy Imaging and Model Management System (TIMMS). TIMMS aims at modeling all data exchanges and decision processes within the surgery room (see Figure 1.3). It contains a communication platform and modules that model and control all components of a procedure, ranging from images and signals to the visualization devices, including the procedure's workflow and knowledge databases. The system needs to be scalable, distributed and to provide interchangeable objects. Design and experiments with an initial TIMMS framework are reported e.g. in [Mayoral et al., 2008].

The adoption of a digital OR could improve the operative efficiency during the surgery, by allowing smoother interaction, communication and decision processes. This could also give way to an accelerated development of telecollaboration, since all information will be available on-line with low transmission delays.

1.1.4 Workflow Optimization

Due to the increasing amount of information, systems and communication requirements, optimizing the workflow is a major objective in the design of the Operating Room of the Future [Cleary et al., 2005, Lemke et al., 2005]. This should permit an improved and more quantifiable treatment as well as cost reduction through better scheduling within the surgery department and increased patient throughput. Workflow in the OR is conditioned by communication interfaces, spatial layout, personnel skills and teamwork abilities. The workflow can first be optimized *off-line*. This usually involves the redesign of the OR and of the personnel guidelines by experts, based on statistics of the past surgeries. Indeed, optimal integration and positioning of systems along with an optimal ordering of personnel tasks can reduce setup, communication and usage delays.

Depending on situations and human factors, the surgical staff may however not always behave optimally at each time-step, e.g. with respect to tool usage and scheduling. Staff may also be overloaded by incoming information. For this reason, workflow should also

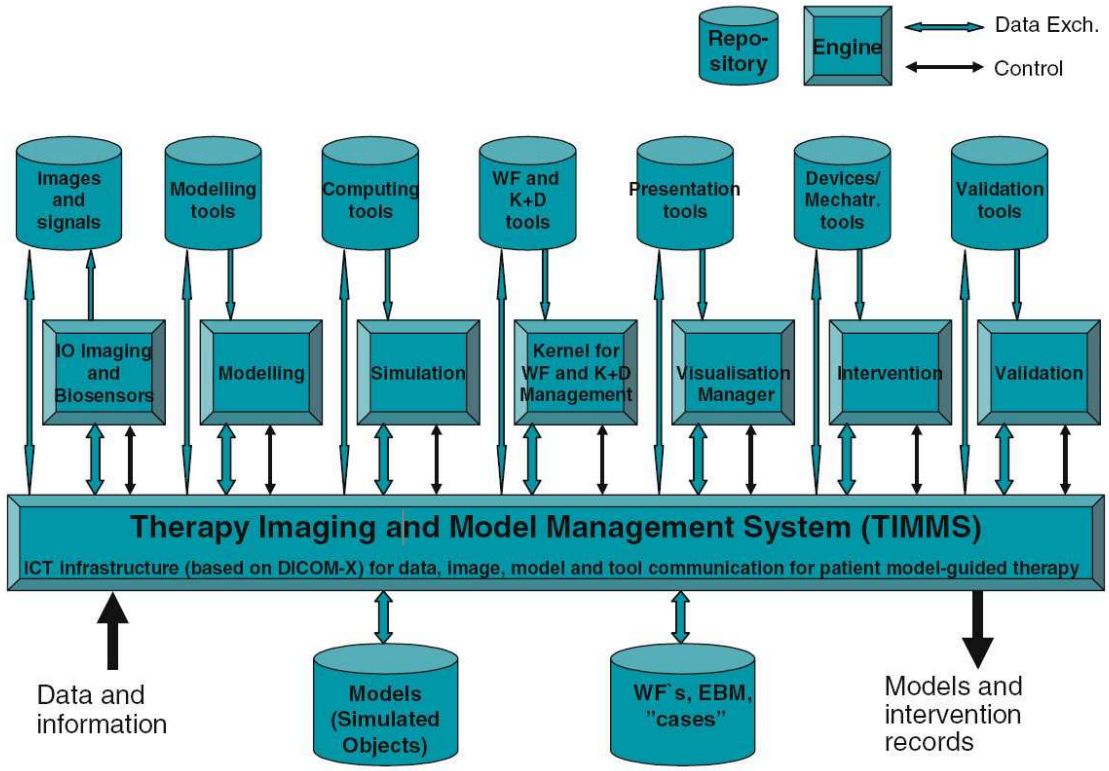


Figure 1.3: Therapy Imaging and Model Management System (TIMMS). Courtesy of [Lemke and Berliner, 2007].

be optimized *on-line* by providing the OR staff with context-sensitive support that relies on actual conditions within the OR. Such support can be the display of the right information at the right time, the triggering of reminders to the personnel, or the real-time transmission of information about personnel and room availability. There is now a need for intelligent systems that will use information from both previous surgeries and models provided by experts to propose opinionated and non-opinionated guidance throughout the procedure (corresponding to the module "Kernel for Workflow and Knowledge + Decision Management" in Figure 1.3). Off-line and on-line optimizations are naturally intertwined, but the on-line nature of such systems is becoming increasingly crucial with the numerous tasks that have to be accomplished, e.g. in image guided surgery [Navab et al., 2007, Lemke and Berliner, 2007].

1.2 Signal-based Surgical Workflow Analysis

1.2.1 Sensor-enriched Operating Rooms

The trends described in the previous section suggest that the Operating Room of the Future will host a tremendous set of sensors and electronic devices that will provide information about almost all activities occurring in the OR, performed either by surgeons,

assistants or nurses. Surgical activity will be indicated by signals coming from the electric activation of tools, from tool and staff tracking systems, robot movements, intra-operative imaging devices, patient monitoring systems, etc. Peri-surgical activity will be perceived by additional sensors, such as video cameras, RFID sensors and the usage of the hospital information system. In addition to these real-time signals, patient specific information, including age, previous treatments and pre-operative images will be digitally accessible. When this information is available in real-time from a unique computer interface and in a synchronized manner, the preliminary requirements for the development of a fully aware operating room will have been met. The OR will then be equipped with multiple sensors similarly to the Aware Home [Kidd et al., 1999], a testbed for developing context-aware assistance in the home.

1.2.2 Surgical Workflows

In most current medical procedures and applications, the sequence of surgical steps and actions that are accomplished to perform a treatment follows a repetitive schema. This schema is usually called a *workflow* (see Section 1.2.4). Workflows do not only structure the performance of the operation, from the first incision on the patient to the last suturing, but also the complete usage of the surgery room, from the entrance of one patient to the next one. In regular operations, patient and surgeon specificities influence the surgical steps in their details. But they do not alter the overall workflow.

Workflows are described in medical books, formalized by protocols and learnt by the personnel during medical studies and training. They are also enforced by the specialization of surgery rooms and cost efficiency requirements [Herfarth, 2003]. In other industries, for instance in production lines, formalization and optimization of workflows as well as information technology integration are more advanced and have shown to be a key point for reducing costs and improving quality. But contrary to each instance of an industrial product, a patient cannot be processed by a simple automate. His variability and his condition of human being raise additional technical and ethical issues.

1.2.3 Potential Applications

The presence of rich OR signals, combined with the constraints and repetitivity imposed by surgical workflows, offer new possibilities for analyzing, understanding and monitoring workflows using machine learning approaches. Several applications are provided in the following sections.

1.2.3.1 Situational Awareness

The huge amount of available information will permit the construction of decision systems that can interpret the current signals based on a database of signals acquired from previous interventions. Possible applications differ both in scope and coarseness of their monitoring as well as in the amount of expert supervision required in the system construction. Some application examples are:

Scheduling: The current surgery stage can be synchronized in real-time to the standard surgery workflow. This would permit to follow visually on a workflow diagram the progress of an operation in a remote control room. Such information could be used to adapt room and personnel scheduling within the surgery department.

Context-sensitive user-interfaces: User-interfaces that adapt to the context can permit to display the right information to the right OR member at the right moment. They could permit to trigger reminders and notifications to new surgical personnel about the next steps that have to be performed. This could ease the preparation of instruments at the right moment. In the long-term, this could even permit a certain degree of automation.

Documentation: Key events and their times of occurrence can be written down automatically, easing the hand out of the final report by the surgeon. Together with the complete archiving of the data, this also provides objective material for legal matters.

Anomaly detection: Unpredicted variations from the signals with respect to the previous procedures can be used to trigger warnings to the surgical staff. More trivially, this could be used to notify the personnel when immediate equipment maintenance or resource replacement (for instance oxygen bottles) are required.

1.2.3.2 Surgical Data Mining

Traditional modeling and design of workflows is based on a manual approach performed by experts. When the objective is to analyze and redraft an existing workflow, e.g. for optimization purposes, objective and statistical information about the existing workflows is very valuable. Experts can indeed be subject to bias, by observing the workflow as it should be rather than how it really is [van der Aalst et al., 2003]. Set of signals, which can also be referred to as process logs, can be analyzed automatically to provide their underlying structure and statistics. This is precious information for experts, who can more easily analyze and find out used resources, causal dependencies, sources of delays and exceptions within the previous procedures.

In addition to their usage for the redrafting of a workflow to make it more efficient, objective inputs can be used to design benchmarks for workflow comparisons. This would be very useful to evaluate the impact and benefits of the introduction of a new medical device.

1.2.3.3 Databases Analysis and Training

The archiving of all signals permits several applications such as planning based on statistical analysis and search for previous similar cases. This can be achieved by analyzing pre-operative images of previous patients, but also by visualizing previous surgeries, including their specific potential issues. This is also particularly valuable for training. Indeed, trainees could visualize surgeries of the same kind performed by different surgeons, displayed in a synchronized fashion. This could permit comparison and comments

by a teacher. This would also allow the trainee to compare himself objectively to more experienced surgeons.

1.2.4 Definitions Related to the Problem

Several terms related to recognition within workflows will be used throughout this thesis. Terms referring to actions, their sub-parts and their compositions are not well-defined, since their distinction and coarseness are a matter of application and time-scale. We present below a usual terminology, based on a-priori semantic knowledge about the application domain.

An *action* is defined as the fundamental *element* in the semantic interpretation of a scene; it consists for example in picking up, handing, cutting. When several of these actions are considered together, they form *activities* or *behaviors* [Bobick, 1997]. With increasing complexity and timescale, we define a *phase* as a semantically meaningful group of activities occurring somewhere inside a temporal sequence of actions, i.e. a "step" in a process. Phases occur *repetitively* or constantly across different sequences; the order in which they appear matters. Phases can have huge differences in durations, especially across recordings. The semantic relevance which leads to considering a group of activities as a phase, is purely determined on the basis of domain knowledge. Finally, a set of phases along with their temporal relations is named a *workflow*.

In this thesis, the workflow analysis approaches are signal-based, i.e. the workflow is observed by a time-series of multidimensional vectors containing signals (information units), for instance obtained from video images and electronic systems. A multidimensional vector of signals at a certain time-step can be referred to as an *observation*. The time-series or sequence of observations obtained during the execution of a workflow is called a *workflow instance* or *workflow exemplar*.

The regular recording of a workflow within a surgery room yields a database of workflow exemplars, illustrating the workflow and its variations, depending on patient, environment and staff differences. Machine learning approaches on this database are used to construct recognition systems. Recognition is said to be *on-line*, when it is performed for each time step of the workflow instance, using all signals of the instance acquired up to that particular time step. It is said to be *off-line*, when it is performed for each time-step, but based on all signals of the workflow instance up to the last time step. In the latter case, the recognition can only occur after the end of the execution of the workflow instance. Real-time recognition is by nature an on-line process.

1.3 Contributions

The analysis of surgical workflow based on real-time signals is a recent research field. To our knowledge, there is no published work presenting on-line workflow recognition approaches for complete surgeries. In this thesis, we introduce and formalize the problem of recognizing phases within a surgical workflow, using representations of interventions in terms of multidimensional time-series formed by synchronized signals acquired over time. We represent the entire surgical workflow within a single model that is learnt from fully

or partially annotated data and permits either off-line segmentation, on-line recognition or both.

The first contribution is a surgical workflow representation, which can be used for off-line surgical phase segmentation of a sequential workflow. The approach combines high-level annotation with average surgical signals and is based on surgery synchronization with the dynamic time warping algorithm. The model is called annotated virtual surgery representation (AVSR) and several approaches are proposed to construct it.

The second contribution is a representation based on hidden Markov models, called annotated workflow hidden Markov model (AWHMM), that permits both off-line segmentation and on-line recognition of the phases of a workflow, possibly containing alternative courses of activities. Several approaches are proposed to construct it. Like the AVSR, the AWHMM is based on phase probability variables that retain semantic information. These variables store the annotation and permit both to prevent semantic information from getting blurred during training and to conveniently construct the model under partial labeling.

The third contribution is to provide a solution for workflow monitoring of laparoscopic cholecystectomies, based on tool usage during the procedure. Results are presented in terms of recognition, modeling and remaining-time prediction, using recordings from real procedures performed at the Hospital Rechts der Isar, Munich. We also present initial steps towards the usage of the endoscopic video as input signal as well as a joint effort with our medical partner to develop a trocar-camera system for real-time tool recognition.

The fourth contribution is an approach based on 4D features for recognition in a complex OR workflow. To recognizes the phases, we propose to use 3D-flow computed directly on multi-view reconstruction data. The approach is evaluated on surgeries containing alternative workflows and performed in a mock-up OR. This application is motivated by the installation of a 16-camera system within an interventional room at Hospital Grosshadern, Munich.

Additionally, we present in Appendix A some early work to introduce wearable sensors in the surgery room and to define automatically a motion vocabulary from this data for action recognition in percutaneous vertebroplasties. The list of publications I authored and co-authored during this thesis is available in appendix E. The abstracts of several major publications that are not discussed within this thesis are available in appendix F.

1.4 Outline

We first present in Chapter 2 related work in the domains of action recognition within workflows, of recognition within medical applications and of modeling of surgical workflows. Chapter 3 introduces the two applications addressed in this thesis. The laparoscopic cholecystectomy procedure, the daily OR workflow scenario as well as the setups and signal acquisition approaches are presented. Chapter 4 describes methods based on dynamic time warping to synchronize surgeries, to generate an annotated virtual surgery representation (AVSR) and to segment the surgical phases of interest. Results are provided for the cholecystectomy application. Work related to this chapter has been published in [Padoy et al., 2007a]. In chapter 5, hid-

den Markov models are presented, as they are the core of the recognition methods proposed in the next two chapters. Chapter 6 introduces the annotated workflow HMM (AWHMM) in the context of laparoscopic procedures for modeling and performing online recognition. Several methods described in this chapter have been published in [Padoy et al., 2007b, Padoy et al., 2008, Blum et al., 2008a, Padoy et al., 2009]. Chapter 7 presents the usage of 3D reconstruction data for recognition in a daily OR scenario. An effective AWHMM initialization is described as well as experiments on a workflow containing alternative paths. Work related to this chapter has been published in [Padoy et al., 2009]. Finally, conclusions of the thesis are presented in chapter 8.

Early work on defining a motion vocabulary from inertial sensor data obtained from vertebropasty procedures is presented in Appendix A. Work related to this appendix has been published in [Ahmadi et al., 2008, Ahmadi et al., 2009]. Major publications not discussed in this thesis are briefly presented in Appendix F.

CHAPTER 2

Related Work in Surgical Workflow Analysis

Human activity recognition is a large research domain that has been actively investigated in the last decade. The trend is higly motivated by the wide variety of applications concerned with understanding and modeling human behaviors, such as surveillance, human-machine interaction and medical monitoring.

The surgical application left aside, extensive literature exist in the computer vision and pattern recognition communities on human activity recognition. The proposed methods are principally based on the usage of video images or wearable sensors. It is however important to note that actions in daily life are less constrained and less subtle than in the operating room. For this reason, most work address the recognition of independent actions, but do not consider the constraints of a workflow. In the operating room, recognition of surgical activity is a recent field of research. Only few work have directly addressed this application under the specific constraints and difficulties implied by the surgical environment. Closely related is literature focusing on the evaluation of surgical skills in minimally invasive surgery and on the design of robotic assistants for the OR. It provides insights both in the OR signals that can be obtained and in the methods used for their analysis. At a more abstract level, the increasing literature about formal modeling of surgical processes stresses the important role played by the surgical workflow inside the operating room.

As the surgical workflow recognition problem is highly multi-disciplinary, we decided to present the different related areas. We first briefly present literature addressing general human activity recognition, with a focus on vision-based approaches using workflow constraints and on approaches using wearable sensors. We then focus on recognition within the department of surgery and on the formal modeling of surgical processes. We briefly position our work with respect to the existing literature at the end of this chapter.

2.1 Human Activity Recognition

Outside the surgery room, previous works have addressed problems such as the recognition of human actions [Aggarwal and Cai, 1999, Yan et al., 2008]; the activity and anomaly detection in public environments, [Grimson et al., 1998, Bremond et al., 2006, Wang et al., 2007]; the modeling and identification of the primitive actions composing activities [Moore and Essa, 2002, Shi et al., 2006]; the automatic discovery of the activities [Hamid et al., 2007, Xie et al., 2004] and the recognition of events in groups such as in meetings [Oliver et al., 2004, Brdiczka et al., 2007].

These approaches rely on sensors placed within the environment or directly on the observed persons. They differ either in the signal combinations and representations, or in the machine learning approaches used for recognition. The Aware Home project [Kidd et al., 1999] is representative of the sensors available for activity analysis. In this project, a house containing ubiquitous sensing was built to study everyday's activities and develop context-aware services. Sensors included video cameras, RFIDs, microphones and weight floor captors. Combining wearable computers with the external sensors is also proposed. Currently, video is still the most widely studied support for research in activity analysis, as cameras are non-invasive, cheap and already present in many locations.

A common underlying feature in the above-cited literature is the focus on recognizing isolated actions or activities (e.g. pick up, wave, fight, etc). In this way, they comply with most long-term applications, where relevant actions need to be detected but periods of inactivity and uninteresting actions should be discarded. In this thesis, we address the activity recognition problem in the context of a workflow. In this case, activities follow a well-defined structure in a long period of time and can be semantically grouped in relevant phases. The major characteristic of the phase recognition problem is the temporal dependencies between phases and their highly varying durations.

In Section 2.1.1, we will first review the subpart of the computer vision literature in which constraints of the workflow are considered in some ways. We will then present, in section 2.1.2, existing work using wearable sensors. These sensors are prone to be miniaturized and may get used soon into the OR [Yang, 2006].

2.1.1 Vision based

For a recent review of general action recognition methods based on video images, we refer the reader to [Weinland, 2008]. In the following, only articles considering workflow aspects are presented.

Pinhanez and Bobick [Pinhanez and Bobick, 1997] were among the first to analyze complex flows of actions. In their seminal work, framing in a TV studio was proposed to be done automatically, by triggering cameras based on the detection of events specified in a script. Unfortunately, their vision system was not able to detect the events and the authors were forced to generate them manually. Although progress has been done ever since, using vision systems for recognition in complex environments remains difficult. The problem is usually simplified by limiting the number of actors or that of actions, using a distinctive background, restricting the activity area or discriminating the actions by their spatial location.

[Koile et al., 2003] introduced the concept of *activity zones*: regions in an environment that are linked to specific activities. Likewise, [Nguyen et al., 2005] represented a room as a collection of cells. In both cases a tracking system was used to determine the presence of a person in a zone or the occupation of a cell. The goal of [Nguyen et al., 2005] was to recognize behaviors that differed in the occupied cells and in the sequence of their occupation. Hierarchical HMMs were used to recognize three behaviors, namely having a snack, a normal meal or a short meal. [Moore et al., 1999] used a camera installed above a desk to track a person's hands and objects on the table. An HMM was used to detect actions based on interactions of the hands with the objects. The use of model constraints to recognize complex events has been suggested in works like [Xiang and Gong, 2008, Moore and Essa, 2002, Vu et al., 2003, Shi et al., 2006]. [Xiang and Gong, 2008] addressed structure learning in HMMs to obtain temporal dependencies between a few high-level events for video segmentation. An HMM modeled the simultaneous output of event-classifiers to filter their wrong detections. [Moore and Essa, 2002] used stochastic context-free grammars to recognize separable multi-tasked activities in a card game from video. Production rules were manually defined to describe all the relations between the tracked events. [Shi et al., 2004, Shi et al., 2006] proposed propagation networks to model and detect from video the primitive actions of a task performed by a tracked person. Propagation networks are graphical models that explicitly model parallel streams of events and are used for classification. The detailed topology is handcrafted and trained from partially annotated data. [Vu et al., 2003] used a symbolic approach to recognize complex activities in surveillance applications. For each activity, a formal scenario was provided by hand, including actors, objects and their spatio-temporal dependencies. Concerning the computation of features to observe the scene from video data, the aforementioned literature mainly relied on detection and tracking of the persons and objects. Failure of one of these components hinders the recognition of the events. These articles imposed constraints on the flow of events to design a better model and improve recognition results. In most of them, the constraints were provided manually with an handcrafted topology and annotated data. In [Nguyen et al., 2005, Xiang and Gong, 2008], the structure is learnt from the data. To simplify the problem, the data either comes from constrained activities (e.g. an action corresponds to a spatial area) or has been pre-processed to contain high level semantic information. The main focus is on action classification. For this reason, on-line detection results are rarely provided. Additionally, the actions usually have similar short durations.

2.1.2 Wearable-sensor based

Even though cameras are available in many places, not all open areas can be covered to observe human activities. Moreover, it remains difficult to estimate accurately human postures and movements from videos in unconstrained environments. Wearable sensors tend to become more comfortable than some years ago with the miniaturization of the involved technologies, in particular of power supplies. They can thus provide additional and complementary information on human activities.

[Bao and Intille, 2004] used two or five biaxial accelerometers worn on the body to

recognize various physical activities, such as running, brushing teeth, reading, vacuum cleaning. Features extracted with the fast Fourier transform were used by decision tree for classification. [Huynh et al., 2007] used three biaxial accelerometers to classify a wide range of daily activities. Classification results were compared using clustering algorithms, support vector machines and HMMs. For the recognition of free-weight activities, [Chang et al., 2007] used two three-dimensional accelerometers, one in a glove and one on the waist. The approach, based either on Bayesian classifiers or on HMMs, resulted in the automatic counting of repetitions. In [Kern et al., 2003], an architecture for gathering data from multiple 3D accelerometers was presented. Physical activity was classified with Bayesian classifiers. [Krause et al., 2003] used an armband containing two biaxial accelerometers and sensors of physiological data to classify daily activities using self-organizing maps. In [Lester et al., 2006], the authors proposed to use a single body location, but a rich sensor-board including among others accelerometers, a barometer, a compass and a microphone. Data was classified with HMMs after feature selection. To recognize actions performed during an assembly task in a workshop, [Ward et al., 2006] used a combination of wearable microphones and accelerometers. The different activities were classified with linear discriminant analysis and HMMs. In [Subramanya et al., 2006], activities were recognized together with locations, using a GPS and a sensor board. A dynamic Bayesian network was trained on partially annotated data to model dependencies between activities and locations.

Medical applications of wearable sensors are for instance the analysis of gait [Morris and Paradiso, 2002], the monitoring of chronic patient [Patel et al., 2007, Bravo et al., 2008] and the monitoring of disease processes [Yang, 2006] .

These work show that recognition of individual activities using wearable sensors is possible. The methods however do not take any workflow constraints into account. Since such sensors are prone to be introduced in the OR [Yang, 2006], they are good candidates for being used for action recognition in surgical workflows.

2.2 Surgical Activity Recognition

In Section 2.2.1, we present works aiming generally at the recognition of events or activity within the OR for various applications. Recognition focused on activities performed by the surgeon during a minimally invasive surgery will be presented in Section 2.2.2. In Section 2.2.3, we present some work addressing the evaluation of surgical skills in endoscopy, since the acquired signals are particularly of interest for workflow recognition.

2.2.1 Recognition in the Operating Room

In [Agarwal et al., 2007], a system was presented for the automatic generation of an electronic medical record (EMR). The system was recording information about the administration of medicines, the presence of surgical staff and the occurrence of medically significant events. For the detection of medical events, physiological signals were queried using a rule-based knowledge system based on fuzzy logic and designed by experts. Experiments were carried out in a trauma scenario with four events using physiological signals

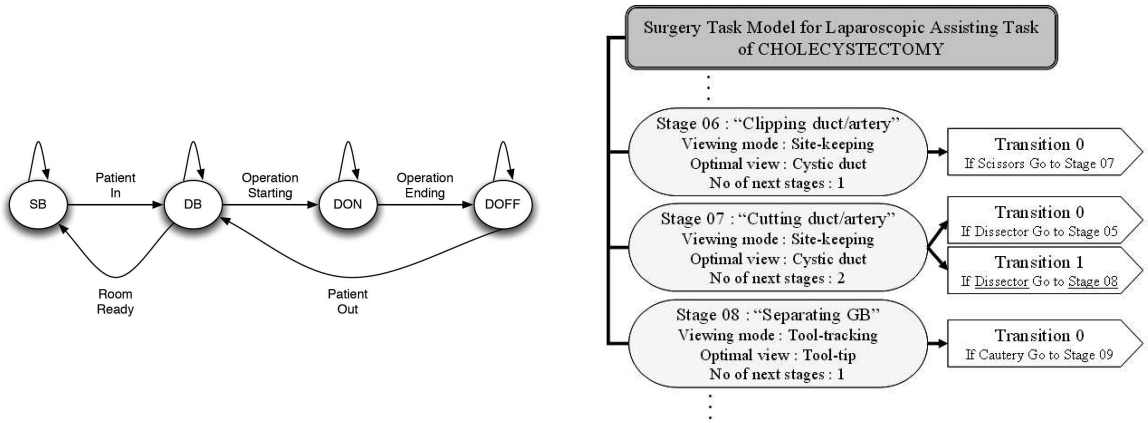
provided by a simulator. Detailed results were not presented, but the authors mentioned high recognition rates with a generally high latency. Tracking of medical supplies and of presence/absence of patient or staff was achieved using RFID sensors. An application of the system is the automatic EMR generation within an unmanned OR. The development of such an OR was targeted by the traumapod project for providing healthcare on the battlefield, using telesurgery with the da Vinci robot from Intuitive Surgical and robotic assistants [Friedman et al., 2007]. In [Agarwal et al., 2007], the system offered the display of tracking information and recognized events, elementary checks such as the presence of the right patient, and also convenient replay of videos recorded during the procedure.

[Meyer et al., 2007] pursued a similar objective. In collaboration with human factors engineers, an integrated display was developed for the Operating Room of the Future at the Massachusetts General Hospital, showing dynamic information gathered from RFID tags and OR systems. The display adapted to the stage of the procedure, which was obtained from the milestone data entered by the nursing personnel. In addition to the smart display of contextual information to all staff members present in the OR, a mentioned long-term objective was the integration of decision support into the system.

Location information Locations provide useful clues about the activities. So far, RFID has been the most widely investigated technique for material and person tracking within the hospital. Objectives were mainly intendance and reporting, but also activity recognition. For instance, in [Bardram, 2004], a context-aware system using RFIDs was presented for several nursing applications including monitoring a pill container, monitoring the patient bed and using efficiently the electronic patient record. In [Favela et al., 2007, Sanchez et al., 2008], input data containing personnel location and interactions was used with neural networks or hidden Markov models to classify daily activities of medical workers inside the hospital, such as taking care of patients or performing clinical case assessment. The data was acquired manually from a study conducted at the hospital, but could be obtained with RFIDs.

Within the surgery room, RFIDs have been mainly proposed to check the presence of the right patient and staff and to perform inventory tasks [Nagy et al., 2006, Egan and Sandberg, 2007]. An important potential application is also the automatic tracking and counting of sponges, to make sure none was left inside the patient [Macario et al., 2006, Rogers et al., 2007]. In [Houliston and Parry, 2008], a setup based on RFIDs was proposed to record the activities in the anesthesia department. The mentionned longer term objective was to monitor the anesthetist's activity.

The usage of RFIDs for tracking the surgical tools would be highly beneficial, both for inventory matters and for recognition of surgical activities. In other environments, such as in the house, interactions between persons and objects have proven useful for activity recognition [Wyatt, 2005]. However, the integration of RFID tags within the small surgical instruments is still a technical issue. [Smith et al., 2005] proposes a glove equipped with an RFID reader. It was tested on medical volunteers [Fishkin et al., 2004] but was reported to be intrusive. Practical solutions for tool recognition in endoscopy are at the moment video-based, by using autoclavable color markers attached to the instruments [Ko et al., 2007, Radrich, 2008].[Tahar et al., 2008] attempted to employ the tools



(a) Courtesy of [Bhatia et al., 2007]. (b) Courtesy of [Ko et al., 2007].

Figure 2.1: (a) Four-states HMM model for patient presence detection. (b) Decision rules for endoscopy holder control in cholecystectomy procedure.

in open surgery as current interruptors to detect their usage, but this approach has too many contraints to become practicable. Several kinds of minimally invasive surgeries, e.g. imaged-guided neurosurgery, need the tools to be tracked. Tool presence is then directly available from the guidance systems, such as VectorVision (Brainlab [1]), in which tools are tracked optically using infrared reflectant markers placed at the tip of the instruments.

Physiological signals Physiological signals reflect the patient health state. They are used to detect anomalies, by the anesthetist in the OR, but also by portable monitors in daily life for patients at risk [Bar-Or et al., 2004]. Within regular and non-crucial surgeries, they do not provide much contextual information, especially as they contain many artifacts. An original use of these signals was made in [Xiao et al., 2005] to automatically detect when the patient entered and exited the OR. The detection was achieved by processing the pulse oximetry, electrocardiogram and temperature readings, but was however very sensitive to signal artifacts. The aim was to provide a hospital-wide system for passively monitoring the status of the surgery rooms [Hu et al., 2006].

Videos In [Bhatia et al., 2007], four different OR states describing the surgery room occupancy with respect to the patient bed presence and occupation were recognized using external video information. Hue histograms of the video images were classified by support vector machines and then filtered temporally using a four-states hidden Markov model, illustrated in Figure 2.1(a).

2.2.2 Recognition in Endoscopy

Due to the simpler and more convenient access to surgical information that endoscopic surgeries provide, especially in terms of signal acquisitions, they are the first kind of

1. `http://www.brainlab.com`

surgeries that are addressed for recognition in the surgical domain. Additionally, the existing acceptance and knowledge by the surgeon of the many advanced tools required during the procedure simplifies the introduction of new technical systems that would be based on surgical steps recognition.

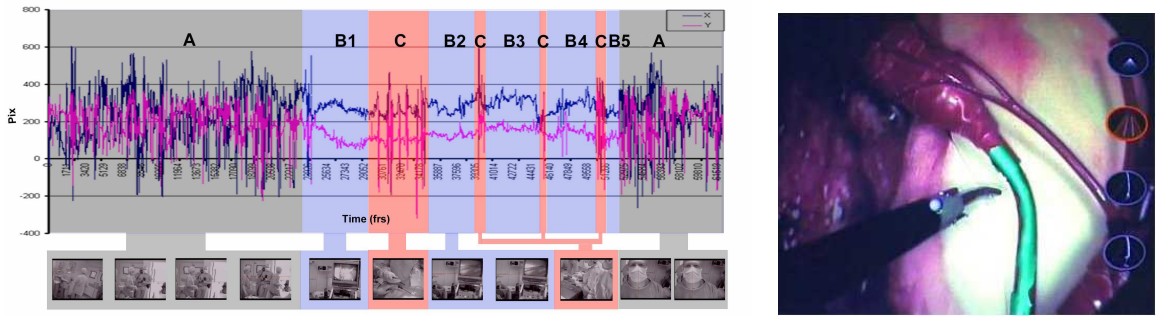
The robotic community has an increasing interest in the recognition of surgical steps, both for automation and development of context aware robotic assistants. Targeting the development of a robotic scrub nurse that would automatically provide the correct endoscopic instrument to the surgeon, [Miyawaki et al., 2005] analyzed movements of the surgeon and of the nurse during a surgery to design timed-automata that model the surgeon's activities, the scrub nurse's activities and the surgical steps. The conception of the model was done by hand, based on recorded videos and on motion tracking obtained from visual markers, which was time consuming. In this initial work, the suggestion was to use the resulting model in conjunction with parametric model checking methods to control the path planning of the robot. In further work [Yoshimitsu et al., 2007], they presented a vision-based system using color markers recognizing when the surgeon is requesting a laparoscopic instrument. The robot handed in the instrument, but it only worked with a tiny set of instruments provided in a predefined order. This recognition was not used in combination with the model.

In [Ko et al., 2007], a task model of cholecystectomy was designed for guidance of a laparoscopic robot which was controlling the camera pose. A viewing mode was assigned to each surgical stage and transition rules between the stages were defined manually based on the active surgical tool detected using color markers. An excerpt of the rule system is illustrated in Figure 2.1(b). It was clearly mentioned that a surgical step cannot be always uniquely recognized from the current surgical tool. They could not distinguish the ambiguity and they planned to address this problem in future work. An additional issue to consider is that quite often different surgeons use the tools in different ways or for different purposes than what these were originally intended for [Mehta et al., 2001]. A learning-based approach, as we propose, has the potential to adapt to the specific technique of each surgeon.

In [Lo et al., 2003], the endoscopic video was used to classify four elementary tissue/instrument interactions, namely idle, retraction, cauterization and suturing. Cues including shape, deformation and illumination information were used within a Bayesian network to perform the video frame classification. In this work, the objective was to assess the surgical skills and results are provided on several sub-sequences of real surgeries. Two of the four interactions, idle and cauterization obtained low classification results. Within complete surgeries, classification rates obtained by using solely video information are expected to be worse, due to the complex and cluttered scenes.

[James et al., 2007] addressed the recognition of one surgical step of a pig cholecystectomy: the clipping of the cystic duct. The surgeon's focus was tracked by an eye-gaze system and the endoscopic video was processed to detect the presence/absence of an instrument inside a portion of the video. These two pieces of information were combined and used with artificial neural networks to detect the step. The signals provided by the eye-gaze tracker are illustrated in Figure 2.2(a).

In [Speidel et al., 2008], an approach for surgical risk assessment was presented and



(a) Courtesy of [James et al., 2007].

(b) Courtesy of [Speidel et al., 2008].

Figure 2.2: (a) Eye gaze tracking signals during a porcine cholecystectomy. (b) Risk assessment in minimally invasive surgery. The system detects the proximity of the instrument with a risk structure.

tested in an experimental setup. The authors proposed to describe high-risk surgical tasks with a logic system that models interactions between anatomical objects and surgical tools recognized from the endoscopic video. An augmented reality system was used to visually notify the surgeon about detected risk situations by highlighting risk structures (see Figure 2.2(b)).

As can be seen from these works, motivation for recognition of endoscopic activities is high. However, methods and signals are so far specific and totally dependent on the targeted application and respective experimental setup.

2.2.3 Surgical Skills Evaluation

Interesting signals for the analysis of surgical gestures are the positions of tools or the forces applied to them. These can be obtained indirectly by using a tracking system or directly using a robot which can also provide the positioning information. Up to now, such signals have been seldom available during regular surgeries and have been mainly used for evaluating and comparing surgeons performing on a simulator. The need for objective methods assessing the surgical dexterity has indeed increased with the introduction of minimally invasive surgeries [Darzi et al., 1999].

In [Cao et al., 1996], endoscopic gestures were video-taped and decomposed into elementary subtasks (called surgemes in [Lin et al., 2006]) to manually mark explicit differences in the movements performed by expert and novice surgeons. [Rosen et al., 2006] and [Megali et al., 2006] used force and torque or kinematic information from a simulator to build hidden Markov models representing surgical dexterity. This allowed the definition of metrics for evaluating surgeons. [Leong et al., 2007] assessed surgeons' hand-eyes coordination by comparing the trajectories of surgical instruments recovered using a Polaris tracking system (NDI [2]). Two HMMs were built to model respectively the trajectories performed by experts and by novices. These were then used to classify a new trajectory

2. `http://www.ndigital.com`

into one of the two modeled skills classes. [Lin et al., 2006] acquired signals from a da Vinci robot to classify surgeons' skill levels on a suturing task. The approach used linear discriminant analysis and Bayesian classification. In [Sielhorst et al., 2005], a method was proposed to qualitatively compare the performance of a trainee using a birth simulator to the performance of an expert. Trajectories of tracked forceps were temporally synchronized using dynamic time warping and displayed simultaneously within an augmented reality system.

The literature cited above focuses on particular gestures but not on complete surgeries. However, similar signals as the ones acquired and used by these authors could be employed for surgical workflow recognition, especially in case of robotic surgery, where such signals are easily available without additional setups.

2.3 Formal Modeling of Surgical Workflows

Tools to formally describe surgical processes are needed for the analysis, visualization and optimization of surgical workflows. There were a few attempts to automatically generate workflow models from process logs, based on approaches from business process modeling [van der Aalst et al., 2003]. For instance [Maruster et al., 2001] attempted to derive a Petri-net from data describing hospital events. Another approach from [Blum et al., 2008b] based on HMMs will be presented in section 6.6.3. However, limitations arised from the lack of semantics present in the process logs. For this reason most approaches are based on expert knowledge.

[MacKenzie et al., 2001] proposed a hierarchical decomposition of the tasks occurring during a laparoscopic surgery. This was obtained manually based on a dataset of recorded surgeries and helped in the understanding of tasks and interactions for optimizing the workflow. In [Jannin et al., 2001], a model based on Unified Markup Language (UML) was proposed in order to understand and optimize the usage of imaging modalities during a neurosurgical procedure. The model was illustrated in Figure 2.3. Using a database of instances, this model was proven to be useful for pre-operative planning [Jannin and Morandi, 2007]. [Neumuth et al., 2006a, Burgert et al., 2006, Neumuth et al., 2006c] presented ontologies and tools to describe and record surgeries in a formal manner. The work steps and interactions occurring in surgeries could be recorded manually by assistants using a software that helps generating standardized descriptions, which could in turn be used for an in-depth analysis of the workflow. Different workflow visualization methods were proposed in [Neumuth et al., 2006b]. An integration of such workflow descriptions within the DICOM standard was presented in [Burgert et al., 2007, Lemke, 2007]. For querying and analyzing recorded surgical workflows, a data warehousing approach was proposed in [Neumuth et al., 2008]. Finally, [Qi et al., 2006] proposed to build the hospital information system on a workflow model that describes as a dependency graph the tasks that have to be accomplished by each department.

The broadest and most widespread formalization for surgical processes is the concept of a Therapy Imaging and Model Management System [Lemke and Berliner, 2007], illustrated in Figure 1.3 on page 8. It was proposed to extend the Picture Archiving and

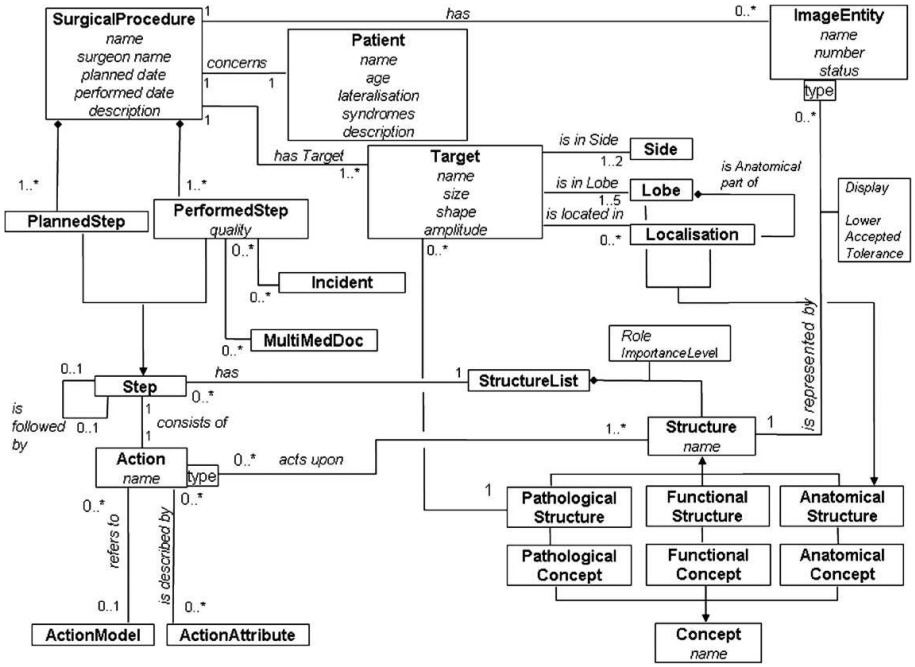


Figure 2.3: UML class diagram representing a surgical ontology in image-guided neurosurgery. Courtesy of [Jannin and Morandi, 2007].

Communication System (PACS) concept used in radiology for the needs of surgery. At term, it should model all surgical processes and objects, and contain a surgical workflow module.

While these modelings provide interesting descriptions of the workflow, a link to real signals still needs to be made for the models to be used in recognition systems.

2.4 Thesis Positioning

Previous sections describe several recent directions that have been taken in the literature for analyzing the surgical workflow. In this thesis, we investigate a new direction, which is complementary to existing work. We namely address the problem of automatically recognizing and classifying the surgical phases during complete operations. In [Neumuth et al., 2006a, Jannin and Morandi, 2007], detailed modelings of surgical workflow are proposed. These models are however not related to real surgical signals. Therefore, they cannot be used for recognition. As opposed to these approaches, we propose statistical models which directly rely on existing surgical signals and which are designed for off-line and on-line recognition. Since they require less manual supervision, our models do contain less semantic information, though.

The design of such a recognition system can be regarded as the development of a "Kernel for Workflow and Knowledge+Decision Management" module within the TIMMS concept introduced in [Lemke and Berliner, 2007]. In [Lin et al., 2006], the recognition of surgical gestures is addressed. The focus is however on surgical skills evaluation. For this reason,

experiments are carried out in experimental setups featuring generic surgical tasks. Instead, we focus on the recognition of phases during the complete surgery, for the purpose of designing context aware support systems. In parallel work, [James et al., 2007] address recognition in cholecystectomy. The approach is however designed to recognize only a single phase. It also requires an eye-gaze system and results are so far only presented on pig surgeries. In contrary, within our cholecystectomy application, we use generic signals and present results based on acquisitions of real surgeries.

Another objective in this thesis is to *learn* how to recognize surgical activities during complete surgeries, based on labeled recordings of the targeted application. As can be seen from Section 2.1, HMMs are widely used for activity recognition in the computer vision community, since they conveniently deal with time-series. In our applications, the considered activities are constrained by a workflow. Classification without considering the temporal contraints of the workflow is prone to errors due to ambiguities. For this reason, we do not train HMMs for each activitiy separately, but train a global form of HMM, in which the activities can be tracked using additional phase probability variables. This modeling directly copes with variations in duration, takes workflow constraints into accounts and does not require parameters for windowing on the data during on-line processing.

We demonstrate our methods on two novel and complementary applications, each featuring a different level of coarseness in the recognized semantic information. In the cholecystectomy application, we recognize the endoscopic phases, based on tool usage signals. This is the first work presenting an on-line approach and results of phase detection evaluated on real and complete surgeries. The second application consists in the recognition of the phases of a generic daily OR workflow. A multi-camera system is used to observe the scene, motivated by the constraints of the complex and occluded surgical environment. This is the first work proposing 4D features, not only for recognition in surgical workflow, but also for recognition in complex reconstructed scenes containing several objects and persons.

Within the following chapters, additional differences to closest work as well as further discussions on their complementarity to ours are presented in more details.

CHAPTER 3

Applications and Setups

Two main applications are addressed in this thesis: the cholecystectomy procedure and the daily OR workflow. This chapter describes the scenarios, as well as the signal acquisitions that we performed during real operations or in a mock-up OR.

Cholecystectomy is a common procedure, consisting in the removal of the gallbladder. In most of the cases, it is performed in a minimally invasive setup using laparoscopy. Even though the methods developed in this thesis can apply to other surgeries, either endoscopic or open, provided that signals such as tool usage are available, this procedure was the main medical application targeted in this work. It was at the origin of much effort to acquire data in the OR during real procedures. It is indeed convenient for recordings since it is frequent and safe, but also suitable for demonstration as it contains a complex workflow.

Motivated by the future possibility of gathering rich and real-time reconstruction data from a multi-camera system under development at one of our partner hospital, we further investigated the usage of this data for recognition of the generic daily workflow in the OR. In addition to the different kind of signals available, a major difference to the cholecystectomy application is that the scenario is coarser and contains alternative paths of activities.

Our clinical partners for the two applications were respectively from Hospital Rechts der Isar and from Hospital Grosshadern, both located in Munich.

3.1 Laparoscopic Cholecystectomy

3.1.1 Description

3.1.1.1 Clinical Indication

The gallbladder is an organ located in the upper-right part of the abdomen, under the liver, as illustrated in Figure 3.1. It stores the bile produced by the liver and releases

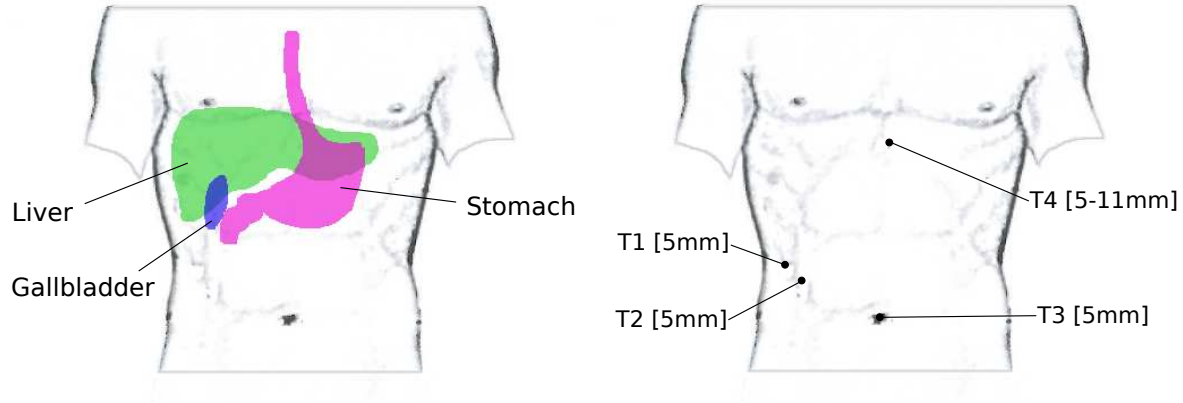


Figure 3.1: Abdominal anatomy (left) and trocar positioning (right).

it into the intestine to help the digestion process. Gallstone disease, also know as the biliary calculus disease, is the most common biliary disorder. It is caused by hard stones appearing in the gallbladder or in the bile duct. They are formed by chemicals from the bile when their proportions get unbalanced. They provoke no symptoms in most cases, but may also cause irritation or pain. In the latter case, the common medical indication is the removal of the gallbladder, as humans can live normally without it.

3.1.1.2 Brief History

The first open cholecystectomy is attributed to the german surgeon Carl Langenbuch and took place in Berlin in 1882 [Morgenstern, 1992, Olsen, 2006]. The first *laparoscopic* cholecystectomies were performed approximately one century later, independently by Dr. Erich Muhe in 1985 in Germany and by Dr. Philippe Mouret in 1987 in France. The success of laparoscopic cholecystectomies generated an unprecedented research effort towards minimally invasive techniques that boosted the developments of endoscopic surgery [Riskin et al., 2006]. Due to its recognized benefits, in particular faster recovery time and less pain for the patient, laparoscopy is now the standard for cholecystectomy.

Most recent developments now permit the performance of cholecystectomy without leaving any scar on the patient, either by using single port access through the umbilicus [Bucher et al., 2009], transluminal access [Jacques et al., 2007] using natural openings with NOTES (natural orifice transluminal endoscopic surgery [Bucher et al., 2009]) or the previous two techniques simultaneously [Zornig et al., 2008].

3.1.1.3 Procedure

The following paragraph describes the details of this procedure in non-medical terms. Different variations in this procedure exist depending on patient specificities and technical preferences of each medical school. Complete medical descriptions are e.g. available in [Feussner et al., 1991, Dubois, 1993].

Before the beginning of the procedure, the patient first gets prepared for full-anaesthesia. He is then positioned for surgery and connected to anaesthesia system for

monitoring of vital signs. In the same time, patient data are displayed on the computer interface and surgical tools are prepared in a sterile zone. The endoscopic environment, including endoscopic tower, monitors, and electric coagulation and cutting devices are connected and placed at their respective positions. The patient abdominal area is disinfected and the surgery is ready to start, upon arrival of the surgeon.

In laparoscopic cholecystectomy, the endoscopic procedure is performed with four trocars which permit the introduction of the instruments inside the abdominal area. Their exact positions depend on patient specificities and on surgeons' preferences, but they are usually placed in a configuration shown in Figure 3.1. The procedure starts with a first incision at the umbilicus navel and the introduction of a needle for inflation of the abdominal cavity with CO2. Then, the first trocar, of size 5mm, is introduced inside this opening, and a first inspection of the abdomen is performed with the endoscope. Upon this inspection, the other trocars with size 5mm or 10mm are placed and inserted, one after each other under endoscopic visualization.

At University Hospital Rechts der Isar, this endoscopic procedure is usually performed with two assistants and one scrub nurse, in a setup called german position: the operating surgeon uses trocars 2 and 3, while trocar 1 is used by an assistant to hold the liver with a metal rod in order to reveal the gallbladder. Trocar 4 is used by a second assistant holding the endoscopic camera.

The next steps, which are also the most delicate, consist in the dissection, clipping and cutting of the bile duct and of the cystic artery. Using a grasper and a dissecting device, the two vessels are revealed by dissection of neighboring tissues. Then, a clipping device is introduced to ligate the two vessels by using two to three clips. The two vessels are cut and the gallbladder is ready to be separated from the liver. Its dissection is performed with an electronic cutting tool that applies high-frequency currents to the tissues, in combination with an electronic coagulation device that stops bleedings. This part of the procedure generates smoke that may slightly obscure the vision through the endoscopic camera.

Finally, a retraction sac is introduced through the larger umbilical incision and the gallbladder inserted into the bag using graspers. After removal of the trocar, the sac is extracted. For it to pass through the hole, first the gallstones that caused the operation are removed from it one by one. This is followed by the drainage of irrigation fluids, a final control phase of the abdominal area, the removal of all instruments and the suturing of the incisions.

An illustration of the main phases of the surgery, in pictures, is given in Figure 3.3 on page 32.

3.1.1.4 Personnel

At the university Hospital Rechts der Isar, six persons are usually involved in the procedure. The nurse and the scrub nurse first prepare the room and the surgical instruments. Once the patient arrives in the room, he gets prepared for surgery by the assistants and the anesthetist. After the start of the procedure, the sterile zone is occupied by the surgeon performing the surgery, the assistant holding and controlling the camera, a second assistant retaining the liver with a liver rod and the scrub nurse preparing and passing

surgical tools to the surgeon. Additionally, a nurse is present outside the sterile zone to take care of missing items, schedules, phone calls and other needs. In the anesthesia area, an anesthetist monitors patient vital signs.

3.1.1.5 Relevance of the Procedure for Workflow Analysis

The cholecystectomy is a very convenient surgery for demonstration of workflow analysis methods. It is indeed a common and safe procedure, for which recordings can regularly take place. In the study [Feussner et al., 1991] conducted at our partner hospital Rechts der Isar, it is reported that less than 3% of 178 cases have required conversion to open surgery due to complications. Additionally, the procedure is complex, in the sense that it requires many different surgical instruments and also many surgical steps to be performed. This surgery is also widely used in the literature related to surgical workflow analysis, for instance in [Miyawaki et al., 2005, James et al., 2007, Ko et al., 2007].

3.1.2 Representation

All signals providing information about the surgery room, the patient state, and the staff actions are interesting for recognition during the surgery. Unfortunately, it is not possible yet to obtain all of them simultaneously in a synchronized fashion. For recognition of the phases occurring during cholecystectomy, we will show in the next chapters that the endoscopic tool usage provides enough information.

3.1.2.1 Tool Usage

The history and combination of used tools correlate with the underlying workflow in endoscopic surgeries. For this reason, we rely on tool usage to infer the actions performed by the surgeon.

We represent an endoscopic surgery of time length T by a multidimensional time-series $\mathbb{O}$ where $\mathbb{O}_t \in \{0,1\}^K$:

$$\mathbb{O}_{t,k} = 1 \quad \textbf{if and only if} \quad \text{signal } k \text{ is active at time } t.$$

The signals represent the instruments' usage during the laparoscopic surgery and in our case the number of recorded instruments is $K = 17$. Signals obtained in one of the surgeries we have recorded are displayed in figure 3.2. Here the signals from the operating room are binary and recorded at a temporal resolution of one second.

3.1.2.2 Other Signals of Interest

Within the surgery room, other surgical information could provide interesting signals related to the workflow of the procedure. Signals from the anesthesia devices could for instance be of interest, especially for the detection of anomalies. The material usage entered by the nurse into the hospital information system and the state of the operating or room lights could provide further information about the progress of the surgery. Even though these values are very dependent on the surgical case, the amount of water used

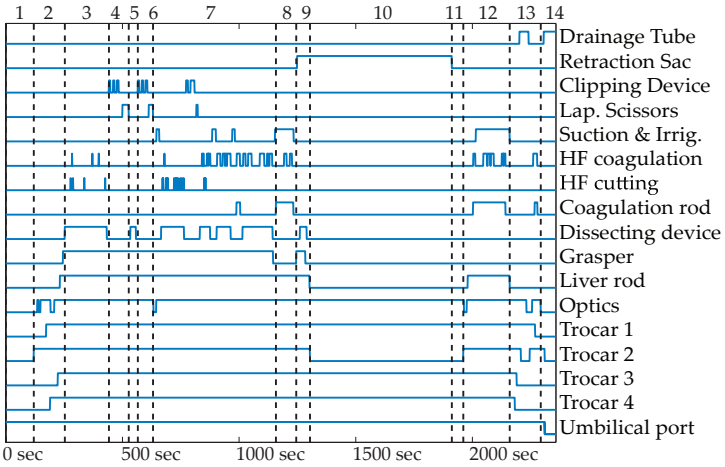


Figure 3.2: Signals recorded from a surgery. The numbers on the top of the figure indicate the different phases.

for irrigation, the quantity of liquid retrieved with the suction device and the amount of CO2 give also interesting clues about the evolution within the workflow. Any other additional signal, binary or not, could be included into the modeling with only little further modification.

3.1.2.3 Phases

From the workflow described in section 3.1.1.3, together with our partner surgeon we selected fourteen generic phases displayed in table 3.1. Even though these phases can have high variations in duration, they can be easily and reproducibly identified within the surgery instances by an expert. They are semantically meaningful by representing each a different medical objective. These phases are the ones that will be used for segmentation and detection in Chapters 4 and 6. They are displayed next to their corresponding signals in Figure 3.2 and visually illustrated in Figure 3.3 on page 32.

3.1.3 Data Acquisition

In this section, we first describe the data acquisitions performed at Hospital Rechts der Isar. We then present ongoing work for the automation of the acquisition process.

3.1.3.1 Approach

The first recordings were performed using two to three camcorders. One recorder was connected to the external output of the endoscopic tower to collect the endoscopic video. The second camrecorder was set up in the back of the room, in order to capture the complete scene. In some cases, we used a third camrecorder to capture a zoomed view of the surgical theater. Synchronization was performed by using a stopwatch with

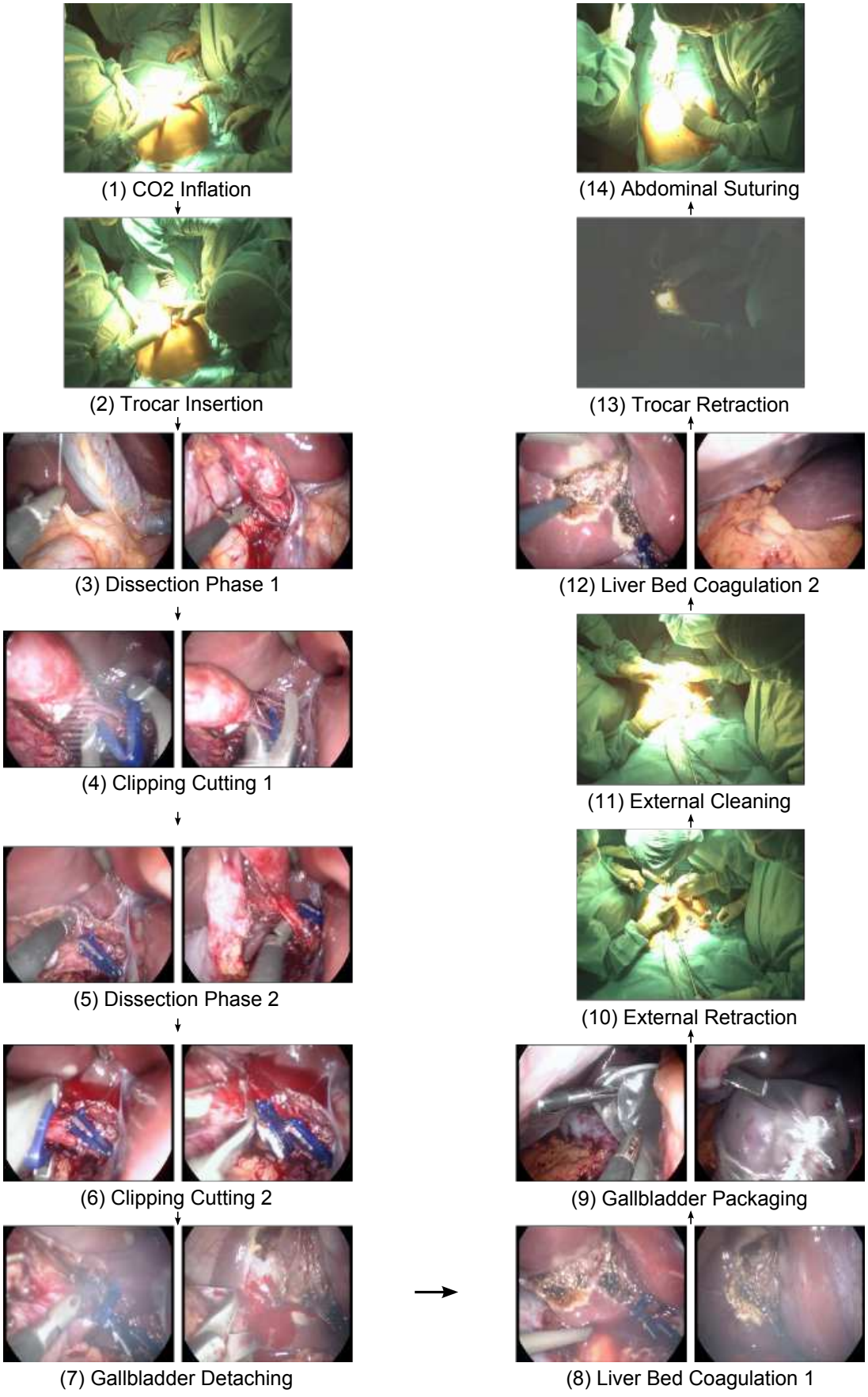


Figure 3.3: Illustration of a laparoscopic cholecystectomy.

	Phase Name	Duration
1	CO2 Inflation	188 (±93)
2	Trocar Insertion	186 (±40)
3	Dissection Phase 1	498 (±265)
4	Clipping Cutting 1	110 (±47)
5	Dissection Phase 2	110 (±146)
6	Clipping Cutting 2	113 (±55)
7	Gallbladder Detaching	550 (±383)
8	Liver Bed Coagulation 1	207 (±129)
9	Gallbladder Packaging	121 (±82)
10	External Retraction	342 (±341)
11	External Cleaning	104 (±120)
12	Liver Bed Coagulation 2	181 (±62)
13	Trocar Retraction	78 (±55)
14	Abdominal Suturing	114 (±99)

Table 3.1: The fourteen cholecystectomy phases used in the recognition. The *duration* column displays the mean and standard deviation of the phases durations, in seconds. These values were computed for the 16 surgeries that are used in our experiments. The average total duration is 48.5 (±18.5) minutes.

100Hz temporal resolution viewed at the beginning and at the end of the recordings by all cameras.

In the last phase of the recordings, in a joined effort with Helmuth Radrich [Radrich, 2008], we used a more sophisticated setup providing four precisely synchronized videos. The first two videos provided endoscopic and room view, the two second videos acquired direct views of the operating theater, for stereo tracking of surgical tools [Radrich, 2008]. Pictures from the surgery rooms at hospital Rechts der Isar can be seen in Figure 3.4.

The surgeries that have been recorded were performed by four different surgeons from

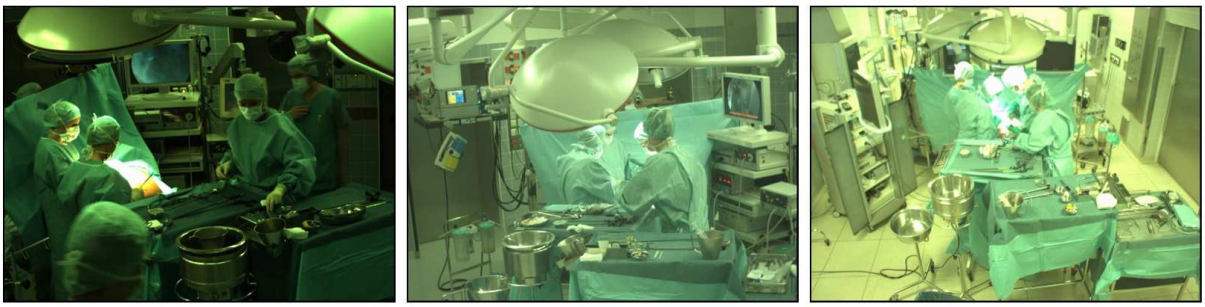


Figure 3.4: External views of three different operating rooms from Hospital Rechts der Isar, Munich, where recordings have been performed.

(a) Trocar camera

(b) Instrument with barcodes

Figure 3.5: Pictures from the trocar camera and from an instrument covered with barcodes.

the same medical school. Additionally, we have attended several cholecystectomy procedures carried out at another hospital (Klinikum Innenstadt, Munich), but could not perform regularly recordings there. Note that the phases were the same. However, instruments were sometimes used in different ways. Such differences should be learnt by the recognition system given that enough training data is available.

The instrument tool usage was obtained for the experiments presented in this work by manual labeling of the videos using a dedicated software. The next section presents an effort for real-time acquisition of these signals.

3.1.3.2 Trocar Camera

Even though the technology exists, at the moment no practical, reliable and cost-effective solution exists for identification of endoscopic tools usage at each timestep. Several companies, including Karl Storz, have been considering the integration of RFIDs inside the tools, but they do not see enough market for it at the moment.

The trocar camera is a device investigated by the team of Prof. Feussner (Institute for Minimally Invasive Therapy and Intervention, Hospital Rechts der Isar, Munich) that should permit the recognition of the tools inserted in a trocar. It should be used in the future for logistic and workflow monitoring purposes. It consists in components similar to a flexible endoscope, as can be seen in Figure 3.5(a). A flexible tube contains two optic cables: one for transmitting a light source, one for observation with a camera. The tube can be connected to the side of the trocar for observation of the tools inserted into it. Its design has to permit the usage in the sterile operating room conditions. An additional advantage is that no electric cables need to be brought in proximity of the patient.

The device was manufactured by the company PolyDiagnost [1]. Tool recognition is performed using barcode detection from the video image using a software that we devel-

1. `http://www.polydiagnost.com`

oped in collaboration with the company MVTech [2]. Issues arise from the poor contrast and the reduced field of view of the camera, which does not capture the complete trocar area, e.g. when a trocar of size 10mm is used. We proposed to solve this issue by using short barcodes of type 2/5 Interleave, which do not incorporate redundancy, but fit fully inside the video image. The camera used is a Guppy F046B from Allied Vision Technologies [3], which provides a framerate of 40fps. The instrument is covered with barcodes for reliable detection, as shown in fig. 3.5(b). In a final setup, the barcodes would be printed on the instruments by laser. Tests in a fully controlled environment resulted in a correct detection. In animal experiment conditions, the device has shown unstability, for technical reasons. A better fixation of the tube to the trocar is required, as well as a fixed emplacement for the camera, so that the image stays stable. When this is improved, further animal experiments will be carried out to test the system in real conditions, where e.g. blood can alter the view of the instruments.

3.2 Surgery Monitoring with a Multi-camera System

The second application addressed in this thesis is the recognition of coarse daily OR activities. This is complementary to the phase recognition problem in endoscopy, since here the cholecystectomy procedure would consist in one of the phases. We investigate the usage of multi-view reconstruction data provided by a permanent system that would continuously observe the OR.

3.2.1 Description

3.2.1.1 Motivation for a Multi-camera System

The monitoring of surgical processes occuring within a surgery room requires a sensing system that on the one hand does not impair the workflow, and on the other hand captures enough information about the activities. For recognition at the scale of a surgery, tool usage can be used, as in previous section. For recognition of overall daily activities, signals that continuously observe the OR are required. Cameras are a natural choice, since they are cheap, widespread, non-intrusive and can be easily installed on the ceiling. The workflow in the OR comprises multiple, precise and complex activities usually involving the interaction of several people and objects. As the crowded scene implies multiple occlusions, a set of cameras is required to keep an overview of the whole activities.

Originally, the installation of the multi-camera system was launched within a research project at the Chair for Computer Aided Medical Procedures of Prof. Navab to monitor a robotic C-arm, the Zeego robot developed by Siemens (illustrated in Figure 3.6). The objective was to prevent possible collisions between the system and the staff and objects present in the room [Ladikos et al., 2008b]. This is a way to enforce safety requirements, but also to potentially permit faster movements of the robots, as required for better imaging results and usage of less contrast product. The solution proposed by

2. http://www.mvtec.com
3. http://www.alliedvisiontec.de

Figure 3.6: Left) Zeego robot from Siemens. Right) Picture from the intervention room at Klinikum Grosshadern, Munich, where the multi-camera system is installed.

[Ladikos et al., 2008b] is to compute the occupation of the room in real-time, generating a reconstruction volume in which potential collisions between objects can be predicted and therefore avoided, as illustrated in Figure 3.7.

We investigate in this thesis a further application of the system, namely the usage of the reconstructed data for activity recognition. When observing the scene with a camera, colors on clothes and tissues are similar and multiple occlusions occur as the personnel principally work around a small area around the patient table. For these reasons, tracking and recognition of fine-grained human actions in this specific environment are extremely difficult and furthermore, not absolutely necessary for phase recognition. We postulate that a global model of the scene obtained by the reconstruction is sufficient for coarse recognition since the whole activity focus is on the patient. To capture the coherence of this activity, we use the real-time reconstruction algorithm. This choice guarantees that the system will not interfere with the normal behavior of the medical staff in the OR and permits the usage of generic low-level features to characterize the phases.

The system at the hospital is currently under development. To prove the concept, we used data that we recorded with a replicate multi-camera system present in one of our laboratories.

3.2.1.2 System Setup

We describe here the multi-camera system used in our laboratory to perform the recordings. Further details are available in [Ladikos et al., 2008a]. The architecture of the system is depicted in fig. 3.8. It consists of one server and several clients. Locally, each client handles up to four cameras and performs a partial reconstruction. Partial reconstructions are then sent over the network to the server, which handles synchronization, global reconstruction and display. All cameras are fully calibrated and mounted on the ceiling to surround the room. The system is currently able to handle 16 cameras with a framerate of 30fps, yielding a reconstruction volume of size $3.7 \times 3.2 \times 2.2$ meters.

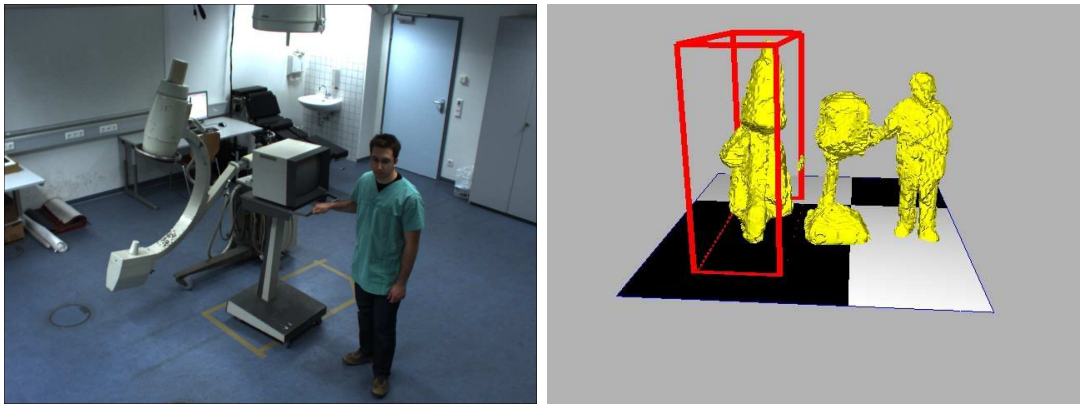


Figure 3.7: External view of a potential collision between a screen and a C-arm. Courtesy of [Ladikos et al., 2008b].

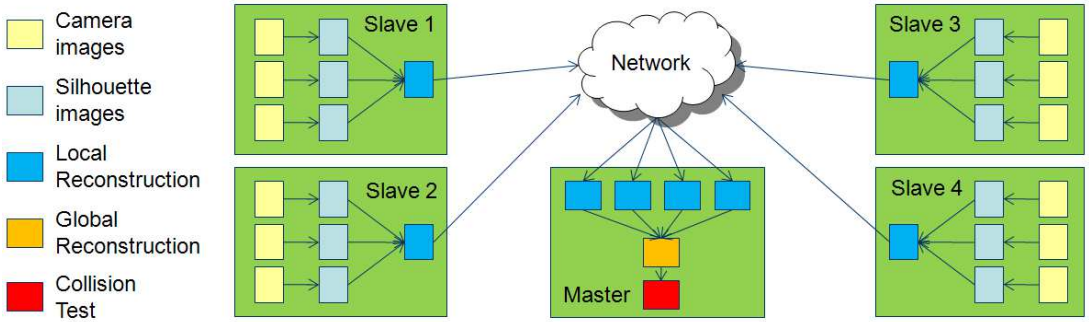


Figure 3.8: Multi-camera reconstruction system layout, describing tasks accomplished by the clients and server. Courtesy of [Ladikos et al., 2008b].

Calibration The calibration process is performed once for a fixed camera configuration. Following a method proposed by [Svoboda et al., 2005], a point light source is moved inside the dimmed room to easily compute point correspondences between the temporally synchronized cameras. This permits the computation of the intrinsic and extrinsic parameters of each camera [Hartley and Zisserman, 2004]. Registration to the room coordinate system is obtained by using a fixed calibration pattern, located on the floor of the room.

Reconstruction For reconstruction, a visual hull approach in voxel representation [Laurentini, 1994, Szeliski, 1993] is used to meet real-time requirements. Segmentation of silhouettes is performed using a robust background subtraction algorithm [Fukui et al., 2006]. Reconstruction is then obtained by testing the occupancy of each voxel in the reconstruction volume. This is done by backprojecting the voxel position into the segmented silhouette image, using a GPU implementation on each client computer for fast computation.

3.2.2 Data Representation and Acquisition

3.2.2.1 Visual Hulls

For our work, we consider the input to be the reconstructed volumes delivered by the existing real-time algorithm. Additional information, such as color would be of interest. Textured reconstruction was however not available in real-time, but if needed, the texture information could be introduced in the framework with few changes.

Let $\Omega \subset \mathbb{R}^3$ be the spatial area that can be reconstructed by the system. The output of the reconstruction at each time step is a 3D volume containing occupation probabilities. The probabilities are thresholded to obtain the effective visual hull. In a sequence of length T, we denote the sequence of computed visual hulls by

$$\{\mathbf{r}_{1:T}\}, \quad \mathbf{r}(\mathbf{v}) : \Omega \rightarrow \{0, 1\}, \tag{3.1}$$

where $\mathbf{v}$ denotes the position of a voxel.

Examples of visual hulls reconstructed by the system can be seen in Figures 3.7, 3.10 and 3.11.

3.2.2.2 Scenario and Acquisitions

Figure 3.9 displays the scenario and phases that we have defined for this application. The scenario incorporates the main activities that occur daily in a surgery room before and after surgery and is complementary to the cholecystectomy scenario studied previously. We consider two kinds of surgeries: endoscopy and open surgery, both requiring different setups. The aim is to prove the concept of using reconstructed data for recognition, until data from the hospital system becomes available. We recorded in a mock-up OR different instances of the workflow, performed by actors familiar with the OR. The phases of this workflow are illustrated in Figure 3.11 on page 40. The scene contained simultaneously up to three persons, three tables and a ceiling OR light.

A nine-camera system based on three clients was used in our experiments. As we recorded the videos from all cameras in addition to the reconstructed volumes, we used a frame-rate of 15fps to spare disk space. A higher framerate was not necessary for our recognition objectives. The reconstruction grid resolution is $128 \times 128 \times 128$ with a voxel size of about $2.8 \times 2.5 \times 1.7$ cm^3.

3.3 Conclusion

In this chapter, we have introduced the two applications considered for workflow analysis in this thesis: the laparoscopic cholecystectomy and the daily OR workflow. We have presented the setups that are used for the recordings, as well as ongoing efforts to automatically gather signals in the OR. We have also indicated the scenarios and the observations that are going to be used as input data for recognition in the next part.

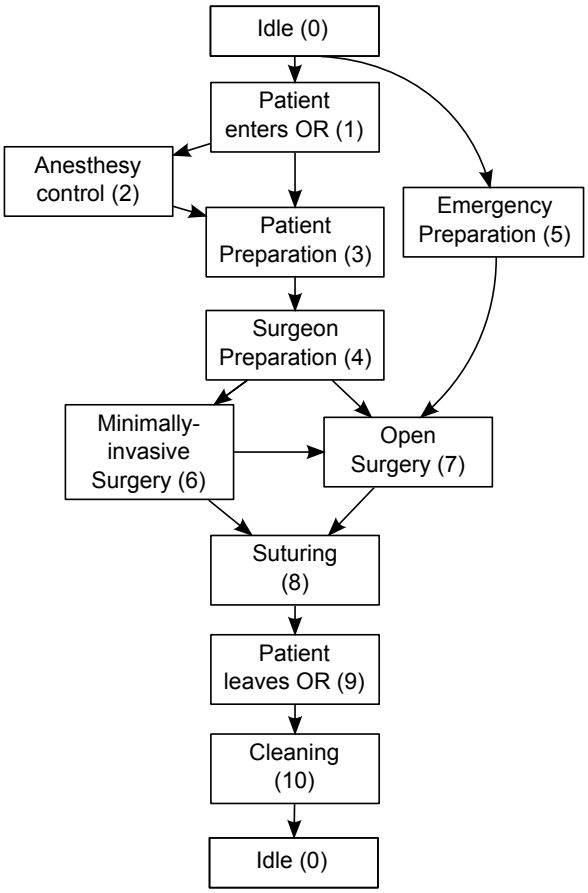


Figure 3.9: Scenario describing the surgery workflow with its alternative paths. Phase labels are given in parentheses.

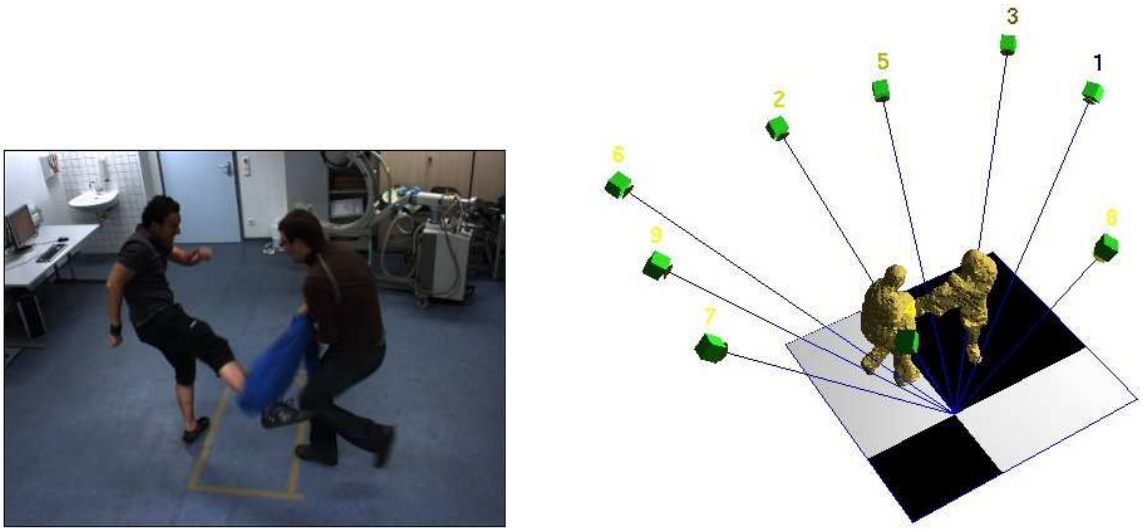


Figure 3.10: External view and reconstruction of a fighting scene, with indication of the nine camera positions. The reconstruction is viewed from the direction opposite to the camera view shown on the left.

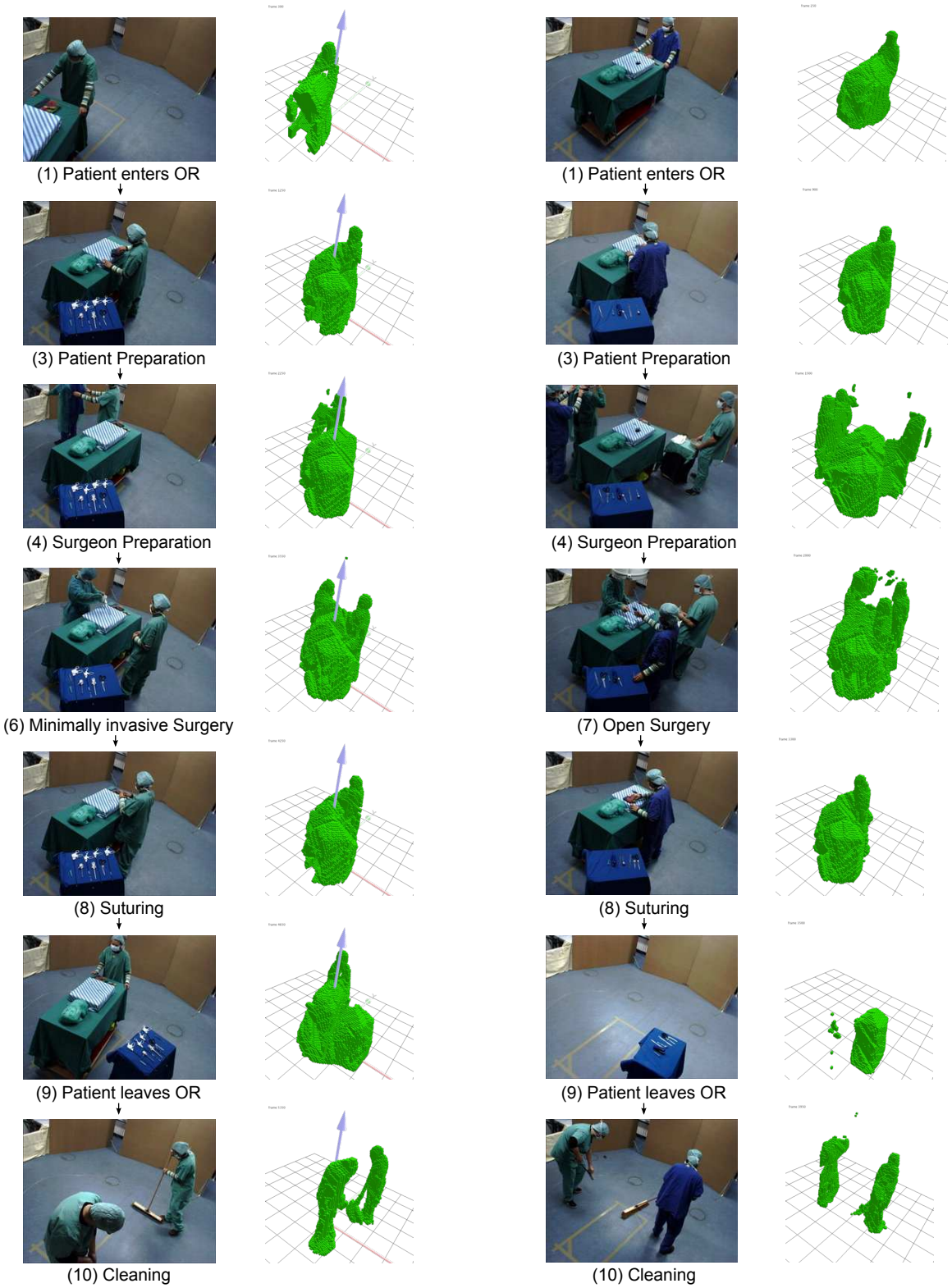


Figure 3.11: Illustration of the daily OR scenario for two workflow instances on a bavarian patient. Images from one view and associated 3D real-time reconstructions. Left) A minimally-invasive surgery. Right) An open-surgery.

Part II

Methods for Monitoring in the Surgery Room

In Chapter 4, we present off-line methods based on dynamic time warping synchronization for segmentation of the surgical phases in a sequential workflow. We compare the methods and show results for the cholecystectomy application. In order to perform on-line phase recognition, and also to cope with non-sequential workflows, we later use a different modeling based on hidden Markov models. An introduction to hidden Markov models is given in Chapter 5. The on-line recognition methods are presented in Chapter 6 and first demonstrated on the cholecystectomy application. The daily OR workflow application is addressed in Chapter 7. Computation of 4D features is described as well as the adaption of the method to a workflow containing alternative paths.

CHAPTER 4

Synchronization and Segmentation of Endoscopic Surgeries

In this chapter, we present several off-line approaches for synchronizing endoscopic surgeries and for segmenting their phases. The instruments used during a phase vary and the presence of a particular instrument is generally not sufficient to characterize the phase. The temporal sequence of actions plays indeed a decisive role. For this reason, the proposed approaches use a temporal model of the surgery. They are based on extensions of the dynamic time warping (DTW) algorithm and permit several applications such as the simultaneous visual replay of surgeries and the drafting of surgical reports. We first present the DTW algorithm in section 4.2.1 and describe related work. Synchronization between surgeries for the creation of an *virtual surgery representation* is presented in section 4.2.3. In section 4.3, this representation is used to construct an *annotated virtual surgery representation*, which is further employed for segmentation. Three construction approaches are presented, which are based on a different usage of supervised information. In section 4.3.4, an adaptive weighting method for DTW based on AdaBoost is proposed. This allows identification of the discriminative instruments and results in improved segmentation.

4.1 Objectives

In the following, we denote surgeries by the time-series $\mathbb{O}$ of their observations, as defined in section 3.1.2.1.

4.1.1 Synchronization

A first objective is to synchronize several surgeries. Synchronization can for instance be used to display surgeries simultaneously in a visually meaningful manner. Given l surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$ of length $T^1, \ldots, T^l$, this implies the computation of a common timeline $\mathcal{T} =$

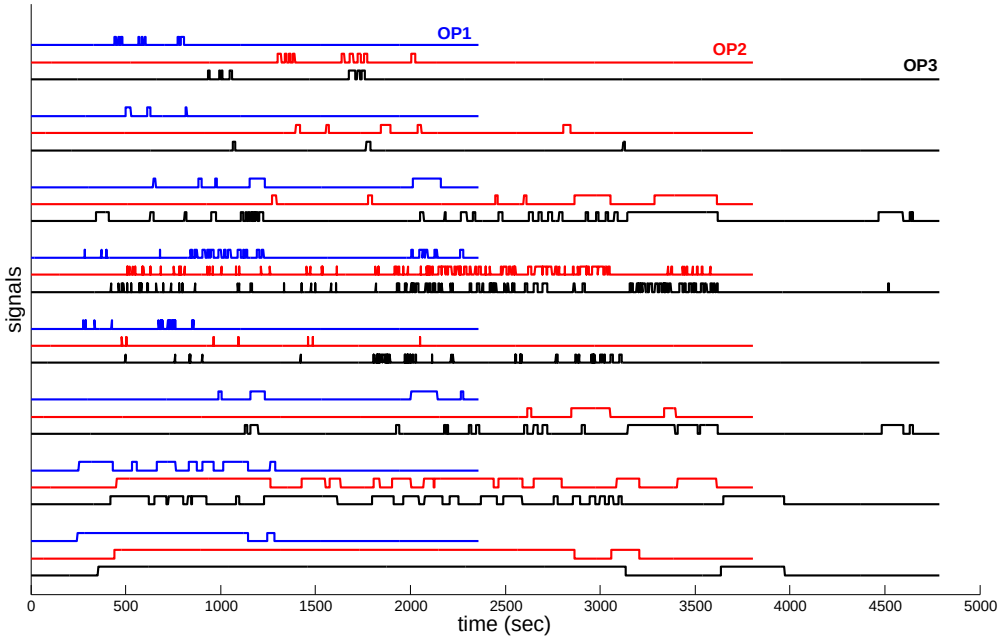


Figure 4.1: Overlaid signals from three surgeries, showing the signal variations and need for non-linear time warping. Only a subset of the signals is represented for better visualization.

$\{1, \ldots, \overline{T}\}$ of length $\overline{T}$, and of synchronization functions indicating the mapping between the timeline and each surgery $\mathbb{O}^i$:

$$sync_i : \mathcal{T} \rightarrow \{1, \ldots, T^i\} \quad . \tag{4.1}$$

Surgical gestures are unique, depending on patient and surgeon. For this reason, the same surgical task can be performed differently in different cases, even when the same surgeon is performing. For instance, clipping of a vessel can be done using from one up to three clips. This implies that there exists no ground truth synchronization between different surgeries and that "good" synchronizations are not unique. A synchronization can be evaluated subjectively at the semantical level, by verifying that certain kinds of semantic information about the surgeries match at each time step of the reference timeline, in all surgeries. Semantic information that we consider here are the surgical phases. Synchronization at the lower level, even though not unique, can for instance provide a nice and smooth visualization. It is visually much better than a simple scaling between validated synchronization points. Figure 4.1 shows several signals from three surgeries simultaneously to illustrate the need for non-linear synchronization between the surgeries.

4.1.2 Segmentation

A second objective is the accurate identification of the surgical phases of a new surgery $\mathbb{O}$ of length T, based on the knowledge of previously acquired surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$. If we assign to each phase a label $p \in \mathcal{L}$, this involves the computation of a labeling function

$$\mathcal{P}_{\mathbb{O}} : \{1, \ldots, T\} \rightarrow \mathcal{L} \ . \tag{4.2}$$

An example of labels for cholecystectomy phases is indicated in Table 3.1 on page 33.

4.2 Dynamic Time Warping Averaging

In this section, we present a method to construct the multi-dimensional signals representing a virtual surgery. The surgeries in the training set are all synchronized using the dynamic time warping algorithm. This is then employed to generate a virtual surgery on a virtual timeline, preserving the average length of the surgeries and of the actions. This virtual surgery can be used for off-line synchronization and segmentation of new surgeries. It can also be used for an efficient generation of the hidden Markov models. The presentation addresses complete surgeries. But the method can also be applied directly to only subparts of surgeries, such as single phases (see Section 4.3.1.2).

4.2.1 Dynamic Time Warping

The dynamic time warping (DTW) algorithm [Sakoe and Chiba, 1978] is both a time-invariant similarity measure and a method to synchronize two time series by finding a non-linear warping path. It warps 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 that minimizes the sum of the distances between all points that are warped onto each other.

Formally, let $u = (u_1, \ldots, u_{T^u})$ and $v = (v_1, \ldots, v_{T^v})$ be two time-series of length T^u and T^v, and $d(\cdot, \cdot)$ be a distance. The algorithm uses dynamic programming to find the discrete timeline $\{1, \ldots, T\}$ and temporal warping path $\text{sync}_{u \leftrightarrow v} = (t_u, t_v)$ with

$$t_u : \{1, \ldots, T\} \rightarrow \{1, \ldots, T^u\} \tag{4.3}$$
$$t_v : \{1, \ldots, T\} \rightarrow \{1, \ldots, T^v\} \tag{4.4}$$

minimizing

$$\sum_{t=1}^{T} d(u_{t_u(t)}, v_{t_v(t)}) \tag{4.5}$$

under the warping constraints:

$$\text{boundary} \quad \begin{cases} (t_u(1), t_v(1)) &= (1, 1) \\ (t_u(T), t_v(T)) &= (T^u, T^v) \end{cases} \tag{4.6}$$

$$\text{continuity} \quad \begin{cases} |t_u(t+1) - t_u(t)| &\leq 1 \\ |t_v(t+1) - t_v(t)| &\leq 1 \ . \end{cases} \tag{4.7}$$

Different warping functions can be obtained by using variations of the continuity constraints [Sakoe and Chiba, 1978]. The optimal solution is found using dynamic programming in $O(T^u \times T^v)$. A distance matrix D is computed according to the following

recursive formula

$$D_{i,j} = d(u_i, v_j) + min(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1}) \ . \tag{4.8}$$

Backtracking the optimal path provides the warping, as illustrated in Figure 4.2. The timeline $\{1, \ldots, T\}$ corresponds to a discretization of the optimal path through the valley.

4.2.2 Related Work on Dynamic Time Warping

The DTW algorithm originates from the speech recognition community [Juang, 1984, Sakoe and Chiba, 1978, Yaniv and Burshtein, 2003] where it was extensively used to model spoken words. It was further used in many different research areas to synchronize application-dependent time-series. It is for instance used in computer vision for the modeling of actions [Darrell et al., 1996], in chemistry for the synchronization of batch processes [Kassidas et al., 1998] or for the mining of large databases [Keogh and Ratanamahatana, 2005].

Our purpose is to use DTW for the modeling of surgeries represented by binary signals. For a medical application, a DTW synchronization approach was proposed in our group by [Sielhorst et al., 2005]. The movements performed by a trainee and an expert surgeon on a birth simulator were recorded with a tracking system. DTW permitted to replay and compare synchronously on an augmented reality display the superimposed movements performed by the two surgeons. An initial work to synchronize six endoscopic surgeries was proposed in [Ahmadi et al., 2006]. A surgical model was created using DTW and manually labeled to segment a new surgery. Our approach extends this work by improving the virtual model creation, by proposing a convenient framework for dealing with phase labeling information, and also by presenting cross-validated results on a larger dataset.

4.2.3 Averaging

The construction of a virtual representation is based on a curve averaging method by [Wang and Gasser, 1997]. In the following, we adapt the presentation to the discrete case that is considered in this chapter. Let the l surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$ be of length $T^1, \ldots, T^l$ and $\mathbb{O}^{ref}$ of length T^{ref} be a surgery taken as reference.

The DTW algorithm with euclidean distance is first used to synchronize the reference $\mathbb{O}^{ref}$ to each training surgery $\mathbb{O}^i$, yielding for $i \in \{1, \ldots, l\}$ the synchronization functions

$$\text{sync}_{ref \leftrightarrow i}(y) = (t_{ref}(y), t_i(y)) \ . \tag{4.9}$$

These functions give discrete correspondences between the timelines of the reference (t_{ref}) and of the surgeries (t_i).
The virtual timeline is computed from the reference as the function $\text{avg}_{time}(t)$:

$$\begin{array}{rcl} \{1, \ldots, T^{ref}\} & \rightarrow & [1, \overline{T}] \\ t & \rightarrow & \frac{1}{l} \sum_{i=1}^{l} \frac{1}{\#\{y:\ t_{ref}(y)=t\}} \sum_{\{y:\ t_{ref}(y)=t\}} t_i(y) \ , \end{array} \tag{4.10}$$

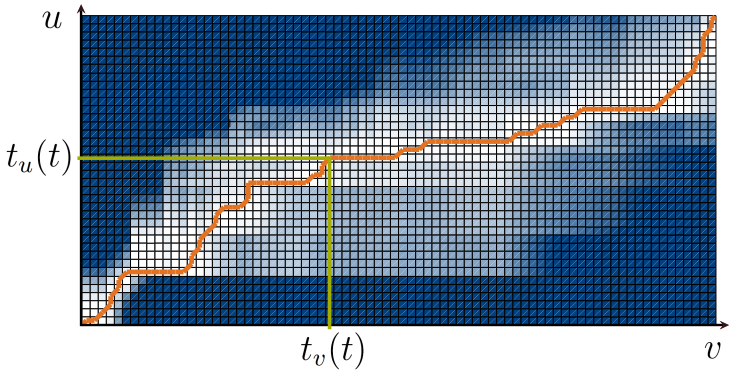


Figure 4.2: DTW distance matrix. The path drawn through the valley displays the optimal synchronization between time series u and v.

where $\#\cdot$ denotes the cardinality operator. The function avg_{time} takes real values, is monotonically increasing between 0 and $\overline{T} = \frac{1}{l}\sum_i^l T^i$ and can therefore be inverted. It is used to compute the virtual surgery representation $\overline{\mathbb{O}}$ on the discrete timeline $\{1, \ldots, \lfloor \overline{T} \rfloor\}$ 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 $t \in [t_a, t_b]$. Noting $\mathcal{I}_a = \{y : \ t_{ref}(y) = \text{avg}_{time}^{-1}(t_a)\}$ and $\mathcal{I}_b = \{y : \ t_{ref}(y) = \text{avg}_{time}^{-1}(t_b)\}$, we then define

$$\overline{\mathbb{O}}_t = \frac{1}{l}\sum_{i=1}^{l} \frac{(t-t_a)}{(t_b-t_a)} \frac{1}{\#\mathcal{I}_a} \sum_{y\in\mathcal{I}_a} \mathbb{O}^i_{t_i(y)} + \frac{(t_b-t)}{(t_b-t_a)} \frac{1}{\#\mathcal{I}_b} \sum_{y\in\mathcal{I}_b} \mathbb{O}^i_{t_i(y)} \ . \tag{4.11}$$

The virtual surgery $\overline{\mathbb{O}}$ has a timeline whose length is the average time between all training surgeries. Its signals $\overline{\mathbb{O}_t}$ can be interpreted as the probability of instrument usage at virtual time t. Provided an action was correctly synchronized between all surgeries, it then appears with its average length within $\overline{\mathbb{O}}$.

Similarly to [Wang and Gasser, 1997], we use three steps for our virtual surgery computation:

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

There are various ways to choose the initial reference. It could simply be one of the training surgeries. Experiments have shown the following approach to yield the best results. The reference is computed recursively by averaging pairwise the surgeries using the approach described above. In order to average the representation of two surgeries $\mathbb{O}^i$ and $\mathbb{O}^j$, we introduce a special surgery as the reference, which we will refer to as the exhaustive surgery. This is useful in order to avoid that an activity from one surgery is reduced to zero time-span if it is synchronized to a reference in which this activity did not occur. Consequently, the exhaustive model preserves all information from both surgeries by merely taking the raw result of their DTW synchronization, which has the same length

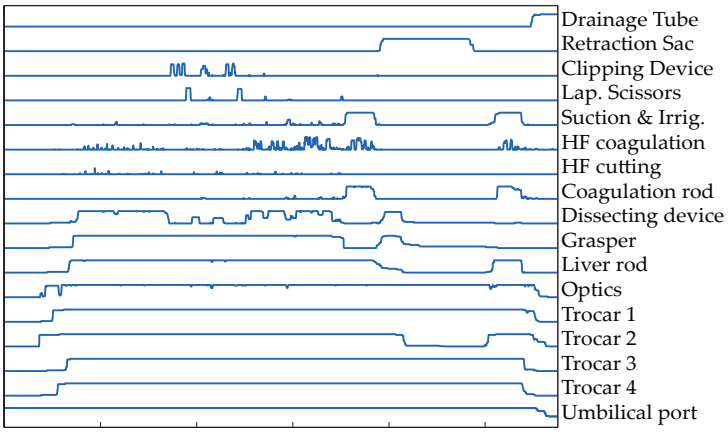


Figure 4.3: Signals of a virtual surgery representation.

as the size of the output range, namely $\text{sync}_{i \leftrightarrow j}(y) = (t_i(y), t_j(y))$:

$$\mathbb{O}_y^{ref\_exhaust} = \frac{\mathbb{O}^i_{t_i(y)} + \mathbb{O}^j_{t_j(y)}}{2} \quad . \tag{4.12}$$

Using this reference prevents the complete disappearance of actions that might not occur in one surgery. Moreover, when the virtual timeline is computed, Formula 4.10 ensures that the virtual surgery keeps the average length of training surgeries.
An example of resulting virtual representation $\overline{\mathbb{O}}$ is shown in figure 4.3.

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

$$SIM_t = \frac{1}{K} \sum_{k=1}^{K} max(\overline{\mathbb{O}}_{t,k}, 1 - \overline{\mathbb{O}}_{t,k}) \quad , \tag{4.13}$$

where K is the size of the observation vectors. 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.

4.3 Segmentation

For segmentation purposes, we present a formalization that combines both the resulting virtual surgery representation and the labeling information. We define an *annotated virtual surgery representation (AVSR)* as a pair

$$\Gamma = (\overline{\mathbb{O}}, \xi) \quad , \tag{4.14}$$

where

$$\xi_t : \{1, \ldots, |\mathcal{L}|\} \rightarrow [0, 1] \tag{4.15}$$

refer to a model annotation with phase probabilities. $\xi_t(p)$ is the probability of being in phase $p \in \mathcal{L}$ at virtual timestamp t from the virtual representation $\overline{\mathbb{O}}$.

The results of segmentation depend on how the AVSR is constructed. The methods presented below differ by the required amount of labeled data and by the steps in the construction process where the labels are used.

4.3.1 Different Approaches for Model Construction

4.3.1.1 Manual Annotation

A virtual surgery representation $\overline{\mathbb{O}}$ is first computed out of all training surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$. An annotation $\xi_t \in \{0, 1\}$ is then determined manually by looking at the signals on the virtual timeline. This possibility was initially used in [Ahmadi et al., 2006]. However, even though the virtual representation visually resembles a surgery, cues that permit a precise annotation within the recorded surgeries are not so precise within the virtual surgery. Manual annotation has also the major drawback that it has to be performed each time a new surgery is included, or for each test in cross-validation experiments. Additionally, backprojecting the labels from the virtual representation onto the initial surgeries it was generated from showed poor synchronization at the label boundaries. The two methods below perform the model annotation more accurately and in an automatic fashion.

4.3.1.2 Pre-annotation

We refer by the term *pre-annotation* to the fact that labeling information is used prior to the construction of the virtual representation. This is a fully supervised framework, in which the DTW averaging approach is applied phase-wise. The virtual surgery representation is then constructed by the concatenation of virtual phase representations computed out of all training surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$. All training surgeries are labeled so that the phases are identified.

Let $\mathbb{P}^i_p$ be the phase from surgery $\mathbb{O}^i$ corresponding to label $p \in \mathcal{L}$. The virtual phase $\overline{\mathbb{P}_p}$ is constructed following the averaging framework described in Section 4.2.3, by using as input the l subsequences $\mathbb{P}^1_p, \ldots, \mathbb{P}^l_p$.

The resulting virtual surgery representation $\overline{\mathbb{O}}$ is the concatenation of the virtual phase representations $\overline{\mathbb{P}_1}, \ldots, \overline{\mathbb{P}_{|\mathcal{L}|}}$. For each time t on the resulting virtual timeline, the annotation ξ_t assigns the probability 1 to the phase p, given that the virtual time t originally stemmed from the virtual phase representation p before concatenation. To the other phases, the probability 0 is assigned.

4.3.1.3 Post-annotation

A virtual surgery representation $\overline{\mathbb{O}}$ is first computed out of all training surgeries as in Section 4.2.3. No surgery labeling information is required for this step. The annotation ξ_t is obtained by synchronizing the virtual representation to the training surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$ that contain labeling information. For these surgeries, labeling functions

$\mathcal{P}_{\mathbb{O}^i} : \{1, \ldots, T^i\} \rightarrow \mathcal{L}$ are available. The synchronization performed using DTW yields l functions

$$\text{sync}_{\overline{\mathbb{O}} \leftrightarrow \mathbb{O}^i}(y) = (t_{avg}(y), t_i(y)) \ . \tag{4.16}$$

The annotation is then determined from the labeling of these surgeries:

$$\xi_t(p) \propto \sum_{\{y:\ t_{avg}(y)=t\}} \sum_{i=1}^{l} \delta\left(\mathcal{P}_{\mathbb{O}^i}(t_i(y)), p\right) \tag{4.17}$$

with

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \ . \end{cases} \tag{4.18}$$

For each virtual timestep t, $\xi_t(\cdot)$ is normalized to one.

4.3.2 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 $\mathcal{P}_{\mathbb{O}^{test}}(t)$ at each time step t while knowing the *complete* signals $\mathbb{O}^{test}_1 \ldots \mathbb{O}^{test}_{T^{test}}$, where T^{test} denotes the end of the surgery. If an annotated virtual surgery representation $\Gamma = (\overline{\mathbb{O}}, \xi)$ is available from training data, the process is the following: the time-series $\mathbb{O}^{test}$ is synchronized to the virtual representation $\overline{\mathbb{O}}$ using DTW. This gives a synchronization function

$$\text{sync}_{\mathbb{O}^{test} \leftrightarrow \overline{\mathbb{O}}}(y) = (t_{test}(y), t_{avg}(y)).$$

The labels from the annotated virtual representation 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 virtual representation annotated with different most likely phases, the overall most likely annotation is used:

$$\mathcal{P}_{\mathbb{O}^{test}}(t) = \underset{p}{\text{argmax}} \sum_{\{y:\ t_{test}(y)=t\}} \xi_{t_{avg}(y)}(p) \ . \tag{4.19}$$

4.3.3 Evaluation

Segmentation is evaluated by comparison to a ground truth labeling. We present below evaluation measures and the obtained results.

4.3.3.1 Measures

We first use **accuracy** as general evaluation measure. It indicates the percentage of correct detections in the complete surgery, compared to ground truth information. As sizes between phases can vary largely, wrong detections inside short phases tend to be hidden within the accuracy. For this reason, we also use two measures defined *per phase*. **Recall** is the number of true positives divided by the total number of positives for the phase in the ground truth. In other words, recall is the percentage of correct detections inside each phase.

Precision is the sum of true positives divided by the number of true and false positives. This is complementary to recall be indicating whether parts of other phases are detected incorrectly as the considered phase.

In order to present summarized results, we will use **accuracy** together with **average recall** and **average precision**, corresponding to recall and precision averaged over all phases.

4.3.3.2 Results

We have evaluated the methods on data of 16 cholecystectomies performed by 4 different surgeons, using the acquisitions detailed in Chapter 3. For statistical relevance, the results are computed with cross-validation using the leave-one-out approach: for each of the 16 surgeries, the training set contains the remaining 15 ones. Displayed values are the mean results with standard deviations over all surgeries.

Comparison of the two evaluated construction approaches, pre-annotation (Section 4.3.1.2) and post-annotation (Section 4.3.1.3) is provided in Table 4.1. Both methods yield a very good segmentation with measures above 94%. The pre-annotation method, which fully relies on supervised construction, has the best values and has also a slightly lower variance. Interestingly, only a few labeled surgeries are required to obtain such results with the post-annotation method. This can be observed in Figure 4.4 showing all measures as function of the number of labeled surgeries used for annotation. During this experiment, tests are averaged on randomly selected subsets of varying size. The figure suggests that labeling only forty percent of the available surgeries is sufficient to obtain good segmentation results for this dataset.

The detailed results per phase, using the pre-annotation method are displayed in Table 4.2. All phases have measures above 90%. Lowest rates principally come from short phases. This can be seen from the average relative length of each phase with respect to the complete surgeries. It indeed occurs that some short phases become extremely short in some surgeries. 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, which causes higher relative errors. Due to the variations of phase lengths, relative errors are more indicative of the goodness of the segmentation than absolute errors in minutes. In practice, they however hide the temporal effect on the application timeline.

Figure 4.5 displays in absolute time the average length of each phase and the *mean error per phase*, corresponding to the duration in minutes of incorrect detections per phase. These errors were obtained for the pre-annotation method. The maximum mean error per phase occurs for phase 8 (liver bed coagulation 1) and is of 9 seconds. The average length among all surgeries is 48 minutes.

4.3.4 Adaptive DTW

In [Ahmadi et al., 2006], a method has been proposed to weight signals iteratively during the virtual surgery construction. Signals that are correctly synchronized to the virtual surgery get iteratively a higher weight. An issue with this approach, as reported in [Ahmadi, 2005], is that instruments which should intuitively get higher weights, such

	Accuracy (%)	Average Recall (%)	Average Precision (%)
AVSR (pre)	97.3 (±6.6)	97.6 (±5.6)	97.0 (±5.7)
AVSR (post)	95.1 (±6.6)	95.5 (±6.0)	94.0 (±6.2)

Table 4.1: Leave-one-out cross-validation on 16 surgeries performed by 4 surgeons. Global measures with mean and standard deviation over all surgeries. (pre) indicates pre-annotation, (post) construction with post-annotation.

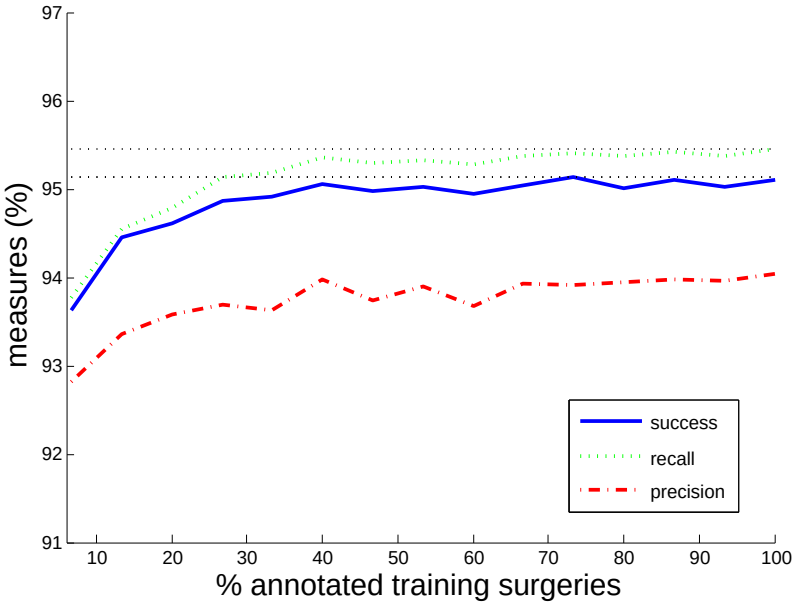


Figure 4.4: Accuracy (rate of success), average recall and average precision for AVSR with post-annotation. Influence of number of annotated surgeries. The horizontal lines refer to the best result for each curve, obtained when all training surgeries are labeled.

	Phase Label	Rel. Len. (%)	Recall (%)	Precision (%)
1	CO2 Inflation	6.5 (±3.2)	100.0 (± 0.0)	100.0 (± 0.0)
2	Trocar Insertion	6.4 (±1.4)	100.0 (± 0.0)	99.3 (± 1.8)
3	Dissection Phase 1	17.1 (±9.1)	99.9 (± 0.4)	100.0 (± 0.0)
4	Clipping Cutting 1	3.8 (±1.6)	100.0 (± 0.0)	96.6 (± 13.2)
5	Dissection Phase 2	3.7 (±5.0)	93.8 (± 24.2)	92.8 (± 24.3)
6	Clipping Cutting 2	3.9 (±1.9)	92.8 (± 24.3)	90.9 (± 25.9)
7	Gallbladder Detaching	18.9 (±13.1)	95.1 (± 17.6)	93.5 (± 21.2)
8	Liver Bed Coagulation 1	7.1 (±4.4)	90.8 (± 20.5)	100.0 (± 0.0)
9	Gallbladder Packaging	4.1 (±2.8)	99.6 (± 1.6)	97.4 (± 5.2)
10	External Retraction	11.7 (±11.7)	96.7 (± 12.8)	99.5 (± 1.6)
11	External Cleaning	3.6 (±4.1)	95.8 (± 12.2)	94.5 (± 21.2)
12	Liver Bed Coagulation 2	6.2 (±2.1)	99.6 (± 1.0)	98.5 (± 5.3)
13	Trocar Retraction	2.6 (±1.9)	99.5 (± 1.3)	98.9 (± 4.2)
14	Abdominal Suturing	3.9 (±3.4)	100.0 (± 0.0)	93.5 (± 19.5)

Table 4.2: Detailed results per phase for the pre-annotation approach, with mean and standard deviation over all surgeries. The third column (*Rel. Len.*) indicates the average relative length of each phase with its standard deviation.

as the dissecting device, effectively get lower weights than instruments deemed less informative, such as a trocar. This comes from the fact that instruments whose state almost does not change have much higher chances to be well synchronized, and therefore to get a higher weight. An instrument which is unused and also does not change state at all, receives an infinite weight with this procedure. All other instruments which do provide the contextual information will be ignored. For instance, instruments that are used frequently during a phase have less chances to be correctly synchronized for each use, especially as the number of uses can vary highly within different surgeries. However, such an instrument brings interesting discriminative information for the phase it is used in.

This discussion suggests that to compute a signal weighting, the signals' potential to *discriminate between the different phases* has to be evaluated, instead of the signals' potential to be well synchronized to the virtual surgery. We propose a discriminative weighting approach based on weights *per phase* [Padoy et al., 2007a].

The idea is to apply the DTW algorithm 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 [Freund and Schapire, 1995] using the labeled training surgeries. AdaBoost has been widely used for feature selection [Viola and Jones, 2004] and provides a natural way for feature weighting.

This information is then included within the DTW averaging process to create the virtual surgery representation out of labeled training surgeries. To synchronize an unsegmented surgery to the model, an adaptive version of DTW, called ADTW, is used. As above, labels from the virtual model can be carried over to an unsegmented surgery using this synchronization.

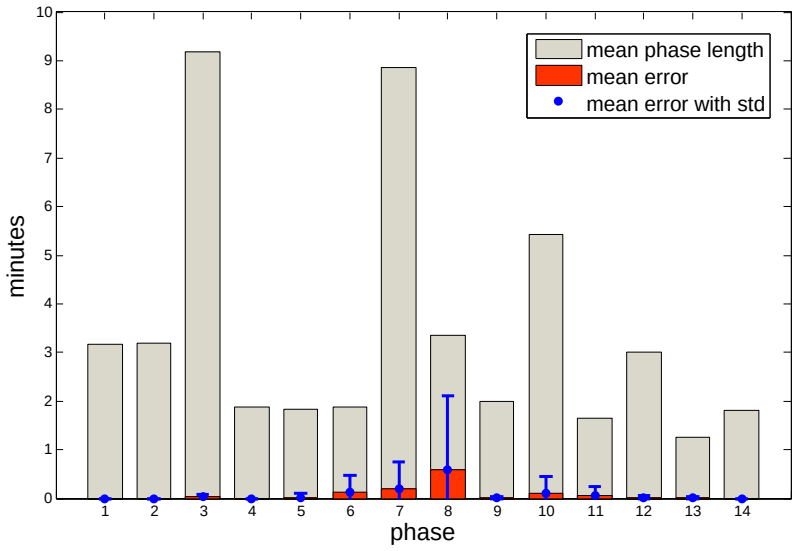


Figure 4.5: For each phase, average length over all surgeries overlayed with mean detection error per phase, in *minutes*. Errors are computed for AVSR with pre-annotation construction.

4.3.4.1 Discriminative Weighting

In this section, we weight the instruments to reflect their ability to discriminate between neighboring phases. When synchronizing a surgery to the virtual surgery model, using these weights, the ADTW algorithm will put a higher priority on the most significant instruments for each phase.

AdaBoost [Freund and Schapire, 1995] 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 $p \in \mathcal{L}$, 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 $C_{n,s}$ classifies an instrument vector O according to whether the state of the instrument n within the vector is equal to s. AdaBoost selects, at each step i, a classifier C_{n_i,s_i} and a weight θ_i to construct the strong classifier:

$$SC = \sum_i \theta_i C_{n_i,s_i} \quad . \tag{4.20}$$

The variables n_i and s_i indicate the instrument and its state selected at step i. As the algorithm re-weights 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:

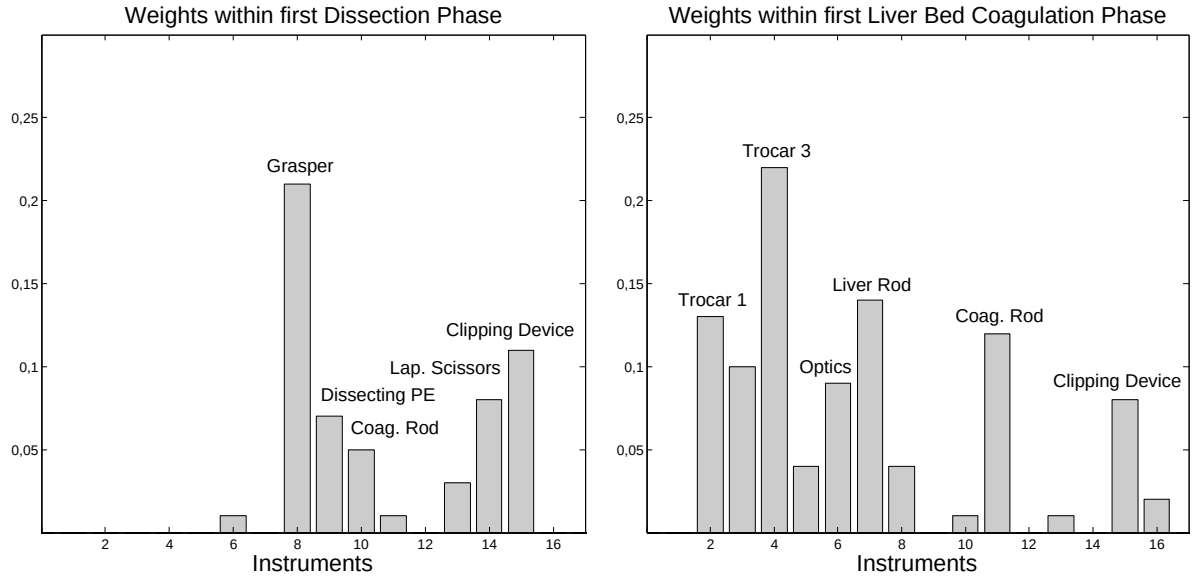


Figure 4.6: Instrument weights computed for the first dissection phase (phase 3) and the first liver bed coagulation phase (phase 12)

$$w_k^{(p)} = \left| \sum_{n_i=k, s_i=1} \theta_i - \sum_{n_i=k, s_i=0} \theta_i \right| \quad . \tag{4.21}$$

They are then normalized to one. As they are computed for each phase, this leads to weights $w_k^{(p)}$, for all phase $p \in \mathcal{L}$ 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. 4.6 displays the computed weights for two phases. In the first dissection phase (see section 3.1.2.3), the most significant three 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 trocars 1 and 3 as well as the liver rod, which all have to be in use.

4.3.4.2 Segmentation with ADTW

An annotated virtual surgery representation is first computed using the pre-annotation approach, with the difference that for each virtual phase $\overline{\mathbb{P}_p}$, DTW is used with the weighted distance d_p corresponding to the phase p:

$$d_p(O_1, O_2) = \sqrt{\sum_{k=1}^{K} w_k^{(p)} (O_{1,k} - O_{2,k})^2} \quad , \tag{4.22}$$

where O_1 and O_2 are two observation vectors. We call the extension of DTW by using a phase-dependent distance d_p Adaptive DTW (ADTW).

For segmentation, an unsegmented surgery is warped with ADTW onto the annotated virtual surgery representation. As in section 4.3.2, this allows carrying over the annotation.

4.3.4.3 Evaluation

The approach was used to determine the most significant instruments in a phase, as shown in Figure 4.6. Additionally, segmentation results using ADTW compared to the previous DTW approaches are presented in Table 4.3. The three success measures are higher and standard deviations much lower. Interestingly, all phases were detected for all surgeries during the cross-validation, while for the pre-annotation approach, for one surgery a few very short phases were not detected.

	Accuracy (%)	Average Recall (%)	Average Precision (%)
AVSR (pre)	97.3 (±6.6)	97.6 (±5.6)	97.0 (±5.7)
AVSR (post)	95.1 (±6.6)	95.5 (±6.0)	94.0 (±6.2)
AVSR (adap)	98.5 (±3.9)	99.2 (±1.5)	99.2 (±1.2)

Table 4.3: Leave-one-out cross-validation on 16 surgeries performed by 4 surgeons. Global measures with mean and standard deviation over all surgeries. (pre) indicates pre-annotation, (post) construction with post-annotation, (adap) construction with ADTW.

4.4 Applications

4.4.1 Synchronous Visual Replay

Using the virtual representation computed with the method presented in section 4.2.3, a synchronization between the surgeries is available. To replay simultaneously several surgeries (see Figure 4.7), the surgeries are played based on the virtual timeline, using their respective time warping functions. When only two surgeries need to be simultaneously displayed, it is visually more accurate to use the global virtual representation built from all available surgeries than to use a DTW synchronization between these two surgeries only.

4.4.2 Training

The presented methods are also very interesting for training purposes. Indeed, they could provide quantitative results on the performance of a training surgeon throughout the surgery, which is possible and relevant when the model has been built from data recorded from expert surgeons. Additionally, synchronizing two surgeries to the virtual surgery representation provides an accurate synchronization between them. This 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 by an expert surgeon.

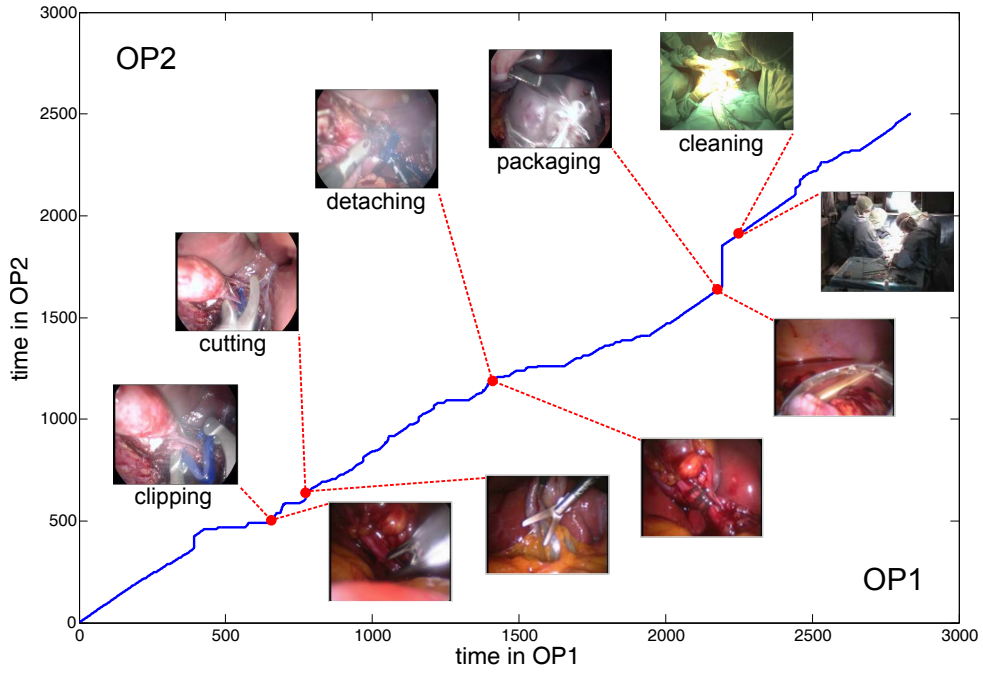


Figure 4.7: Simultaneous video replay of two surgeries, based on the temporal warping path displayed in the middle.

4.4.3 Reporting

Writing the report is a time-consuming post-operative task for the surgeon, but is inevitable due to legal reasons. The accurate phase identification in off-line segmentation can be used to sketch a precise and objective report, containing each phase and their starting times. All data recorded during the surgery could also be linked to an electronic report of a surgical procedure. Automatic report generation can certainly not completely replace manual writing. But it can shorten the report writing by providing a partially filled report template.

4.5 Conclusion

The methods presented above are based on the construction of a surgical model that we named *annotated virtual surgery representation*. They permit simultaneous replay and accurate segmentation of surgeries, as needed for instance for the automatic drafting of surgical reports. When using the ADTW approach, recall and precision above 99% can be obtained for the segmentation. When automatic signal acquisitions become available on a daily basis, it will be possible to test the system directly in the OR. Such a system will also allow the regular recording of additional surgeries, which can be used for broader statistical validation and also comparison between trainee and expert surgeons.

The modeling, however, does not easily permit on-line recognition of the phases and cannot naturally cope with non sequential workflows. A more flexible modeling of both

observations and temporal constraints is required and could be provided by well-chosen probabilistic frameworks. For the requirements of on-line recognition within a workflow, we will use a variant of HMMs. The next chapter gives an introduction to these models, since they will be used later on.

CHAPTER 5

Hidden Markov Models

Hidden Markov models (HMMs) and variations thereof are probabilistic models that have been widely used in recognition problems involving the modeling of time-series. They are the core of the on-line recognition methods presented in chapters 6 and 7 for the detection of the current surgical phase, based on data acquired from cholecystectomies or from the multi-view reconstruction system. In this chapter, we introduce the common notations and algorithms related to HMMs. They belong to a class of Graphical Models that can be very conveniently represented in terms of Dynamic Bayesian Networks (DBNs). As many variations of HMMs are usually presented as DBNs in the literature, we briefly introduce DBNs at the end of this chapter. For a more detailed introduction to HMMs and DBNs, we refer the reader to respectively [Rabiner, 1989] and [Murphy, 2002].

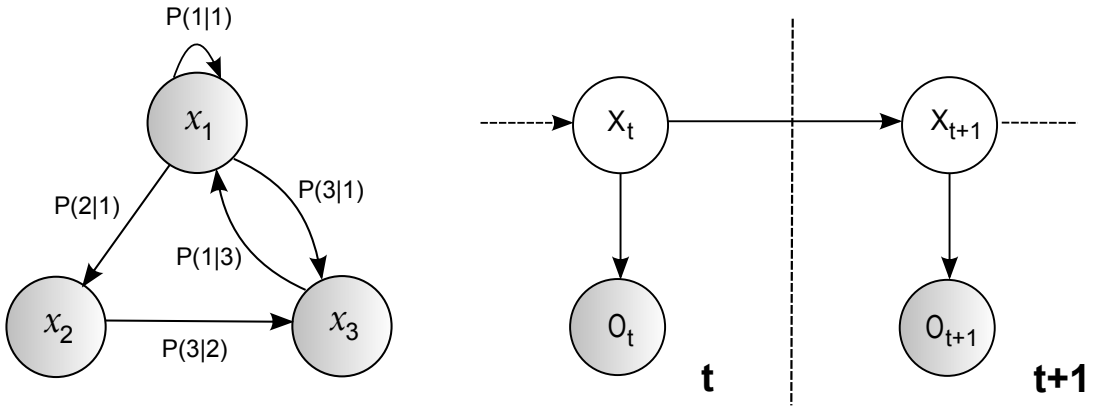
5.1 Hidden Markov Models

We first present Markov chains in Section 5.1.1. Hidden Markov models, which can be viewed as an extension of Markov chains by adding an observation model, are introduced in Section 5.1.2.

5.1.1 Markov Chains

A Markov chain models a discrete-time stochastic process with a finite set of N states $\{x_1, \ldots, x_N\}$ connected by transition probabilities. The evolution in time is represented by random variables $X_t \in \{x_1, \ldots, x_N\}$ for $t \geq 1$ and probabilities $P(X_{t+1} = x_j | X_t = x_i)$. A Markov chain is a Markov process of first order, whose transition probabilities to the next state only depend on the previous state:

$$P(X_{t+1}|X_t, \ldots, X_2, X_1) = P(X_{t+1}|X_t) \ . \tag{5.1}$$



(a) Graphical representation of the transitions between the states of a Markov chain.

(b) Graphical representation of a hidden Markov model as a dynamic Bayesian network.

Figure 5.1: Graphical representation of a Markov chain and of a hidden Markov model. Shaded nodes indicate observed nodes. Unshaded nodes indicate hidden (or latent) nodes.

When these probabilities are also constant in time, the process is homogeneous and is simply modeled by the triple $\lambda = (N, A, \pi)$, where A is the transition matrix defined by

$$a_{i,j} = P(X_{t+1} = x_j | X_t = x_i) \tag{5.2}$$

and π is a vector of initial probabilities: $\pi_i = P(X_1 = x_i)$. A simple Markov chain is represented graphically in Figure 5.1(a).

The probability of a process to be in the sequence of states $\mathbf{x} = x_{k_1}, \ldots, x_{k_T}$ can be computed as:

$$\begin{aligned} P(\mathbf{x}|\lambda) &= P(X_1 = x_{k_1}, \ldots, X_t = x_{k_T} | \lambda) & (5.3) \\ &= p(X_1 = x_{k_1}|\lambda) \prod_t p(X_t = x_{k_t} | X_{t-1} = x_{k_{t-1}}, \lambda) \ . & (5.4) \end{aligned}$$

For recognition, Markov chains can be used in maximum a posteriori (MAP) classification, e.g. by finding the most probable model Λ among k models $\lambda_1, \ldots, \lambda_k$ to have generated the sequence $\mathbf{x}$:

$$\Lambda = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_k\}}{\operatorname{argmax}} \ P(\mathbf{x}|\lambda) \ . \tag{5.5}$$

In many practical problems, such a state sequence is not known or cannot be simply inferred from the data (for instance a video) generated by the process. HMMs handle this difficulty by supposing that the states can only be observed indirectly by observations that they generate (the states are hence called "hidden"). Observations are modeled in each state X by conditional probabilities $P(O|X)$, as defined below.

5.1.2 Hidden Markov Models

A hidden Markov model (HMM) with discrete observations is defined by a quintuplet $\lambda = (N, \mathcal{O}, A, B, \pi)$, where A and π are transition and initial state probabilities as defined

for Markov chains. B is a matrix modeling for each of the N states $\{x_1, \ldots, x_N\}$ an observation distribution over a discrete set of M symbols $\mathcal{O} = \{o_1, \ldots, o_M\}$:

$$b_{i,j} = P(O = o_j | X = x_i) \ . \tag{5.6}$$

Computing the probability of an observation sequence $\mathbb{O} = O_1, \ldots, O_T$ to have been generated by the model λ can be achieved by marginalization over the state variables:

$$\begin{aligned} P(\mathbb{O}|\lambda) &= \sum_{x_{k_1},\ldots,x_{k_T}} P(\mathbb{O}, X_1 = x_{k_1}, \ldots, X_t = x_{k_T}|\lambda) && (5.7)\\ &= \sum_{x_{k_1},\ldots,x_{k_T}} P(\mathbb{O}|x_{k_1}, \ldots, x_{k_T}, \lambda) P(x_{k_1}, \ldots, x_{k_T}|\lambda) && (5.8)\\ &= \sum_{x_{k_1},\ldots,x_{k_T}} P(x_{k_1}) P(O_1|x_{k_1}) \prod_{t=2}^{T} P(x_{k_t}|x_{k_{t-1}}) P(O_t|x_{k_t}) \ . && (5.9) \end{aligned}$$

When no ambiguity exists, dependency on λ is omitted for better readability. This computation is expensive and takes $O(2TN^T)$. Fortunately, a faster computation based on dynamic programming can be obtained in $O(N^2T)$. This computation addresses the first classical problem for HMMs. It permits classification and is presented in Section 5.1.2.1. The description of the HMMs problems and of the related algorithms is based on [Rabiner, 1989]. The second classical problem, namely the computation of the "most likely state path" in the model given an observation sequence, is described in Section 5.1.2.2. The third problem, namely the optimization of the model parameters by expectation maximization (EM), is addressed in Section 5.1.2.3. In Section 5.1.2.4, we present the modeling in case of continuous observations. Some literature addressing the learning of model parameters based on a set of training sequences is presented in Section 5.1.2.5.

5.1.2.1 Forward-Backward Algorithm

The forward-backward algorithm is based on dynamic programming. It computes the forward variables $\alpha_t(i)$ and backward variables $\beta_t(i)$ that can also be used for learning within HMMs. For the observation sequence $\mathbb{O} = O_1, \ldots, O_T$, they are defined as

$$\begin{aligned} \alpha_t(i) &= P(O_1, \ldots, O_t, X_t = x_i|\lambda) && (5.10)\\ \beta_t(i) &= P(O_t, \ldots, O_T|X_t = x_i, \lambda) \ . && (5.11) \end{aligned}$$

They are both computed iteratively. For instance, the initialization and the recursive formula to compute forward variables are

$$\begin{aligned} \alpha_1(i) &= p(O_1|X_1 = x_i) p(X_1 = x_i) && (5.12)\\ \alpha_t(i) &= p(O_t|X_t = x_i) \sum_j \alpha_{t-1}(i) p(X_t = x_i|X_{t-1} = x_j) \ . && (5.13) \end{aligned}$$

The probability of an observation sequence is obtained by marginalization over the last visited state:

$$P(\mathbb{O}|\lambda) = \sum_i \alpha_T(i) \ . \tag{5.14}$$

As can be seen from Equation 5.10, the forward variables also permit inference in the model, namely the computation of $P(X_t|O_1, \ldots, O_t, \lambda)$ for $t \leq T$.

5.1.2.2 Viterbi Path

The Viterbi path is defined as a state sequence $\mathbf{x} = x_{k_1}, \ldots, x_{k_T}$ that maximizes the probability of having produced a given observation sequence $\mathbb{O} = O_1, \ldots, O_T$:

$$\underset{\mathbf{x}}{\operatorname{argmax}} P(\mathbf{x}|\mathbb{O}, \lambda) \quad . \tag{5.15}$$

This is equivalent to the maximization of $P(\mathbf{x}, \mathbb{O}|\lambda)$, obtained in a way similar to the computation of the forward variables. Noting

$$\delta_t(i) = \max_{x_{k_1}, \ldots, x_{k_{t-1}}} P(\mathbb{O}, X_1 = x_{k_1}, \ldots, X_{t-1} = x_{k_{t-1}}, X_t = x_i|\lambda) \quad , \tag{5.16}$$

the best path in terms of Equation 5.15 is obtained by backtracking the maxima in their iterative computation:

$$\begin{aligned} \delta_1(j) &= \pi_j P(O_1|x_j) & (5.17) \\ \delta_{t+1}(j) &= [\max_i \delta_t(i) a_{i,j}] P(O_{t+1}|x_j) \quad . & (5.18) \end{aligned}$$

This results in a synchronization between the sequence $\mathbb{O}$ and the model. The Viterbi path is really a path in the sense that it does not include any zero transition probability. It can be computed in $O(N^2T)$. At each time step t, the state X_t obtained by computation of the Viterbi path may differ from the individually most probable state x maximizing $P(X_t = x|\mathbb{O}, \lambda)$.

This individually most probable state is computed using the variables γ_t:

$$\begin{aligned} \gamma_t(i) &= P(X_t = x_i|\mathbb{O}, \lambda) & (5.19) \\ &= \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} & (5.20) \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)} \quad . & (5.21) \end{aligned}$$

5.1.2.3 Baum-Welch Algorithm

The Baum-Welch algorithm is a formulation of the expectation maximization algorithm to optimize parameters of an HMM. Given a set of l observation sequences $V = \{\mathbb{O}^1, \ldots, \mathbb{O}^l\}$ and an initial set of HMM parameters λ, it updates the parameters in order to optimize the probability $P(V|\lambda)$ that the sequences are generated by λ:

$$P(V|\lambda) = \prod_i P(\mathbb{O}^i|\lambda) \quad . \tag{5.22}$$

The algorithm involves two steps:

1. E-step: computation of expected sufficient statistics
2. M-step: reestimation of parameters

During the E-step, expected sufficient statistics (ESS) are computed based on actual model parameters and observation sequences. This corresponds to the computation of the number of expected transitions between each pair of states and to the number of expected observation of each symbol within each state. During the M-step, the HMM parameters are reestimated from the expected sufficient statistics.

The ESS can be efficiently computed from the forward and backward variables. Note however that although the EM algorithm guarantees convergence to a local maximum, the behavior of the optimization highly depends on the initialization parameters, as the function to optimize is usually strongly non-convex.

5.1.2.4 Continuous Observation Distributions

If the observed data is continuous, the discrete HMM presented above would require discretization of the input space into a finite set of observation symbols. Continuous input data can however be used directly by changing the modeling of the observation probabilities. The most common way is to model the observation distributions with mixtures of gaussians:

$$P(O|X = x_i) = \sum_k c_{ik} N(O|\mu_{ik}, \sigma_{ik}) \ , \tag{5.23}$$

where c_{ik} are the mixture coefficient and $N(\cdot|\mu, \sigma)$ is a normal distribution. Changing the distribution does not affect the Forward-Backward algorithm. In the Baum-Welch algorithm, formulas need to be adapted to account for the different parameters.

Different variants have been considered in the literature, for instance by considering gaussians tied between states of the HMM. Worth to mention for complex input-spaces are examplar-based HMMs [Elgammal et al., 2003], where a modeling distribution centered directly on data values in non-euclidean spaces is proposed.

5.1.2.5 Parameter Learning

As mentioned in section 5.1.2.3, the success of the expectation maximization algorithm highly depends on the initialization. When a large set of training sequences is available, model selection can be performed by randomly initializing the model with a fully connected graph in which transitions between all nodes are allowed, performing EM training with part of the sequences, and then testing the model on the remaining ones. After doing this multiple times, the best model is finally kept. This approach is however time-consuming and random. Many heuristics have been proposed to search the parameter space of HMMs, for instance by merging states [Stolcke and Omohundro, 1994], splitting states [Singer and Ostendorf, 1996], combining state merging and state splitting [Xie et al., 2003], using genetic evolution [Won et al., 2006], or optimizing model scores such as the Bayesian Information Criterion (BIC) [Friedman et al., 1998] or entropy [Brand and Kettnaker, 2000].

Other approaches make guesses or assumptions on the underlying structure of the process and train the model in the respective subspace. For instance [Ghahramani et al., 1997] use factorial HMMs to model independent parallel processes

sharing the same observations and [Brand et al., 1997] propose coupled-HMMs to couple interacting processes. These approaches are presented as extensions to HMMs, since they often use a modeling containing additional random variables. Using the joint space modeling the product of all random variables, they can be modeled as traditional HMMs. It is however more convenient to use a factored state space in which all random variables are identified, as this permits a better presentation and also the design of dedicated learning and inference algorithms. The algorithms can be faster and require less memory space. The dependencies between multiple random variables are conveniently represented in the framework of DBNs, which is introduced in the next section.

5.2 Dynamic Bayesian Networks

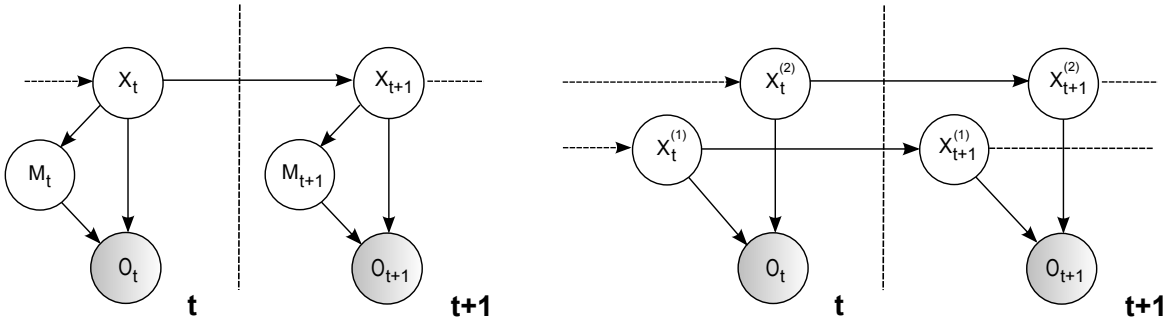
A dynamic Bayesian network [Murphy, 2002] is a type of Graphical Model [Stephenson, 2000] that models probabilistic dependencies over time between a set of random variables used to describe a process. Its variables satisfy the Markov properties. Let $X_{1:T}^{1:N} = \{X_t^1, \ldots, X_t^N \ : \ 1 \leq t \leq T\}$ be a set of random variables, where T is a time boundary and N the number of nodes used to model the process. Some of these nodes can be observed, the others are hidden.

To specify a DBN, initial probability distributions $P(X_1^j)$ for all nodes have to be specified, as well as intra-time-slice probabilities $P(X_t^j|X_t^i)$ and inter-time-slice probabilities $P(X_{t+1}^j|X_t^i)$. The joint distribution up to a bounding time T is then obtained by

$$P(X_{1:T}^{1:N}) = \prod_{i=1}^{N} P(X_1^i|X_1^{1:N}) \prod_{t=2}^{T} \prod_{i=1}^{N} P(X_t^i|X_t^{1:N}) P(X_t^i|X_{t-1}^{1:N}) \ . \tag{5.24}$$

In the graphical representation, nodes represent random variables and arc dependencies. A representation of an HMM as a DBN is given in Figure 5.1(b). At each time step, the observation only depends on the current hidden state, corresponding to the observation distribution B in Section 5.1.2. Due to the Markov assumption, only dependencies between two successive time slices need to be represented. Between two slices, there is a dependency only between the hidden nodes, modeled in Section 5.1.2 by the matrix A. For comparison, an HMM with an observation distribution modeled by a mixture of gaussians is given in Figure 5.2(a). The hidden (or latent) node M_t is used to represent the mixture coefficients (c_{ik} in Section 5.1.2.4). The graphical representation of a factorial HMM, in which the observations depend on independent processes, is given in Figure 5.2(b).

Similar problems as for HMMs are addressed in DBNs: fast estimation of the probability of a sequence of observations, learning and inference. Using the product space (also called joint space) of the random variables, DBNs can be "flattened" and modeled as HMMs. But an advantage of working within the factorized state space of DBNs is the improved modularity and interpretability of the model. The representation permits easily enforcing a-priori constraints in the learning and optimization algorithms. It can be used to design faster algorithms. It also provides a unique framework to incorporate existing HMM variants. DBNs generalize Bayesian networks by modeling the temporal



(a) Graphical representation of a hidden Markov model with an observation distribution modeled by a mixture of gaussians.

(b) Graphical representation of a factorial hidden Markov model with two layers.

Figure 5.2: Example graphical representations of hidden Markov models as dynamic Bayesian networks. Shaded nodes indicate observed nodes. Unshaded nodes indicate hidden (or latent) nodes.

dependencies. An interesting description of the extension of Bayesian networks to DBNs is available in [Mihajlovic and Petkovic, 2001].

5.3 Conclusion

In this chapter, we have presented the hidden Markov models and several related fundamental algorithms. We have also described the dynamic Bayesian network framework, in which HMMs and their extensions are conveniently represented. HMMs will be used in the next chapters for the modeling of workflow processes. The probabilistic framework yields a natural formulation of the on-line phase recognition problem. Additionally, the possibility to define arbitrary topologies permits us to address non-sequential workflows (Chapter 7).

CHAPTER 6

Monitoring Endoscopic Surgeries

Context-awareness requires on-line identification of the current surgical events, which is not provided by the framework described in chapter 4. In this chapter, we introduce an approach based on hidden Markov models that permits the on-line recognition of the surgical phases during endoscopy. Similarly to chapter 4, where we augmented a virtual surgery representation with an annotation indicating phase probabilities, we introduce phase probability variables that can be seen as a latent node within the HMM architecture represented as a DBN. The approach, named annotated workflow-HMM (AWHMM), is presented in Section 6.2. Three different methods are proposed for its construction, which differ in their use of the a-priori knowledge about the phases. Endoscopic videos can also potentially provide interesting additional information about the activities. Some experiments incorporating visual signals obtained from these videos are given in Section 6.5. Finally, section 6.6.3 shows that the HMM topology construction can provide interesting information about the underlying workflow.

6.1 Objectives

HMMs offer a much more general statistical description of a process than the virtual surgery representation, especially by allowing non-sequential modeling and by explicitly including probabilities, as needed for on-line recognition. A detailed relationship between HMM and DTW can be found in [Juang, 1984]. Traditionally, for instance in speech recognition [Rabiner, 1989], HMMs are used to model the stochastic properties of a training set of time series. The likelihood $P(\mathbb{O}|\lambda)$ that the time series $\mathbb{O}$ has been generated by the HMM λ can for instance be used for classification, if different HMMs are used to represent the classes. Most approaches using variants of HMMs for human action recognition address *isolated* actions and follow this methodology [Weinland, 2008]. In contrary, we model the complete workflow in a single model, so as to take the specific constraints of the workflow into account.

Several works share our initiative of imposing constraints on the flow of events to design a better model and improve recognition results, for example [Moore and Essa, 2002, Vu et al., 2003, Shi et al., 2006, Xiang and Gong, 2008]. Their main focus is however still on detecting single instances of actions. Our focus is on recognizing phases containing multiple actions within a lengthy, high-level workflow. In other words, our application and therefore modeling differs in that we consider either a higher semantical level (the phase) or an environment with more constrained temporal repetition across the sequences due to the surgical protocols.

In this chapter, 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 the phase $p \in \mathcal{L}$ carried out by the surgeon on-line. Based on training surgeries $\mathbb{O}^1, \ldots, \mathbb{O}^l$, we generate a model λ that allows the computation of the probability $P(\text{phase} = p \mid \mathbb{O}_{1:t})$ for each time-step t of another surgery $\mathbb{O}$ which does not belong to the training set. This gives us, as in Chapter 4, an annotation function

$$\mathcal{P}_{\mathbb{O}} : \{1, \ldots, T\} \rightarrow \mathcal{L} \ . \tag{6.1}$$

6.2 Annotated Workflow-HMMs

We define annotated workflow-HMMs (AWHMMs) as a form of HMMs extended by phase probability variables that keep track of the semantic (phases) inside the model. Additionally, these variables are convenient for coping with semantic loss during EM training and for training the model when only partially labeled data is available.

Formally, the AWHMM is a sextuplet $\lambda = (N, \mathcal{O}, A, B, \pi, \phi)$ where N is the number of states $\{x_i : 1 \leq i \leq N\}$ in the model, A the transition probability matrix between the states, modeling the topology, and $\mathcal{O}$ the space of observations. B is the observation model, indicating for any observation $O \in \mathcal{O}$ and state x the probability $B_x(O) = P(O \mid x)$ that O can be observed by x. π is a probability distribution over the initial states.
ϕ is an extension that denotes the *phase probability variables*, indicating the probability $\phi_x(p) = P(p \mid x)$ of *each* state x to belong to a phase $p \in \mathcal{L}$. These variables can be seen as the introduction of a latent node within the generic representation of an HMM as a dynamic Bayesian network (see Figure 6.1).

The results of segmentation and phase recognition largely depend on the approach chosen to construct and initialize the parameters of the AWHMM. There exist different heuristics to initialize and train hidden Markov models. Our choices have been determined by the fact that in our application, only a small set of training surgeries is available. We present in the following section three generic HMM initialization methods used in the AWHMM construction. We then present in section 6.2.2 several approaches for the construction of the AWHMM, based on these generic methods for comparison. The presented approaches differ in the required amount of labeled data and in the way the labels are used.

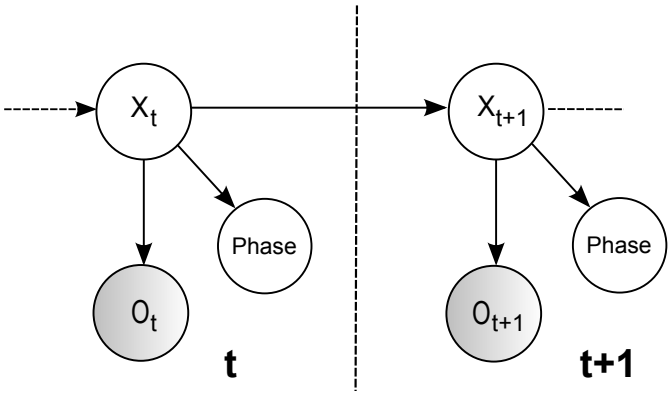


Figure 6.1: Dynamic Bayesian network representation of the generic AWHMMs with addition of a latent node.

6.2.1 HMM Initialization Methods

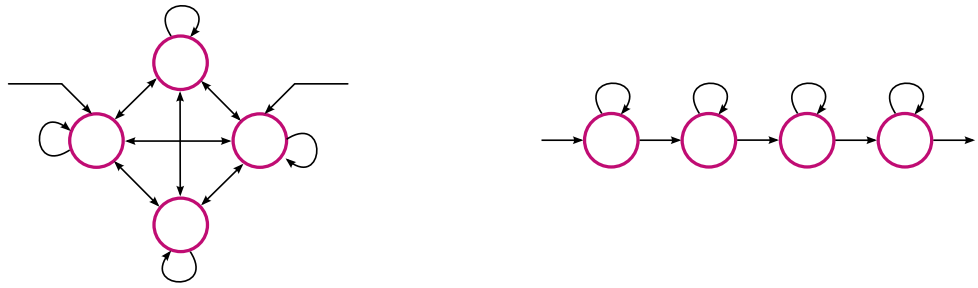
Given a training set of l time-series $\mathbb{O}^1, \ldots, \mathbb{O}^l$, there exist different methods to initialize and train an HMM λ such that the log-likelihood $\sum_i logP(\mathbb{O}^i|\lambda)$ is maximized. The AWHMM constructions described in the Section 6.2.2 are based, for comparison, on the three generic initialization methods described in the next sections. The different resulting topologies of these models are illustrated in Figures 6.2 and 6.3.

6.2.1.1 Fully-connected HMMs

In this initialization, a fixed number of states is chosen, either as a parameter or based on the amount of data. The topology is fully connected, which means that no transition probability is enforced to be zero (see Figure 6.2(a)). All parameters are then randomly initialized and the model is further trained by expectation-maximization (EM). Multiple initializations are performed and each model is evaluated on its ability to represent the training data. The best model is ultimately chosen. This is a common approach usually requiring a large amount of training data [Rabiner, 1989]. It will be refered to as *(HMM-full)* in the experiment sections.

6.2.1.2 Sequential HMMs

Here, the topology is enforced to be sequential with left-right transitions (see Figure 6.2(b)). The number of sequentially connected states is given as a parameter or inferred from the amount of training data. The training sequences are then split regularly into as many segments as the number of states and assigned to the states. The transition probabilities are chosen such that the expected duration time of the model corresponds to the average lengths of the training sequences. The observation distribution is initialized from the corresponding data. EM is finally performed on the model to refine the parameters. This is a natural construction approach, as the sequential topology permits a good initialization of the parameters from the data. In our experiments, the number of states is chosen to be $\sqrt{(\frac{\overline{d}}{2})}$ where $\overline{d}$ is the average length of the training sequences.



(a) Fully connected topology. (b) Left-right sequential topology.

Figure 6.2: Two usual HMM topologies.

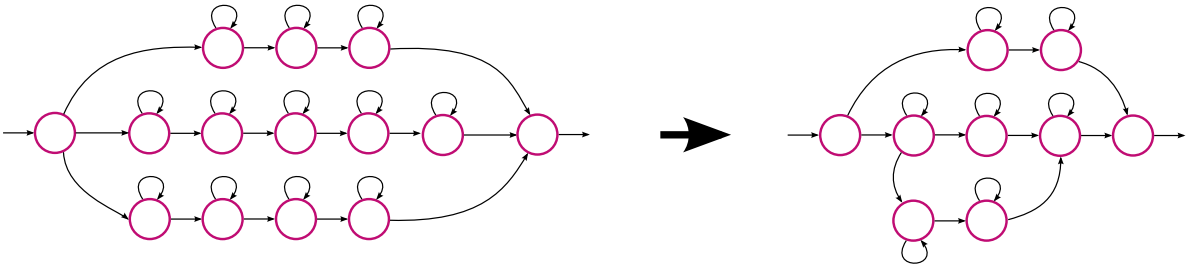


Figure 6.3: Model merging is performed on an exhaustive representation of the training sequences (here three are displayed) by merging pairs of states and updating parameters accordingly.

This is a classical rule of thumb that performs better than choosing any constant, but other classical rules for model selection, like $(\frac{\overline{d}}{log(\overline{d})})^{\frac{1}{3}}$, perform equivalently well. Variants in this model include additional left-right transitions skipping one state. This approach will be referred to as *(HMM-seq)* in the experiment sections

6.2.1.3 Model Merging

This method is described in [Stolcke and Omohundro, 1994]. Initially, sequential left-right topologies are built out of each training sequence, by creating one state for each observation. Only one state is created for consecutive identical observations. These topologies are connected with initial and end states as shown in Figure 6.3 to generate an exhaustive and overfitted model. Then, pairs of states are chosen and merged together to optimize the probability that the model generates the training sequences. The process is stopped when the decrease in log-likelihood exceeds a given threshold. Additional details on the model merging approach can be found in [Blum, 2007]. An advantage of the approach is its direct initialization from the data, which can reveal underlying structures, as described in section 6.6.3. The approach is however time consuming and the data has to be split at meaningful points if sequences are too long. This approach will be referred to as *(HMM-merged)* in the experiment sections.

6.2.2 AWHMMs Construction Approaches

The following sections describe construction methods to automatically build a AWHMM that describes the stochastics of the surgical signals and permits inferring the current phase using labeled training data.

6.2.2.1 Manual Model Annotation

In this case, the training set is not supposed to be labeled. An HMM is constructed in an unsupervised fashion from the data and the phase probability variables are manually defined after its construction. While this is possible for a virtual surgery representation (see section 4.3.1.1), as it resembles a surgery, this is more complicated for HMMs, where the meaning of a state is difficult to infer from its observation distribution. This would however be feasible for models constructed using a left-right topology or model merging, when the signals contain enough semantic to allow for a consistent visualization of the meaning of the states (see section 6.6.3). Another major drawback of this approach is that the model needs to be re-annotated or updated manually each time a new surgery is added to the training set.

6.2.2.2 Pre-annotation

This is a fully supervised framework. It supposes all sequences of the training set to be labeled. In this case, an HMM λ_p for each consecutive phase p is constructed and all these models are concatenated to form an overall model λ as presented in figure 6.4. Artificial starting and ending states are used in each sub-model to facilitate the concatenation. Using one of the three methods presented in section 6.2.1, each sub-model is initialized from the training data of the corresponding phase. For each state x of the overall model λ, the phase probability variables assign the probability 1 to the phase p, given that the state x originally stemmed from the sub-model λ_p before concatenation.

As explained in Section 6.2.2.4, if the EM algorithm is used after concatenation, the annotation ϕ is updated accordingly, leading to phase probability variables that do not take only binary values anymore. Practically, this HMM is a hierarchical HMM with 2 hierarchy levels.

6.2.2.3 Post-annotation

In this case, the model is first directly constructed from all data without using any label information. Afterwards, the phase probability variables are automatically computed using the training sequences that are labeled.

An approach using the best path in terms of the Viterbi algorithm is presented below. A more general approach using all paths to visit more states is used in the next section for completeness, but both approaches yield similar results. To compute the phase probabilities, the labeled time-series $\mathbb{O}^1, \ldots, \mathbb{O}^l$ of length $T^1, \ldots, T^l$ are synchronized to the model using the Viterbi algorithm (see Section 5.1.2.2), giving for $i \in \{1, \ldots, l\}$ paths

$$\text{path}_i : \{1, \ldots, T^i\} \rightarrow \{(x_k)_{1 \leq k \leq N}\} \quad . \tag{6.2}$$

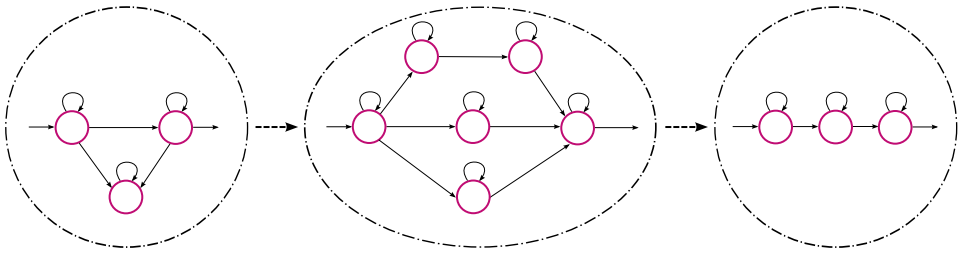


Figure 6.4: Pre-annotation model construction: Sub-HMMs are constructed for each phase or cluster of data and appended as displayed in this image. Start and end states used to facilitate the concatenation are not displayed in this figure.

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

$$\phi_x(p) \propto \frac{\#\{i,t:\ path_i(t)=x \ \ and \ \ p=\mathcal{P}_{\mathbb{O}^i}(t)\}}{\#\{i,t:\ path_i(t)=x\}} \quad . \tag{6.3}$$

where $\#\cdot$ is the cardinality operator. If the state x was visited, ϕ_x is then normalized to one. If not, all probabilities are kept to zero.

When several labeled surgeries are used, the annotation is averaged. When the model is not sequential, there can be multiple, parallel paths leading from the starting state to the terminal state, as can be seen in the graphical representations of Figures 6.2 and 6.3. In this case, it is possible that some states or entire state branches are not visited during the process and thus do not receive any annotation. Consequently, the number of states in the topology has to be kept reasonably small with respect to the amount of training data to avoid graphs containing numerous unannotated states. Alternatively, more labeled data has to be provided to achieve an annotation of these states. If this is not possible or unwanted, we choose to assign an average annotation for an un-labeled state by using all neighboring states with annotation, weighted according to the transition probabilities between the states.

6.2.2.4 Training and Phase Probabilities

It is usual to refine HMM models with the expectation maximization algorithm. When applying global EM, however, observations can shift to neighboring states. In case of pre-annotation, hidden states might no longer exclusively represent single phases. Phase probabilities are hence used to track this shift. In the post-annotation approach, phase probabilities also need to be recomputed.

Different proposals have been made to deal with such a shift in the semantic meaning of the hidden states. For instance [Shi et al., 2006] proposed, for learning their *propagation networks*, the use of a labeled anchor sequence during EM to prevent alteration of semantic information. In their work, sub-actions are represented only through single nodes and the model topology is provided manually. When the topology is derived from the data and the modeling involves longer and more complex sub-phases represented by several nodes,

constraining the EM algorithm with a labeled sequence requires a complex labeling and is less convenient. Instead, we propose a general approach that keeps track of the phases precisely by using the phase probability variables. This provides moreover a convenient formulation during on-line recognition in the AWHMM (see 6.3.2) and is potentially more powerful by enabling the addition of new labels without performing another EM training.

In this approach, ϕ_x is computed a-posteriori using a set of labeled sequences $\mathbb{O}^1, \ldots, \mathbb{O}^l$, in a way similar to a single update of model parameters during EM [Rabiner, 1989]. The approach is refered to as "all-paths", as opposed to the best path approach of Section 6.2.2.3. The variables $\gamma_t^i(x) = P(X_t = x|\mathbb{O}^i)$ are used (see Section 5.1.2.2), indicating the probability to be in state x at time t while knowing the sequence $\mathbb{O}^i$. We count the number of visits that a label makes to each state:

$$\phi_x^i(p) = \frac{\sum_{\{t,phase(\mathbb{O}_t^i)=p\}} \gamma_t^i(x)}{\sum_t \gamma_t^i(x)} \quad . \tag{6.4}$$

Finally, ϕ_x is obtained by summing and normalizing over all available sequences. ϕ can be quickly recomputed without having to perform EM, for instance if more previous sequences are labeled.

This approach is more general than the one of Equation 6.3, by considering all possible paths. In particular, this reduces the likelihood that unannotated nodes occur, as more states get annotated for each labeled sequence. It however does not garantee that all nodes will get an annotation, since this depends on the available annotated sequences. On our data, the results were similar with both approaches, all paths and best path, both for cholecystectomies and the 4D application presented in the next chapter.

In the cholecystectomy application, the phase probabilities are especially interesting for the modeling when only a subset of the training data is labeled. The post-annotation method of section 6.2.2.3 can be used in such case. For the pre-annotation, recomputing the phase probabilities after global EM shows little change in the results for this application. This is due to the linear overall workflow and discrete signals. In the next chapter, where continuous signals are used in a workflow with alternative paths, we present experiments where the usefulness of recomputing the phase probability variables after global EM is apparent.

6.2.3 Observation Distributions

The last parameters that need to be determined in the AWHMMs are the observation distributions for the discrete data. We used two distributions for the discrete signals. The first distribution corresponds to the observation vector frequencies. It is computed by counting the observation occurrences in the data. The second distribution assumes instrument independence and counts the usage frequencies of each instrument. The resulting probability is the product of the probabilities of each instrument. In both cases, a constant small probability is used to account for unobserved data. Since the second distribution performed slightly better in the experiments, this was the distribution we used for the evaluation.

6.2.4 Model Construction Speed-up

Training an HMM model from full data or from data of a long phase can be very computationally expensive, especially when using the model merging method. To alleviate this burden, we split the training data and concatenate the trained sub-models in an unsupervised manner, similarly to the pre-annotation approach (see Section 6.2.2.2). In this approach, the transitions between phases are used as synchronization points. We propose to use the virtual surgery representation to find such consistent splitting points and to construct the full model from subparts. This can be applied to sequential workflows or to phases. Using the surgical similarity SIM_t (see Section 4.2.3), we split at points which correspond to local similarity maxima, where the maximum allowed size of a split is bounded by a single parameter. This provides clusters of training data by back-projecting from the virtual representation to the training time-series, as described in fig. 6.5. HMM sub-models are trained from these clusters of training data and finally concatenated.

6.2.5 Discussion

Within the three construction methods, the pre-annotation construction requires all phases of all training surgeries to be labeled. In contrast, the post-annotation provides the advantage that not all 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 detecting only some subparts of the surgeries is interesting, it is also straightforward to construct the model using only data where these subparts are annotated. The effectiveness of the post-annotation approach depends on the ability of the model constructed with no supervision to adequately model the time-series. For the sequential workflow of cholecystectomy, as will be shown in the evaluation, the approach yields good results. For the 4D application presented in the next chapter, where the workflow contains alternative courses of actions and the observations are continuous, we required a set labeled data. Finally, addition of new surgeries requires the recomputation of the models in all approaches. For post-annotation, there is the possibility to use the labeling information for recomputation of the annotation only.

6.3 Off-line Segmentation and On-line Recognition

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

6.3.1 Off-line Segmentation

Off-line segmentation is the process of segmenting a new surgery after the acquisition of all signals $\mathbb{O}_{t,k}^{test}$. The objective is to compute the phase $\mathcal{P}_{\mathbb{O}^{test}}(t)$ at each time step t

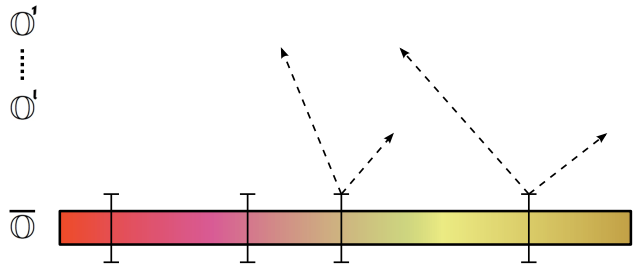


Figure 6.5: Training data obtained from surgical similarity splits. Each subpart of the virtual surgery $\overline{\mathbb{O}}$ corresponds to synchronized subparts of the training surgeries, which are then used to construct the sub-HMM.

while knowing the *complete* signals $\mathbb{O}_1^{test}, \ldots, \mathbb{O}_{T^{test}}^{test}$, where T^{test} denotes the end of the surgery.

If an AWHMM λ 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:

$$\text{path} : \{1, \ldots, T^{test}\} \rightarrow \{(x_i)_{1 \leq i \leq N}\} \ . \tag{6.5}$$

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

$$\mathcal{P}_{\mathbb{O}^{test}}(t) = \underset{p}{\text{argmax}}\, \phi_{\text{path}(t)}(p) \ . \tag{6.6}$$

6.3.2 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.

In the AWHMM λ, the forward probabilities (see Section 5.1.2.1) permit to compute $P(X_t = x_i \mid \mathbb{O}_1^{test}, \ldots, \mathbb{O}_t^{test})$, 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:

$$\mathcal{P}_{\mathbb{O}^{test}}(t) = \underset{p}{\text{argmax}}\, P(\text{phase} = p \mid \mathbb{O}_1^{test}, \ldots, \mathbb{O}_t^{test}) \ . \tag{6.7}$$

From the annotation of the model, we know the probability

$$P(p \mid X = x) =_{def} \phi_x(p) \tag{6.8}$$

of being in phase p while being in the HMM state x, thus

$$\mathcal{P}_{\mathbb{O}^{test}}(t) = \underset{p}{\operatorname{argmax}} \sum_{x_i} P(p \quad | \quad X_t = x_i) \times P(X_t = x_i \quad | \quad \mathbb{O}_1^{test}, \ldots, \mathbb{O}_t^{test}) \tag{6.9}$$

$$= \underset{p}{\operatorname{argmax}} \sum_{x_i} \phi_{x_i}(p) P(X_t = x_i \quad | \quad \mathbb{O}_1^{test}, \ldots, \mathbb{O}_t^{test}) \ . \tag{6.10}$$

The model being given, this on-line computation has a constant complexity per time-step.

6.4 Evaluation

We have evaluated the presented methods on data of 16 cholecystectomies performed by 4 different surgeons, using the evaluation measures introduced in Section 4.3.3.1 (accuracy, recall and precision). Results for off-line segmentation are shown in Table 6.1, results for on-line recognition in Table 6.2. The three different names AWHMM-full, AWHMM-seq and AWHMM-merged refer to the three different generic initialization methods for AWHMMs presented in Section 6.2.1. We performed a leave-one-out cross-validation, where for each of the 16 surgeries the model was built from the remaining 15 ones. The results display the mean values and standard deviations over all surgeries.

Since all information is available, results for off-line segmentation are better than for on-line recognition when the same methods are compared. For off-line segmentation, methods based on DTW have measures above 94%. The AWHMM-seq method is only slightly worse. In these off-line experiments, we use the same AWHMMs as for the on-line results, for fair comparison. HMM topologies are designed general enough for on-line recognition and adaptation to rare observations.

As can be expected, results with post-annotation are worse than with pre-annotation, as less knowledge is used. They are however good, as recall and precision above 90% can be achieved for HMMs. AWHMM-full is based on random initialization. Multiple initializations followed by EM 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 online-recognition, the best results are above 90% and obtained with AWHMM-seq. Post-annotation slightly decreases the values for this method, but the decrease is below 3 percentage points. This is very interesting because when only half of the data is labeled, the errors remain similar for all methods, as shown in Figure 6.6.

While AWHMM-merged performed worse than the other methods, this method has still an important advantage. As explained in Section 6.6.3, using the inherent semantic of the tool usage signals, the topology of the resulting HMM can be interpreted. It can be used to automatically generate visualizations and human understandable statistical models of surgical workflow.

In Table 6.3, on-line results per phase are displayed for AWHMM-seq with pre-annotation construction. Results are usually above 90 %, except for short phases with a

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

Table 6.1: Off-line results. Leave-one-out cross-validation on 16 surgeries performed by 4 surgeons. Mean and standard deviation over all surgeries. *(pre)* indicates pre-annotation construction, *(post)* construction with post-annotation. AVSR refers to the approach in Chapter 4.

relatively high standard deviation compared to the mean duration, as it is the case for phase 4, 5 and 11. For these phases, there are one or two surgeries in which the phases were not recognized.

The results are very promising, since usually all phases are recognized. The errors are mainly caused by a delay of some seconds when detecting a new phase, which is acceptable for most applications. The phases that are completely skipped are mainly phases that are very short and that are not recognized due to the delay. For instance, the external cleaning phase (phase 11) has an average length of 1 minute and 44 seconds. In a few surgeries, its duration is below 15 seconds.
In Fig. 6.7, the average length of each phase and the corresponding mean error per phase (duration of uncorrect detections) are indicated in minutes. These errors were obtained for AWHMM-seq with pre-annotation construction. The maximum mean error per phase occurs for phase 3 and is of 1 minute and 14 seconds. For comparison, AWHMM-seq with post-annotation yields 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.

The fact that these results were obtained using only little 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. With a large enough database of cases, it would be possible to generate surgeon-specific models of the surgery. This, in return, can be expected to further increase recognition accuracy.

6.5 Use of Visual Signals

Real surgeries show little repetition in the endoscopic video within such a small amount of data. This is due to patient and surgeon specificities as well as camera movements. Additionally, the images are very challenging for vision algorithms as many perturbations occur. These include strong specularities, appearance of smoke, tissue deformations,

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

Table 6.2: On-line results. Leave-one-out cross-validation on 16 surgeries performed by 4 surgeons. Mean and standard deviation over all surgeries. *(pre)* indicates pre-annotation construction, *(post)* construction with post-annotation.

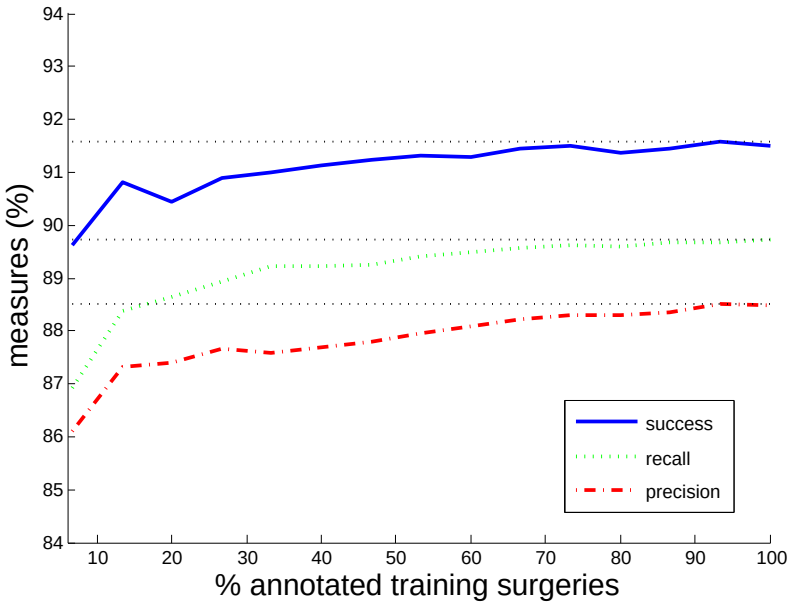


Figure 6.6: Accuracy (rate of success), average recall and average precision for AWHMM-seq with post-annotation evaluated on-line. Influence of number of annotated surgeries. The horizontal lines refer to the best result for each curve, obtained when all training surgeries are labeled.

	Phase Label	Rel. Len. (%)	Recall (%)	Precision (%)
1	CO2 Inflation	6.5 (±3.2)	100.0 (±0.0)	100.0 (±0.0)
2	Trocar Insertion	6.4 (±1.4)	98.3 (±6.7)	99.6 (±1.1)
3	Dissection Phase 1	17.1 (±9.1)	86.4 (±23.9)	98.1 (±7.5)
4	Clipping Cutting 1	3.8 (±1.6)	69.4 (±45.4)	66.4 (±40.6)
5	Dissection Phase 2	3.7 (±5.0)	68.3 (±36.9)	67.7 (±37.9)
6	Clipping Cutting 2	3.9 (±1.9)	91.5 (±15.7)	79.1 (±25.5)
7	Gallbladder Detaching	18.9 (±13.1)	94.8 (±5.9)	88.0 (±18.9)
8	Liver Bed Coagulation 1	7.1 (±4.4)	93.4 (±18.1)	90.0 (±13.6)
9	Gallbladder Packaging	4.1 (±2.8)	99.7 (±1.1)	96.8 (±6.0)
10	External Retraction	11.7 (±11.7)	99.7 (±0.9)	95.9 (±8.4)
11	External Cleaning	3.6 (±4.1)	82.3 (±31.9)	79.8 (±17.3)
12	Liver Bed Coagulation 2	6.2 (±2.1)	96.8 (±7.3)	99.6 (±1.0)
13	Trocar Retraction	2.6 (±1.9)	94.8 (±12.1)	93.9 (±13.7)
14	Abdominal Suturing	3.9 (±3.4)	98.1 (±5.5)	93.3 (±17.9)

Table 6.3: Online results. Leave-one-out cross-validation on 16 surgeries performed by 4 surgeons. Detailed results per phase for the pre-annotation construction, with mean and standard deviation over all surgeries. The third column (*Rel. Len.*) indicates the average relative length of each phase with standard deviation.

occlusions and fast change of field of view (see Figure 6.8). In order to use information from these images, robust visual signals need to be constructed.

In the following, we define two specific detectors whose outputs are computed directly from the endoscopic video and can be used as signals. The first detector indicates whether the camera is present in the body and the second whether clips are detected in the field of view (see Figure 6.9). Their usage shows improvement in the recognition rates.

6.5.1 Signals

Endoscopic camera signal As can be guessed from figures 3.3 on page 32, color is a main cue to obtain the state of the endoscopic camera. Since cameras with different settings are used, a color normalization [Paulus et al., 1998] is performed on the images beforehand, which proved to improve the results. We used a small color histogram of size 20 as visual feature. The first 10 bins contain a hue histogram and the last 10 bins a saturation histogram of the image. Based on labelled images from training surgeries, two Gaussian mixture models (GMMs) G_{endo} and G_{out} are trained to model the color spaces of endoscopic images and of outside images taken by the camera before or after extraction from the trocar. The training consists in an initialization with principal component analysis followed by EM iterations. An image is classified as endoscopic when the probability of its histogram to belong to G_{endo} is higher than for G_{out}. Evaluation based on a few manually labelled complete surgeries provides a success rate of 92%.

Two kinds of images are difficult in the evaluation: images where the camera is half

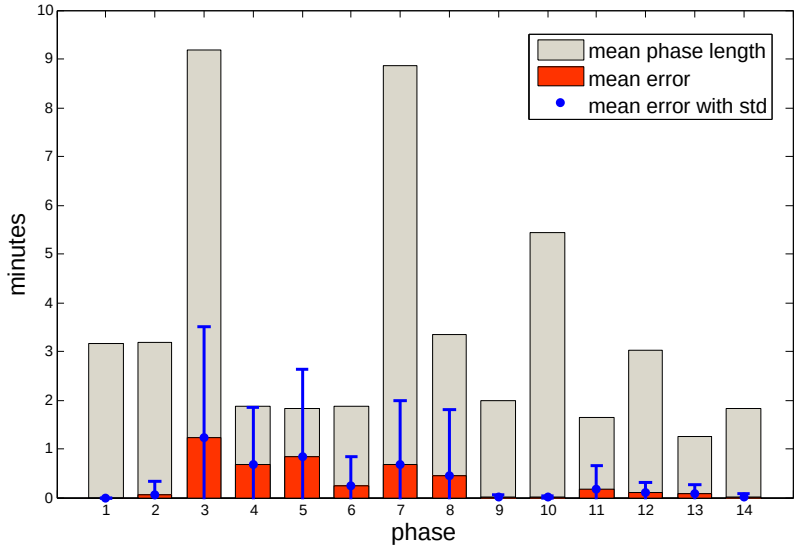


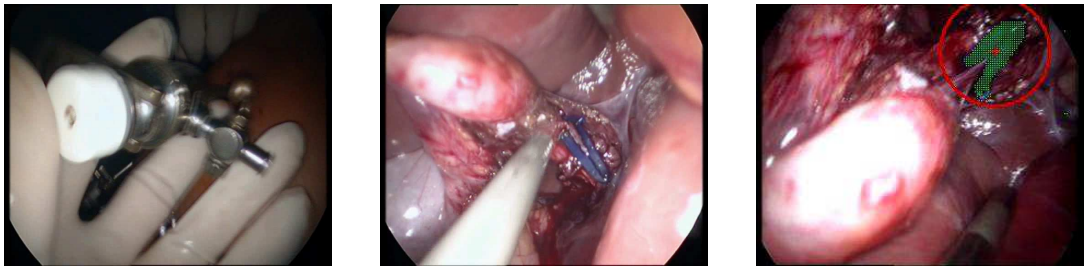
Figure 6.7: For each phase, average length over all surgeries overlaid with mean detection error, in minutes. Errors are computed on-line for AWHMM-seq with pre-annotation construction.

(a) Blur. (b) Smoke. (c) Far view. (d) Specularity.

Figure 6.8: Several challenging views taken by the endoscopic camera.

inserted and as much of the metallic trocar can be seen as of the internal anatomy, and images where the camera is entirely inside the abdomen but fully blinded by specularity due to its proximity to the tissue.

Clip detection signal Clips are small longitudinal metallic objects used to clamp the blood vessels. In most of the surgeries we recorded, dark blue clips were used, sometimes with additional grey titanium clips. To detect the blue metallic clips, we use color classification with a similar approach as before.Two GMMs G_{clips} and G_{bg} modeling the colors of the clips and of the endoscopic background are constructed based on a few training images segmented manually. They are used in a first step to classify the image pixels into clip and background. After morphological closing operations, the connected components are elected as clips depending on two properties: reasonable size and longitudinal shape. When the endoscopic camera is detected to be in the body and such a connected component is found, the clip signal is set as active. Note here, that false positives in the detection



(a) Endoscopic camera outside the body.

(b) View of a clip.

(c) Detected clip.

Figure 6.9: Exemplary pictures for the computation of visual signals.

are more an issue than false negatives. Indeed, detected clips before their introduction in the body give a wrong indication about the possible phase. It however often occurs that clips are small or hidden from the field of view after their introduction in the body. For this reason, the criteria are set such that clearly visible clips are detected, while far or partially hidden clips might not be detected. This reduces the number of false positives. False positives may occur as the dark blue clips have a color similar to coagulated tissues and to smoke. Clips are also sensitive to specularities and reflections. The recognition model appeared however to cope with these false detections. The evaluation is not easy, as manual labeling of the clips is tedious and has to be done image per image, contrary to the instruments for which we only consider the presence in the body, not in the field of view. On three manually labeled surgeries, where also partially and hardly visible clips were noted as present, the automatic detection rate is 76%.

6.5.2 Evaluation

For these experiments, we used 11 surgeries performed by three surgeons where blue clips were used. As before, we perform a cross-validation on the dataset. In the experiment without visual signals, all signals presented in Section 3.1.2.1 are used except the signal indicating the presence of the endoscopic camera. In the experiment with visual signals, the two signals computed from the endoscopic videos as defined in previous section are added. The signal indicating clip presence adds information to the existing signal indicating the presence of the clipping device. We use these two binary signals instead of the color histograms for dimensionality reduction, as a small training set of surgeries is available. On-line results after cross-validation with the pre-annotation approach are presented in table 6.4. They show in particular that addition of signals computed from the videos can improve the recognition results. As noted before, the two added signals are noisy, especially due to the ambiguities arising from the challenging images. However, these encouraging results show that the AWHMMs cope with noise that can be present within the input.

Other signals could potentially be derived from the videos. However, as can also be seen in the work of [Lo et al., 2003, James et al., 2007], detection rates are usually low, even when considering simpler cases such as pig surgeries or parts of videos only. For

	Accuracy (%)	Average Recall (%)	Average Precision (%)
With	91.2 (±7.5)	92.1(±7.9)	86.9 (±10.9)
Without	87.3 (±13.8)	88.8 (±11.9)	83.6 (±15.4)

Table 6.4: Online results for AWHMMs, with mean and standard deviation over the cross-validation tests. Comparison of evaluation measures using pre-annotation construction with and without the visual signals included during the experiments.

this reason, discriminative signals such as instrument usage seem to be more practical, especially if only small training sets are available.

6.6 Applications

Using methods described in the previous sections, we can develop several applications with a direct and potentially large impact on patient welfare, peri-operative hospital organization and surgeon's control over the surgical processes.

6.6.1 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 (see e.g. Figure 6.10). This can also be used to control a user-interface providing context-aware information. Calling the next patient is actually of clinical importance, since if done too soon, the next patient might stay anaesthetized for an unnecessarily long period of 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 (gallbladder detaching). This could therefore be done automatically and reliably, relieving the OR staff from this task. This phase was always recognized in the detection process; its mean recall is 94.8% and mean precision 88.0% in case of AWHMM-seq with pre-annotation. The recall corresponds to a recognition error in absolute time within the phase of 40 seconds, as can be seen in fig. 6.7.

6.6.2 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 planning of schedules. After detecting the current HMM state during a running surgery, the average remaining time can be easily inferred from the HMM transition probabilities. Results are shown in Figure 6.11. For this figure, prediction of remaining time was performed at each time steps and the errors averaged in each phase are presented.

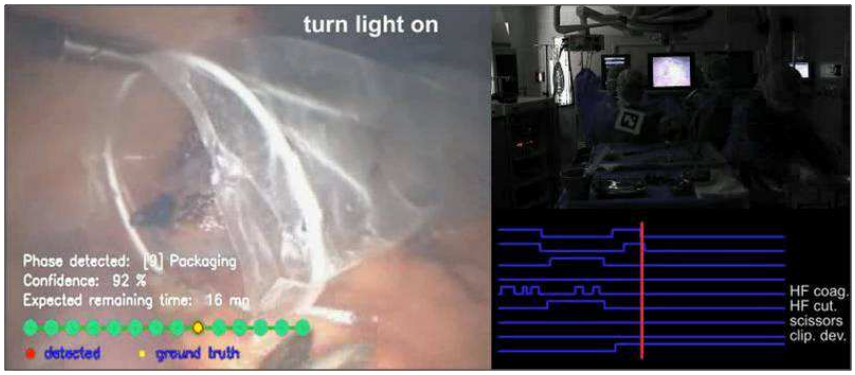


Figure 6.10: Excerpt from an illustrative monitoring video published in [Padoy et al., 2008], showing synchronized input data, recognition and event triggering (here a message asking to switch on the OR lights).

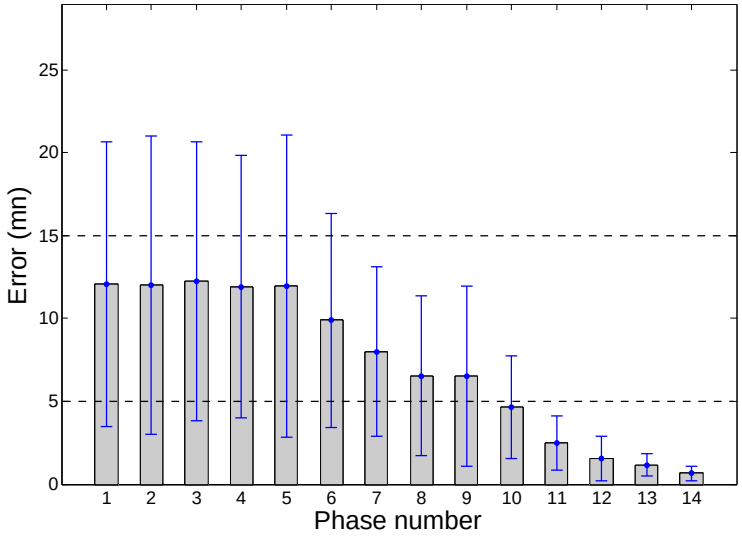
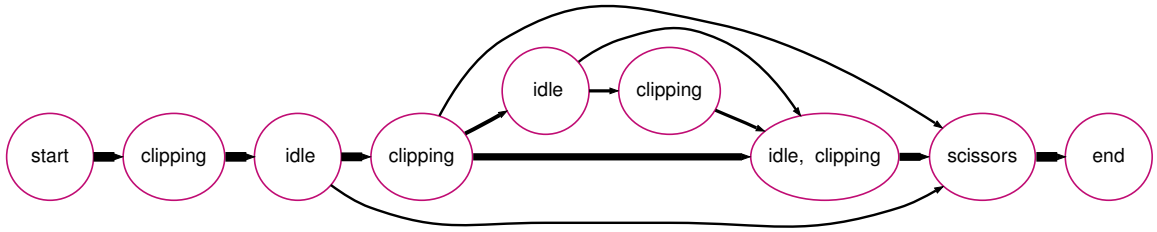


Figure 6.11: Average errors in remaining time prediction, per phase.

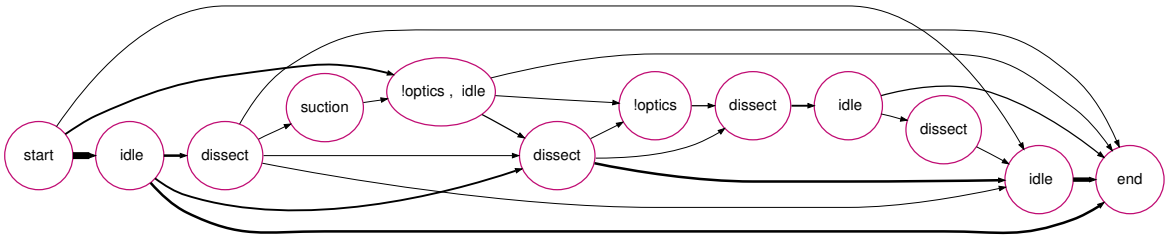
Prediction starts to be accurate with the beginning of phase 10, with a mean prediction error below 5 minutes. The average length among all surgeries is 48 minutes. A more accurate prediction could use information about the patient (e.g. size, weight, state of inflammation) and about the surgeons (e.g. skills for the kind of surgery). We unfortunately lack this data, but hopefully this data will become available in the future.

6.6.3 Towards Workflow Mining

An interesting property of the model merging approach introduced in section 6.2.1, is that the construction of the topology provides hints about the underlying workflow process. Models generated with the model merging approach for two phases are presented in figure 6.12. The method clusters the observations from the training data and permits to identify the different probable paths within the workflow. Widths of the arrows are



(a) Clipping and cutting phase 1.



(b) Dissection phase 2.

Figure 6.12: HMM topologies obtained with the model merging approach. Node labels indicate the main occuring actions and are derived from signal semantics.

proportional to these probabilities. In order to attach semantic meaning to the topology, for each node a label is derived from its observation distribution. The displayed label indicates the most likely activities, derived from the tool usage signals associated to the node by the training process. While the approach is promising, for instance for automatically generating a human-readable model of a complete surgery, several interesting issues need to be tackled for broader usage. First, computational complexity has to be reduced to run the algorithm on signals from complete surgeries, instead of signals from phases. Other structure learning approaches could be faster and also interesting, such as combination of model splitting [Siddiqi et al., 2007] with model merging, or entropy based methods [Brand and Kettnaker, 2000]. Second, continuous signals do not always have inherent semantics attached, such as tool presence. Semantic annotation of the model then requires the presence and incorporation of additional information, such as ontologies. Finally, usability of the results highly depends on the clustering, which may be misleading and may have to be controlled for a better understanding. For instance, in Figure 6.12(a) the different nodes indicating clipping show that the number of clips can vary. But we cannot infer from the clustering how many clips are actually used regularly by the surgeon, as the number of 'clipping' nodes depend on the number of merging steps performed by the process. A graphical interface permitting to expand/contract the model by decreasing/increasing the number of merging steps, as presented in [Blum et al., 2008b], can however ease the interpretation.

6.7 Conclusion

In this chapter, we have presented on-line phase detection approaches for endoscopic cholecystectomies. Results on this data suggest that the development of context-aware applications in the surgery room is possible, if signals providing the tool usage in real time become available in regular routine. Both the current choice of signals and the approach can be applied to other endoscopic surgeries. As in Chapter 4, results were presented on 16 surgeries performed by 4 different surgeons from the same medical school. It will be interesting to see how the results improve when more data is available. With sufficient exemplary training data, the models are expected to provide good results even on surgeries performed at different medical schools. An interesting testbed would be surgeries performed with the da Vinci robot, where signals can be automatically acquired in real time during real surgeries. Once the question of the signal acquisition is solved, other interesting applications would be open surgeries, where workflow plays the same important role. In the next chapter, we apply the methods to another challenging and complementary application, the recognition of the daily workflow in the operating room. In this application, the signals are continuous and the workflow can contain alternative paths. This makes it interesting and complementary to the work presented in this chapter.

CHAPTER 7

Monitoring Workflows using 4D Features

In this chapter, we address the recognition of generic phases occuring during the daily OR workflow, based on signals recorded from a multi-view reconstruction system. Two such systems have been installed in Munich, the first in a laboratory at Technische Universität München, the second within an intervention room at Hospital Grosshadern. A main objective of the system, which was supported by Siemens Healthcare, is the detection of potential collisions between a robotic device and the objects and staff present within the surgery room.

A second and very promising application of the system that we address in this thesis is the recognition within a surgical workflow. Recognition based solely on reconstruction data is also a recent field of research, where existing work has only addressed action classification performed by a single human. The purpose of this chapter is to propose an initial approach for workflow recognition based on reconstruction data, within a complex scene involving several persons and objects. Contrary to the cholecystectomy application, where the workflow is sequential, in this application the workflow contains alternative courses of actions. The setup of the system has been described in Section 3.2. The features derived from the reconstruction data are presented in Section 7.2. The initialisation of the annotated workflow-HMM adapted to these new features and to a workflow containing alternative paths is given in Section 7.3. Finally, experimental results are shown in Section 7.4.

7.1 Objectives

Using the multi-camera reconstruction system, we obtain a sequence of visual hulls $\mathbf{r}_{1:T}$ in an on-line manner, as explained in Section 3.2.2 on page 38.

Even though much work address the reconstruction problem in a multi-camera setup (we refer the reader to [Seitz et al., 2006] for a recent review), only few work address activity recognition using 3D reconstruction data. Existing work address the recognition

of actions performed by a single human, such as kicking or punching. The INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [Weinland et al., 2006] is usually used as reference in experiments. This dataset consists of synchronized acquisitions of 5 multi-view videos capturing actions performed by various actors.

In [Weinland et al., 2006], the motion template approach that had been introduced in [Bobick and Davis, 2001] for 2D action recognition was generalized to 3D. Motion history volumes together with a viewpoint independent representation were proposed to classify human actions. The method was directly applied to 3D reconstruction data. In [Lv and Nevatia, 2007, Weinland et al., 2007, Yan et al., 2008], actions were learned from 3D voxels for view-invariant action recognition in 2D images. For recognition, [Lv and Nevatia, 2007] rendered multiple 2D poses from 3D examples and performed silhouette matching. During matching, temporal smoothness constraints were enforced using a deterministic action graph. [Weinland et al., 2007] proposed a similar approach based on key poses, using HMMs to recognize actions. In [Yan et al., 2008], recognition was performed with spatio-temporal features, computed from 2D projections of the 3D examples over time.

As motivated in section 3.2.1.1, in this work, the purpose is both to learn from the 3D data and to perform recognition directly on this data. The data has the reconstruction quality that can be achieved by our real-time system. We also do not focus on classification of isolated action sequences from a pre-segmented dataset, but on recognition of activities inside the continuous workflow, involving several persons and objects. Due to the spatial constraints present in the OR, especially the fixed position and orientation of the patient table, no view invariance is required in our case.

For recognition, we use several 3D features, namely 3D occupation and 3D motion flow. 2D optical flow was recently used for action recognition by [Efros et al., 2003]. To our knowledge, no previous work has used 3D features for the analysis of complex workflow.

The AWHMMs introduced in the previous chapter are used as recognition approach. To deal with the alternative paths of actions, the a-priori annotation approach is applied and temporal relationships between the phases are automatically computed from the labeled data.
In order to model activities containing several actions occuring in parallel, [Shi et al., 2006] manually provided the relationships between the nodes within their propagation networks. In this model, each node modeled a single action. In [Vu et al., 2003], the relationships are given a-priori in form of formal rules. Contrary to these two work, [Xiang and Gong, 2008] interestingly proposed to learn the relationships between the activities, which were first classified by a dedicated classifier performing dimensionality reduction. The classifier does not use a temporal model. Adding a temporal modeling to each action would be similar to using a hierarchical HMM, which is a model more complicated to train in presence of a small amount of data containing a complex structure and without additional information. This is the main reason why we use a set of labeled data to compute the temporal structure of the workflow.

In the following, we denote as $\mathbb{O}$ the feature sequence computed out of a reconstruction sequence $\mathbf{r}_{1:T}$. Using training sequences $\mathbb{O}^1, \ldots, \mathbb{O}^l$, we generate a model λ that takes alternative courses of action into account and permits the computation of the probability

$P(phase \mid \mathbb{O}_{1:t})$ for each time-step t of the surgery $\mathbb{O}$.

7.2 Observations

The multi-camera system provides sequences of 3D occupation grids, also called visual hulls:

$$\{\mathbf{r}_{1:T}\}, \quad \mathbf{r} : \mathbf{v} \rightarrow \{0,1\}, \quad \mathbf{v} \in \Omega \subset \mathbb{R}^3 \ . \tag{7.1}$$

As explained in Section 3.2.1.1, individual tracking of objects is very challenging in the surgery room. For phase recognition, we compute features which coarsely describe the spatial distribution of occupancy and motion in the OR over time, without linking these to specific instances of persons and objects. We thereby rely on the fact that the different phases are discriminative with respect to different occupancy constellations and motion patterns appearing at key-locations in the OR. For instance, the *patient entering* (see section 3.2.2.2) can be identified by a strong motion close to the entrance of the OR, while the *surgery* phases are characterized through distinct motion patterns appearing in the neighborhood of the operation table.

In our experiments, motion features appeared as the most interesting. Occupancy based features revealed indeed to be poorly discriminative, even in combination with the motion features. Below, we present the computation of these two features.

7.2.1 Occupancy Features

Each visual hull $\mathbf{r}$ is split into $Q \times R \times S$ evenly spaced cells $\mathbf{c}_i$. For each cell we compute occupancy features $\{\mathbf{O}_{1:Q \cdot R \cdot S}\}$ as the number of voxels that fall into the cell. In the following, for clarity we omit indices representing time and sequence number. To avoid quantization effect, voxels vote smoothly for each cell, weighted by radial basis functions centered at the cell's centroid:

$$\mathbf{O}_i = \sum_{\mathbf{v} \in \Omega} \mathbf{r}(\mathbf{v}) \exp\left(-\frac{(\mathbf{v} - \bar{\mathbf{c}}_i)^2}{2\sigma_i^2} \right) \ , \tag{7.2}$$

where $\bar{\mathbf{c}}_i$ is the centroid of cell $\mathbf{c}_i$, and σ_i controls the radius at which voxels contribute to a certain cell.

7.2.2 3D Motion Features

7.2.2.1 3D Optical Flow

In 3D, optical flow has mainly been used for medical image analysis. This is one of the rare domains were 3D+t datasets are available on a regular basis. Recovery of the motion field can be used for registration of several volumes acquired over time. 3D optical flow has been used to find mapping between PET volumes [Klein and Huesman, 1997], cardiac MRI volumes [Barron, 2004] and CT volumes

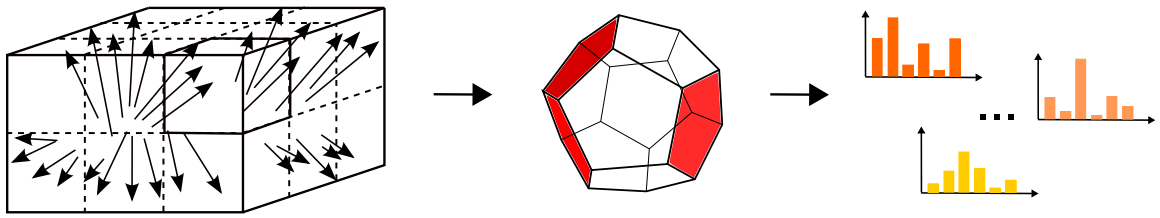


Figure 7.1: Computation of 3D flow histograms within the reconstructed volume, using regular polyhedron quantization.

[Zhang et al., 2008]. In [Chen et al., 2001], 3D optical flow was used on 3D doppler data to compute the velocities in storms. The extensions of the widely used 2D optical flow methods [Horn and Schunck, 1980] and [Lucas and Kanade, 1981] to 3D were presented in [Barron and Thacker, 2005, Simon Baker and Matthews, 2004].

We describe the 3D Lucas-Kanade optical flow, which will be used for the computation of the motion features. Using the intensity constancy assumption, the optical flow constraint at a voxel $\mathbf{v}$ is given as:

$$\frac{\delta \mathbf{r}(\mathbf{v})}{\delta \mathbf{v}} \cdot \mathbf{f}(\mathbf{v}) + \frac{\delta \mathbf{r}(\mathbf{v})}{\delta t} = 0 \quad , \tag{7.3}$$

with $\mathbf{f}(\mathbf{v})$ representing the velocity vectors. For each velocity vector, three parameters have to be estimated. This requires additional equations, obtained by assuming that the velocities to be locally constant around each voxel. For a voxel $\mathbf{v}$, this assumption yields a set of linear equations

$$\left.\begin{array}{lcl} \frac{\delta \mathbf{r}(\mathbf{v_1})}{\delta \mathbf{v}} \cdot \mathbf{f}(\mathbf{v}) & + & \frac{\delta \mathbf{r}(\mathbf{v_1})}{\delta t} = 0 \\ \frac{\delta \mathbf{r}(\mathbf{v_2})}{\delta \mathbf{v}} \cdot \mathbf{f}(\mathbf{v}) & + & \frac{\delta \mathbf{r}(\mathbf{v_2})}{\delta t} = 0 \\ & \vdots & \\ \frac{\delta \mathbf{r}(\mathbf{v_n})}{\delta \mathbf{v}} \cdot \mathbf{f}(\mathbf{v}) & + & \frac{\delta \mathbf{r}(\mathbf{v_n})}{\delta t} = 0 \end{array}\right\} \quad , \tag{7.4}$$

where $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$ are the voxels in a window centered at $\mathbf{v}$. The equation system can be rewritten in the matrix form as $AF = -b$ and solved using the least-square method, which provide the solution $F = (A^\top A)^{-1} A^\top (-b)$. In order to give more weight to the center voxels, an additional gaussian weighting W is used, actually yielding the equation: $WAF = -Wb$. Meaningful velocities are obtained after a reliability test based on a thresholding of the smallest eigenvalue obtained during the decomposition of $(A^\top W^2 A)$. Our implementation extends to 3D the code available for the 2D case in the Piotr matlab toolbox [1].

7.2.2.2 Motion Features

As for occupancy features, we split the reconstruction volume into a set of $Q \times R \times S$ evenly spaced cells $\mathbf{c}$. For each cell, we compute a histogram of 3D motion orientations (see

1. `http://vision.ucsd.edu/~pdollar/toolbox/doc/`

Figure 7.1). To avoid quantization effects at the boundaries of the cells and to introduce invariance to small variations, we use a smooth voting scheme for histogram computation based on radial basis functions. From the occupation grid sequences $\mathbf{r}$ we compute 3D optical-flow sequences as explained in previous section:

$$\{\mathbf{f}_{1:T}\}, \quad \mathbf{f}(\mathbf{v}) : \Omega \rightarrow \mathbb{R}^3, \quad \Omega \subset \mathbb{R}^3 \quad , \tag{7.5}$$

The flow vectors $\mathbf{f}(\mathbf{v})$ for all voxels $\mathbf{v}$ are then quantized into n orientation bins $\tilde{h}_1(\mathbf{v}), \ldots, \tilde{h}_n(\mathbf{v})$ using a soft voting scheme, and accumulated into histograms

$$\{\mathbf{H}_{1:Q \cdot R \cdot S, 1:n}\} \quad . \tag{7.6}$$

One histogram is computed for each of the $Q \times R \times S$ evenly spaced cells $\mathbf{c}_i$:

$$\mathbf{H}_{ij} = \sum_{\mathbf{v} \in \Omega} \tilde{h}_j(\mathbf{v}) \exp\left(-\frac{(\mathbf{v} - \bar{\mathbf{c}}_i)^2}{2\sigma_i^2}\right) \quad , \tag{7.7}$$

where $\bar{\mathbf{c}}_i$ is the centroid of cell $\mathbf{c}_i$, and σ_i controls the radius at which flow vectors contribute to a certain cells.

To quantize the 3D flow vectors, we adapt an approach proposed by [Kläser et al., 2008], which uses regular polyhedrons to define the histogram bins. This quantization scheme avoids well-known problems of standard spherical quantizations, such as singularities close to the poles and decreased accuracy close to the equator [Kläser et al., 2008]. Polyhedrons generalize polygons in 3D. In order to regularly approximate a sphere [2], there are only five regular polyhedrons with respectively 4 (tetrahedron), 6 (cube), 8 (octagon), 12 (dodecahedron) and 20 (icosahedron) faces. In this approach, quantization is performed by projecting 3D flow vectors onto the homogeneously distributed face normals. The face normals can be easily computed out of the face centers. For instance, a dodecahedron centered at the origin can be described by the following 12 face centers:

$$(0, \pm 1, \pm\varphi) \quad (\pm 1, \pm\varphi, 0) \quad (\pm\varphi, 0, \pm 1) \quad , \tag{7.8}$$

with $\varphi = \frac{1+\sqrt{5}}{2}$ being the golden ratio.

Let $\{\mathbf{p}_{1:n}\}$ be the face normals of a regular n-sided polyhedron. A flow vector $\mathbf{f}(\mathbf{v})$ votes for each bin as follows:

$$h_i(\mathbf{v}) \quad = \quad max\left(\frac{\mathbf{f}(\mathbf{v})^{\mathrm{T}} \cdot \mathbf{p}_i}{\|\mathbf{f}(\mathbf{v})\|_2} - q, 0\right) \quad . \tag{7.9}$$

q is chosen such that a flow vector lined up with one of the bin's normals will only vote for this bin, that is $q = \mathbf{p}_j^{\top}\mathbf{p}_k$, with $\mathbf{p}_j, \mathbf{p}_k$ being direct neighbors (see Figure 7.2). Otherwise, it votes for several neighboring bins, in proportion to its proximity to the corresponding bin.

The final contribution of flow vector $\mathbf{f}(\mathbf{v})$ into histogram bin $\tilde{h}(\mathbf{v})_i$ is computed by normalizing the values of $h_i(\mathbf{v})$ to the unit and weighting them by the flow vectors magnitude:

$$\tilde{h}(\mathbf{v})_i = \frac{\|\mathbf{f}(\mathbf{v})\|_2 \cdot h_i(\mathbf{v})}{\sum_i h_i(\mathbf{v})} \quad . \tag{7.10}$$

The resulting motion flow is illustrated in Figure 7.3.

2. Question: what polyhedron is used to build a soccer ball?

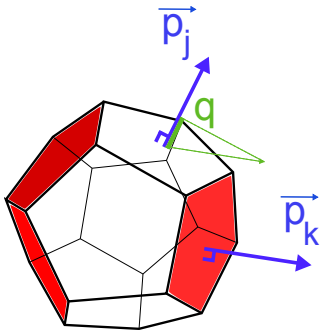


Figure 7.2: Notations for quantization using face normals.

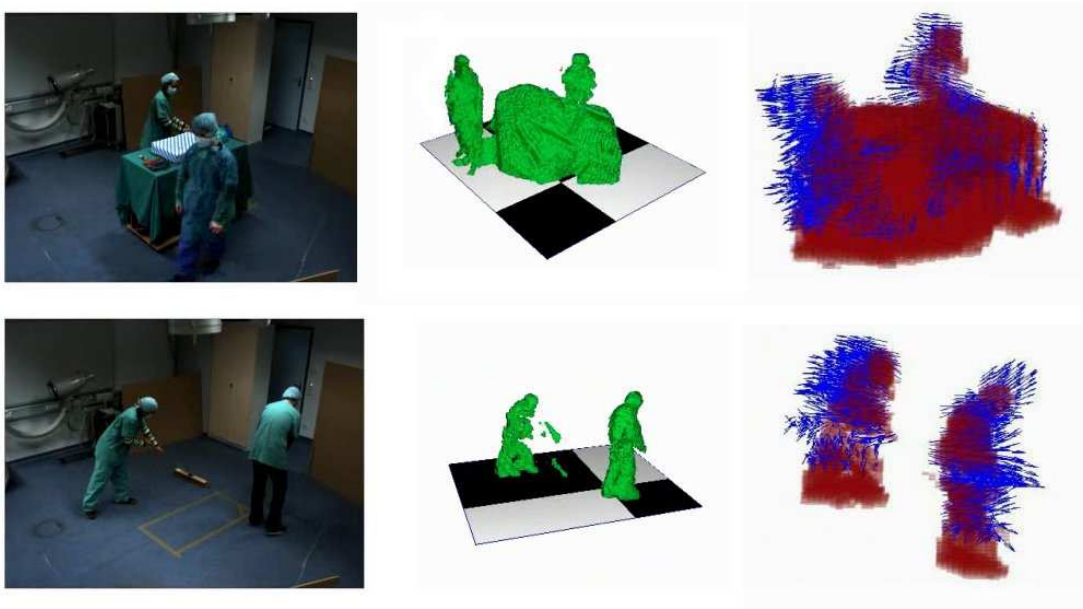


Figure 7.3: Two scenes from a surgery, with reconstruction and motion flow.

7.3 Modeling

In the cholecystectomy application, the surgery is performed continuously without interruptions. The surgical phases are contiguous and for this reason there is no need for an extra phase, modeling background or intermediate activities. In the daily OR workflow, it appeared necessary to use an additional label 'background' assigned to periods of time where nothing semantically interesting with respect to the workflow occurred. This corresponds to parts in the videos that were not labeled.

To deal with the non-linear workflow containing alternative paths, we build the AWHMM structure using a phase-wise construction approach (see Section 6.2.2.2). We use labeled data, and do not rely on structural learning. This is in particular of advantage for this dataset as only a small set of training sequences with continuous observations is available and signals do not contain any semantic. By using the labeling, we can directly derive a meaningful high level topology for concatenation of the sub-models. Effectively,

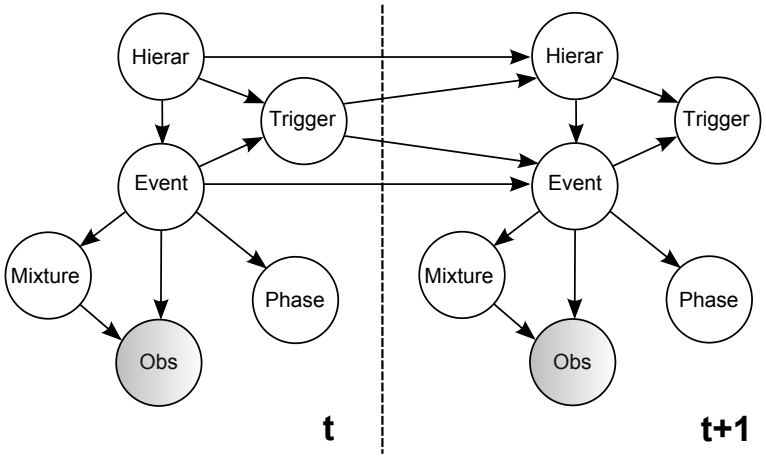


Figure 7.4: Dynamic Bayesian network representation of the AWHMMs: two-level hierarchy with a phase variable.

the model has a two-layer hierarchy: nodes on the top layer represent dependencies between distinct phases, while nodes on the bottom layer represent dependencies within individual phases. We construct the AWHMM using labeled and unlabeled observation sequences of the workflow. The construction involves the generation of the top-level topology, the initialization of the parameters in the lower hierarchy, and the overall training using EM to refine all probabilities.

The detailed structure of our AWHMM is illustrated using the convenient graphical description of dynamic Bayesian networks [Murphy, 2002] in Fig. 7.4. In this formalism, the random variables *Hierar*, *Event* and *Trigger* enforce the two-level hierarchy, which models the phases and their dependencies. *Mixture* and *Obs* represent the observation distributions occurring within the process. Finally, *Phase* models the probabilities of being in a phase knowing the current *Event*. As in the previous chapter, to model the topology we use a conventional HMM with a single state variable x, which encodes all possible combinations of the Markov states of the original model. The topology shown in Figure 7.4 is nevertheless preserved by enforcing its structure on the transition matrix of x, i.e. by setting non-possible transitions to zero. The construction of the model is explained in the next section 7.3.1.

7.3.1 Initialization of Model Parameters

The initialization consists in two steps: 1) generating the top-level topology by enforcing the temporal constraints between the phases using a set of labeled sequences $\mathbb{O}^1, \ldots, \mathbb{O}^l$. 2) initializing the bottom level from the labeled information.

As each labeled sequence provides the temporal relationships between its labels, a directed graph can be deterministically derived from the data, modeling the temporal relationships between the phases, ie. the overall workflow (such a graph is illustrated in Figure 7.5). For the sake of recognition, inter-phases, namely the chunks of observations between two consecutive labels, are also used to build nodes of the graph. Inter-phases

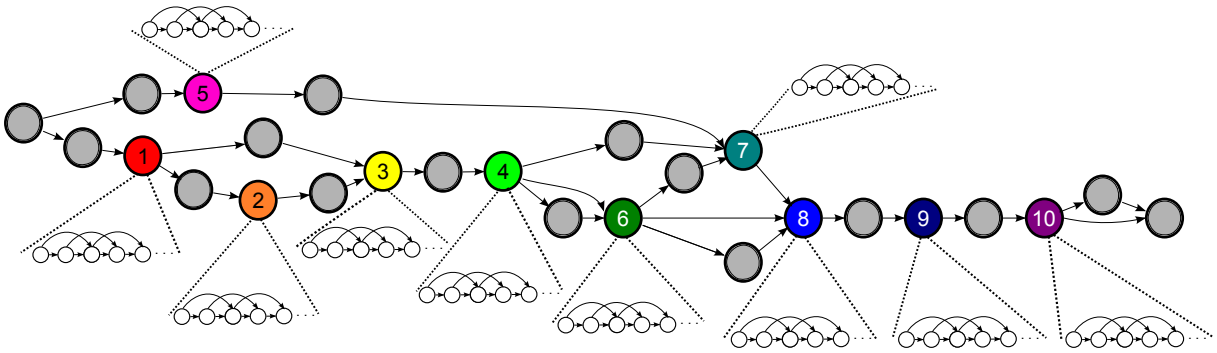


Figure 7.5: Graph representing the temporal relationships between the phases of the workflow in Figure 3.9, as extracted from annotated sequences. Colored nodes stand for labeled phases and gray nodes for inter-phases. The bottom levels of gray nodes are not displayed for visualization purposes.

model background and intermediary activities, which are also repetitive across instances of the workflow.

Figure 7.5 also shows that the resulting initialization is obtained by replacing each node of this graph with a sub-HMM modeling the corresponding group of events, as in Section 6.2.2.2. Each sub-HMM is first initialized using the sub-sequences of data corresponding to its label. We use a left-right model (see Section 6.2.1) with skip-one-ahead transitions and mixtures of two gaussians to model the continuous observations. The number of states is defined based on the average-length of the sub-sequences and the transition probabilities are set so that the expected duration corresponds to the duration of the phase. The mixtures are initialized by using k-means from their respective data after splitting temporally the training subsequences in as many chunks as available states in the sub-model. Each sub-HMM is finally trained independently with EM on the sub-sequences.

7.3.2 Training and Recognition

To train the model, we use EM and the refinement of phase probabilities after EM as described in the previous chapter in Section 6.2.2.4. The computation of the annotation function

$$\mathcal{P}_{\mathbb{O}^{test}} : \{1, \ldots, T\} \rightarrow \mathcal{L} \tag{7.11}$$

for a test surgery $\mathbb{O}^{test}$, both off-line and on-line, is also done as presented in Section 6.3.

7.4 Evaluation

In the experiments, we use a data-set composed of a series of 22 videos (illustrated in Figure 3.11 on page 40) of 3D volumetric reconstructions of resolution 128x128x128. We performed the recordings in a mock-up OR as described in Section 3.2.2.2. The number of occurrences of each phase within our dataset is indicated in the second column of Table

	Accuracy (%)	Average Recall (%)	Average Precision (%)
Off-line	92.5 (±7.0)	92.2 (±5.4)	89.9 (±8.4)
On-line	89.2 (±5.5)	87.8 (±5.2)	85.3 (±7.2)

Table 7.1: On-line results and off-line results for AWHMMs. Mean and standard deviation over 22 sequences using leave-one-out cross-validation.

7.2. We evaluate the phase recognition results both on-line and off-line. In a first set of experiments, we present general results and compare the two features. Then, using this complex dataset containing alternative paths and continuous observations, we illustrate the effect of the training process on phase probabilities, especially in case of partial labeling. Finally, we show the importance of incorporating the temporal constraints of the workflow within the model to obtain better recognition rates.

In practice, we use a dodecahedron with 12 bins for the quantization. Both for occupancy and motion features, the volume is split into a grid of $3 \times 3 \times 2$ cells and we set $\sigma_i = s_i * 0.6$, where s_i is the average spacing between cell $\mathbf{c}_i$ and its neighbors (see section 7.2.1). The resulting observations vectors, obtained by vectorization of the histograms, are passed as observations $\mathbb{O}_{1:T}$ to the recognition system after dimensionality reduction using PCA. For evaluation, we perform a full cross-validation. The presented results are averaged over all tests performed with the leave-one-out method.

7.4.1 General Results

The overall results using motion features only are presented in Table 7.1. They show slightly inferior recognition rates on-line than off-line. On-line results for each phase are given in table 7.2. This table shows reliable detection using 3D-flow, except for phase 5. Interestingly, this phase corresponds to an emergency, which is actually defined as an anomaly in our workflow. Indeed, it consists of the accelerated performance of the first preparation phases and only occurs 4 times in our dataset. When wrongly detected, it is recognized as these similar phases. Standard deviations in the precision measures are higher in the presence of alternatives, since the beginning of the phases are ambiguous, also compared to the background phase.

The occupancy features appear to be less discriminative, especially because in many phases, the actors are at similar places. For comparison, the recognition rates using occupation features only or using a combination of occupation and motion features are presented in Table 7.3.

7.4.2 Partial Labeling

The phase probabilities are used in general to attach semantic information to the model. Their utility to track semantic shift during training appears in particular in experiments conducted with sub-sets of annotated sequences and sub-sets of the labeled phases to be recognized.

	Phase	#	Recall (%)	Precision (%)
B	Background	22	91.1 (±5.4)	85.7 (±6.7)
1	Patient Entering	18	86.0 (±8.3)	76.0 (±38.0)
2	Anesthesy Cont.	5	86.2 (±9.0)	96.8 (±5.7)
3	Patient Prep.	18	86.5 (±16.9)	89.6 (±22.0)
4	Surgeon Prep.	18	89.7 (±11.3)	94.6 (±4.5)
5	Emergency	4	15.8 (±12.5)	91.4 (±14.8)
6	Min. Invasive	11	85.7 (±10.3)	76.9 (±41.7)
7	Open Surgery	15	89.3 (±14.1)	68.8 (±41.4)
8	Suturing	22	87.3 (±10.5)	86.1 (±11.9)
9	Patient Leaving	22	94.8 (±3.5)	87.2 (±10.9)
10	Cleaning	22	96.0 (±7.1)	99.0 (±1.7)

Table 7.2: On-line results presented per phase using AWHMMs. Mean and standard deviation over 22 sequences using leave-one-out cross-validation. Column *Phase* indicates the phase label, as in Figure 3.9. B stands for the phase modeling background activity. Column # shows the number of occurrences of each phase within the dataset.

	Accuracy (%)		Average Recall (%)		Average Precision (%)	
	occ	occ+mot	occ	occ+mot	occ	occ+mot
Off-line	47.4	89.8	41.9	86.8	50.5	84.7
On-line	61.8	85.2	53.4	80.5	64.0	82.8

Table 7.3: On-line and off-line results, using occupancy features (occ) and both occupancy and motion features (occ+mot). Mean over 22 sequences using leave-one-out cross-validation. Results for motion features only are presented in Table 7.1.

		Accuracy (%)	Average Recall (%)	Average Precision (%)
Off-line	NO	90.5 (±13.1)	88.4 (±18.2)	87.4 (±18.6)
	EM	86.6 (±13.9)	84.8 (±16.2)	73.9 (±22.4)
	PV	92.9 (±6.2)	91.3 (±9.1)	75.8 (±15.6)
On-line	NO	78.6 (±12.2)	70.5 (±16.4)	66.7 (±19.3)
	EM	80.7 (±9.5)	75.3 (±12.7)	71.5 (±13.4)
	PV	84.2 (±6.7)	78.8 (±10.7)	74.8 (±11.9)

Table 7.4: Results in percent using solely labels 3, 6, 7 and 8. Mean and standard deviation over all sequences using leave-one-out cross-validation. Comparison on-line and off-line, without EM (NO), with EM only (EM) and with EM followed by recomputation of phase probability variables (PV).

Figure 7.6 shows the influence of EM training and of phase probabilities. We see that performing the phase probability computation after EM training improves the results. Additionally, this figure shows how the overall results vary depending on the percentage of labeled sequences available in the training set. For this experiment, results were computed from a single random split of the sequences into training and test data for each number of labeled sequences and cross-validation test. Results are expected to get smoother if they are averaged on all the possible subsets.

Results using only a subset of the labels are shown in Table 7.4 . In this experiment, only four labels are available in the training sequences, corresponding to phases 3, 6, 7 and 8. Phase probabilities also improve significantly the results, the means and the standard deviations, as a semantic shift during EM can easily occur in the long background phases.

7.4.3 Temporal Constraints

To illustrate the importance of considering the temporal constraints of the workflow with alternative paths, we compare AWHMM with two methods that do not use these constraints. In the first method used for comparison, named MAP-HMMs, all phases are modeled independently by different HMMs trained on sub-windows of the data. Maximum likelihood classification is performed at each time-step using a sliding-window. There is therefore no distinction between off-line and on-line for this approach. In the second method, the sub-HMMs modeling all phases are arranged in parallel and connected via another sub-HMM modeling background activity, as shown in Figure 7.7. We use this form of connections instead of fully inter-connected HMMs, having a single or multiple HMMs modeling the background activities, since it showed better results. We call this approach CO-HMMs.

As can be seen in Table 7.5, AWHMM outperforms both CO-HMMs and MAP-HMMs approaches, showing the importance of using all information provided by the annotation. A difficulty when performing on-line recognition with MAP-HMMs on phases with highly varying lengths, is the choice of the window size. We performed experiments with different sizes but could not match the results of the two other approaches. CO-HMMs

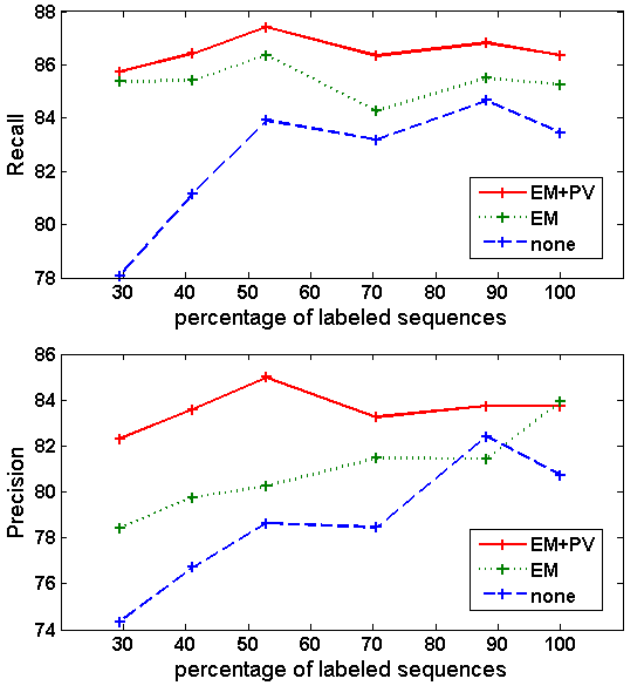


Figure 7.6: Precision and Recall of AWHMMs as a function of the percentage of annotated sequences in the training set. On-line results before EM (none), after global EM (EM) and after global and computation of phase variables (EM+PV). Mean over all sequences using leave-one-out cross-validation.

perform worse than AWHMM as they are less constrained and do not take advantage of all temporal relationships. An additional disadvantage of CO-HMMs are the numerous short and incorrect transitions that occur both off-line and on-line. This effect also occurs on-line for AWHMMs, but in a much smaller scale, as shown in Figure 7.8. This figure shows the number of transitions occurring between all pairs of phases during a complete cross-validation test. Self-transitions on the diagonal were removed for better visualization and the background phase is not taken into account. CO-HMMs produce 4 times more short incorrect transitions than AWHMMs.

7.5 Conclusion

In this chapter, we have given a simple but effective way to initialize a AWHMM from the labeling of a small set of training data. It was demonstrated on continuous data from a workflow containing alternative paths of activities. An interesting future work is to automatically discover the topology of the workflow from the available data without using the labeling information, similarly to the workflow mining application presented in Section 6.6.3. Initial work in this direction using continuous data can be found in

	Accuracy (%)			Average Recall (%)			Average Precision (%)		
	AW	CO	MA	AW	CO	MA	AW	CO	MA
Off-line	92.5	79.4	70.5	92.2	80.7	63.6	89.9	72.8	53.5
On-line	89.2	78.2	70.5	87.8	79.5	63.6	85.3	62.3	53.5

Table 7.5: Summarized results, comparing AWHMMs (AW), CO-HMMs (CO) and MAP-HMMs (MA), on-line and off-line, using EM and computation of phase probabilities. Mean over 22 sequences using leave-one-out cross-validation.

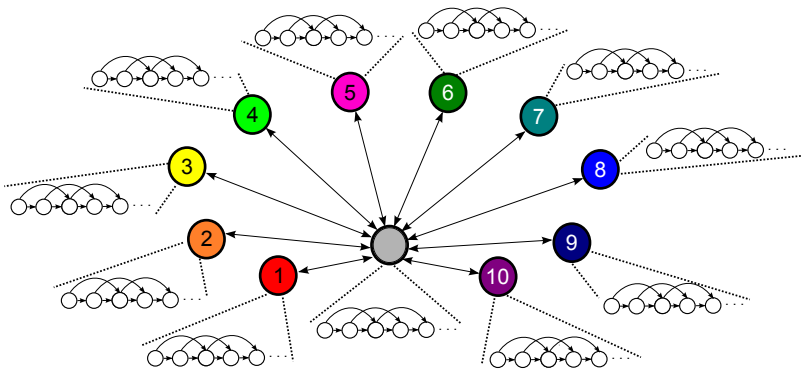


Figure 7.7: Parallel structure of CO-HMMs, where all phases are interconnected through the background phase.

[Brand and Kettnaker, 2000, Xie et al., 2004]. These methods will have to be adapted, implemented and experimented on relevant surgical signals.

We have also presented 4D features that can be used for workflow recognition inside complex scenes observed by a multi-camera reconstruction system. 3D motion flow yielded good results in our experiments. The approaches have been designed as a proof of concept. The major challenge to be tackled for such a direction, to become effective in the real intervention room, is the obtention of a good reconstruction under real clinical conditions. Indeed, additional issues such as highlights, shadows and systems permanently present in the room have to be taken into account. The sequences acquired in real conditions will also bring additional challenges, such as broader variations and more exceptions in the activities. The experiments presented in this chapter show that reconstruction data provides interesting input signals for surgical workflow analysis.

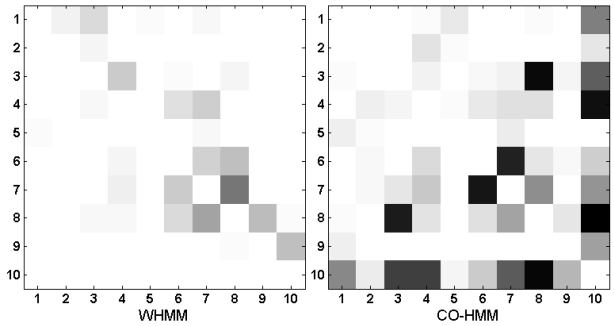


Figure 7.8: Occurring phase transitions, on-line, for AWHMMs and CO-HMMs. White color means absence of transition. The diagonal has been removed for better visualisation. Mean over all sequences using leave-one-out cross-validation.

Part III

Outlook and Conclusion

CHAPTER 8

Conclusion

In this chapter, we summarize the achievements presented within the dissertation. We also discuss the current limitations and present several ideas for future work aiming at recognizing activities within surgical workflow.

8.1 Summary

This dissertation has addressed a novel research field, namely the problem of recognizing activities inside the workflow taking place in Operating Rooms. This is highly motivated by the increasing complexity of ORs, due to the rapid development and introduction of new and high-tech surgical systems, and by the need to optimize the workflow. Using signals acquired in the OR, robust workflow activity recognition will permit the development of new assistance systems for the OR and the department of surgery. Examples of applications are context-aware support, for assistance to the surgical staff and easier monitoring and scheduling inside the department of surgery, automatic report generation and workflow optimization.

Operating rooms bring several new specificities and challenges to the activity recognition problem. Due to sterility requirements, safety standards and the fact that the OR workflow should not be modified or impaired, restrictions apply with respect to the kind of sensory information that can be used. Furthermore, the crowded environment, the small field of operation, the similar clothes and the changing lighting conditions bring additional difficulties for purely vision-based solutions.

Nonetheless, the operating room is getting increasingly high-tech and digital. Thus, there is little doubt about the fact that every activity performed in the OR will soon be sensed by some signal. Moreover, certain constraints, such as the strict and repetitive overall workflow in regular surgeries, can be used to facilitate the recognition task.

In this work, we layed some foundations towards recognition in the operating room and the use of workflow constraints in the modeling of the problem. We represented surgeries

in terms of multidimensional signals acquired over time and considered recognition in the surgical workflow as a learning problem based on previously acquired surgeries. First, we proposed to model the entire workflow within a single model, an annotated average surgery representation or an annotated workflow hidden Markov model, both containing phase probability variables to attach semantic information. The introduction of these variables allowed us to conveniently cope with model construction and training under partial labeling. They also prevent semantic information from getting blurred during training.

Second, we addressed two complementary applications: recognition inside a laparoscopic surgery and recognition inside a daily OR workflow. For each of these applications, we proposed specific signal representations. In the Cholecystectomy application, we model the activity by using instrument usage information. Recordings from real Cholecystectomy were used to demonstrate the concept. In the OR workflow application, we use 3D reconstruction data obtained from a real-time multi-camera system. The data was obtained by simulating an OR workflow containing alternative paths of activities in a mock-up OR. This is the first work that uses reconstruction data and 3D motion flow for the recognition of complex activities.

For both applications, we provided on-line and off-line recognition results. We also introduced several applications, such as the simultaneous replay and synchronisation of surgeries performed on different patients, the drafting of a surgical report and the prediction of remaining operating time.

Finally, we presented several efforts towards the automatic acquisition of additional signals in the OR: endoscopic tool signals using a vision-based approach relying on a trocar-camera and surgeon's hand movements using inertial sensors (appendix A).

8.2 Discussion and Future Work

The applications we have targeted have been largely determined by the availability of signals permitting regular recordings. Despite the interest and collaboration of our surgical partners, our cholecystectomy recordings have been fairly time-consuming and not all available signals could be obtained. For the study of other surgical workflows, future work has to focus on an easier and more automatic acquisition system that can be integrated in the OR and easily started when the patient is introduced. Additionally, for the demonstration of a recognition system during real surgeries, it is necessary that tool usage information and other informative signals become available in real-time. Ideally, further research towards a cognitive operating room will be facilitated by open standards and a fully digital Operating Room of the Future.

Obtaining real-time reconstruction data from the OR also faces important challenges. Indeed, the algorithm will have to cope with the reflective surfaces and illumination changes. Moreover, the permanent devices stationed in the room need to be included within the reconstruction. These difficulties could be alleviated by the complementary usage and fusion of time-of-flight camera information.
Our 4D features were based on motion computed from the visual hulls. It will be interesting to use textured reconstruction data for the computation of new 4D features

incorporating color information. Such data should be available in real-time soon. Furthermore, during real surgeries, reconstruction data shall in the end be only one part of the information. Recognition will highly benefit from the usage of the multiple available signals.

The methods that have been proposed for constructing the AVSR and AWHMM models all require supervision, since they need semantic information provided by the labeled data. Results however show that only a subset of the data needs to be labeled. A practical approach for the OR would be to label incrementally and at regular intervals parts of the new surgeries. Labeling should focus on phases that receive lower recognition accuracy than required by the application. In particular, in case the system is used by surgeons from different surgical schools, training data from the different surgeons is required, since the models can only recognize what has been learnt.
In order to more finely model differences inside phases or between surgeons, the hidden Markov models could be extended with further hierarchies and latent nodes. To avoid the tedious task of very finely labeling the data for training these bigger models, structure discovery approaches will need to be developped for the identification and modeling of these differences. Additionally, even though the models cope with variations in the surgical performance, novel phases cannot be recognized and are associated to the most probable existing phase. Discovering novel phases and anomalies, off-line and especially on-line by analysing the models' output, would be an interesting extension to this work.

The addition of novel semantic labels naturally creates alternative paths of activities in the workflow. The AWHMM approach then becomes indicated for recognition. But, where synchronization makes sense, for instance between common subparts of surgeries or between surgeries that follow the same workflow path, the AVSR approach can also be applied. This can yield more precise results by using only corresponding subsets of the data, e.g. identified with the AWHMM approach. For instance, to replay simultaneously two surgeries which do not exactly follow the same workflow path, AWHMMs can be applied to detect the common parts and also to find other surgeries following the same paths. The AVSR approach can then be used on this data for meaningful synchronization.

In this initial work, the modeling of the workflow has been designed for the on-line recognition of phases provided by an expert. While this level of semantic information already permits the development of interesting medical applications, larger semantic descriptions of surgical workflows already exist, which include different levels of granularity. Future research should focus on *combining* these descriptions with a statistical recognition model and with appropriate signals, so as to deliver recognition of additional and possibly finer events that can also occur simultaneously. Conversely, methods for structure discovery, topology optimization and workflow mining out of OR signals have to be investigated. They could permit us to generate a better model initialization from a small amount of data and also to refine and update the existing descriptions provided by experts with respect to the workflow effectively taking place in the OR. Extending the AWHMMs with respect to these two aspects, incorporation of finer medical semantic information and discovery of the internal structure of the surgical data, should be a main future direction.

In addition to recognition in the workflow, techniques have to be developed in parallel to identify all objects and persons present in the OR. This is a complementary objective,

as recognition in the workflow would highly benefit from additional signals. Reciprocally, disambiguation in the identification process could be achieved using the workflow context. Fusing information from the different systems together with the context would help in this regard. For instance, workflow phase information combined with RFID information about the staff present in the room could result in identifying the role and location of the persons located within the reconstruction volume.

Finally, a recognition system will have little impact on the workflow within the surgery room if not completed by the right interaction mechanisms. Multimedia user interfaces need to be designed and accessible to all personnel, taking into account the specificities of each kind of surgery. Integrated OR suites, as already provided by several companies, are the ideal testbed for the development of such advanced context-sensitive support.

Part IV

Appendix

APPENDIX A

Discovering a Surgical Vocabulary from Inertial Sensor Data

In the previous chapters, we have either used signals from tool usage or from reconstruction data. Tool usage captured activities within a surgical workflow, while reconstruction data captured the overall activity of the surgery room. We have investigated another potential source of information: accelerometer data from sensors attached to the wrists and the waist of the surgeon. We present here briefly some early work and ideas for these signals, published in [Ahmadi et al., 2008, Ahmadi et al., 2009].

As the raw accelerometer data is unstructured, we have experimented an approach to automatically discover a vocabulary of atomic activities from the signals. The objective is to reduce dimensionality by generating discrete signals with semantic information similar to the instrument usage out of this data. Using an automatic approach permits the generation of objective discrete signals without intensive labeling by an expert.

Besides the generation of supplementary signals that can be used for recognition in a surgical workflow, discovering a motion vocabulary from raw data can be potentially beneficial to other tasks such as the evaluation of surgical skills. For instance, in [Lin et al., 2006] surgical gestures are identified by an expert to define words of a "language of surgery". This language is further used to segment a surgeon's motions and to assess his skills. The automatic generation of an objective vocabulary from the data is a step towards the automatic definition of a motion "language" for each kind of surgery.

In section A.1, we present the introduction of inertial sensors in the surgery room. In section A.2, we present an approach from the motif discovery community that we applied on surgical data recorded in the context of the percutaneous vertebroplasties.

Figure A.1: (Left) Attachment of the inertial sensors with bandages. (Right) Setup during a vertebroplasty procedure.

A.1 Inertial Sensors in the OR

For the recordings, we used three 3D actibelt accelerometers designed by the Sylvia Lawry Centre for Multiple Sclerosis Research [1]. These accelerometers are light, wireless, compact and have a sampling rate of 100Hz. The acquired data is stored in the internal memory of the sensor and can in our case be retrieved after the surgery. For on-line data processing and analysis, wireless transmission of the data would be required. Convenient light-weight sensors with wireless transmission are for example available from the company Intersense [2].

An inertial sensor was fixed on the waist of the surgeon using a belt. To attach the sensors on his wrists, we used sterilizable TG tube bandages as displayed in Figure A.1. The sensors were fixed over the sterile bandages with Velcro fasteners. The sides of the bandages were then folded around the sensors. This setup is performed right before the surgeon puts on the surgical coat and the latex gloves over it. This protocol permitted us to ensure the sterility requirements and only introduced a quick additional step in the procedure's workflow.

We recorded several laparoscopic surgeries and percutaneous vertebroplasty procedures under these conditions. The setup was reported as comfortable by the surgeons. In the following, we present early experiments of automatic vocabulary generation in the percutaneous vertebroplasty procedure. This surgical intervention consists of the repair of a damaged vertebra by introduction of cement through a needle [Predey et al., 2002]. This is achieved under visual guidance and control with CT scans.

1. http://www.slcmsr.net/
2. http://www.intersense.com/

A.2 Vocabulary Generation

We consider a motion vocabulary to be interesting if its elements occur within almost all surgeries in the training set. For this reason, we search for similar temporal patterns occuring repetitively across the datasets. In the motif discovery community, such patterns are called motifs [Keogh and Lin, 2005]. Motifs can be seen as clusters in the space of time-series subsequences, using a time-invariant distance.

To discover a surgical vocabulary, we used an algorithm called *subsequence density estimation* [Minnen et al., 2007]. It permits the unsupervised identification of motifs and has been successfully used on accelerometer data [Minnen et al., 2006]. In our experiments, we search for motifs by enforcing the condition that instances are present in all datasets. The algorithm permitted the automatic selection of meaningful patterns from the data. Semantic was attached to these motions by labeling one instance of each motif using the corresponding video data. Interesting motions in the procedure are for instance *idle, tool change, needle positioning, hammering, cement stirring, syringe filling, applying syringe*.

We achieved promising results on data simulated on a phantom vertebra, where most interesting motifs were discovered. Up to now, only three real surgeries could be recorded. For this reason results on real data are poor. Additional data is required to cope with the motion variability. To validate the discovery of the vocabulary, we therefore recorded data of simulated surgeries. The discovered motifs were used to train an Hidden Markov Model for each motion. The resulting HMMs were finally applied for the redetection of these motions in surgeries that were not used in the motif discovery dataset. For six of the interesting motions, including *hammering, stirring, putting off vest*, cross-validated results yielded detection accuracies above 70%. These detections could be used in future work to provide intermediate signals. Further details on the results and experiments can be found in [Ahmadi, 2008, Rybachuk, 2009].

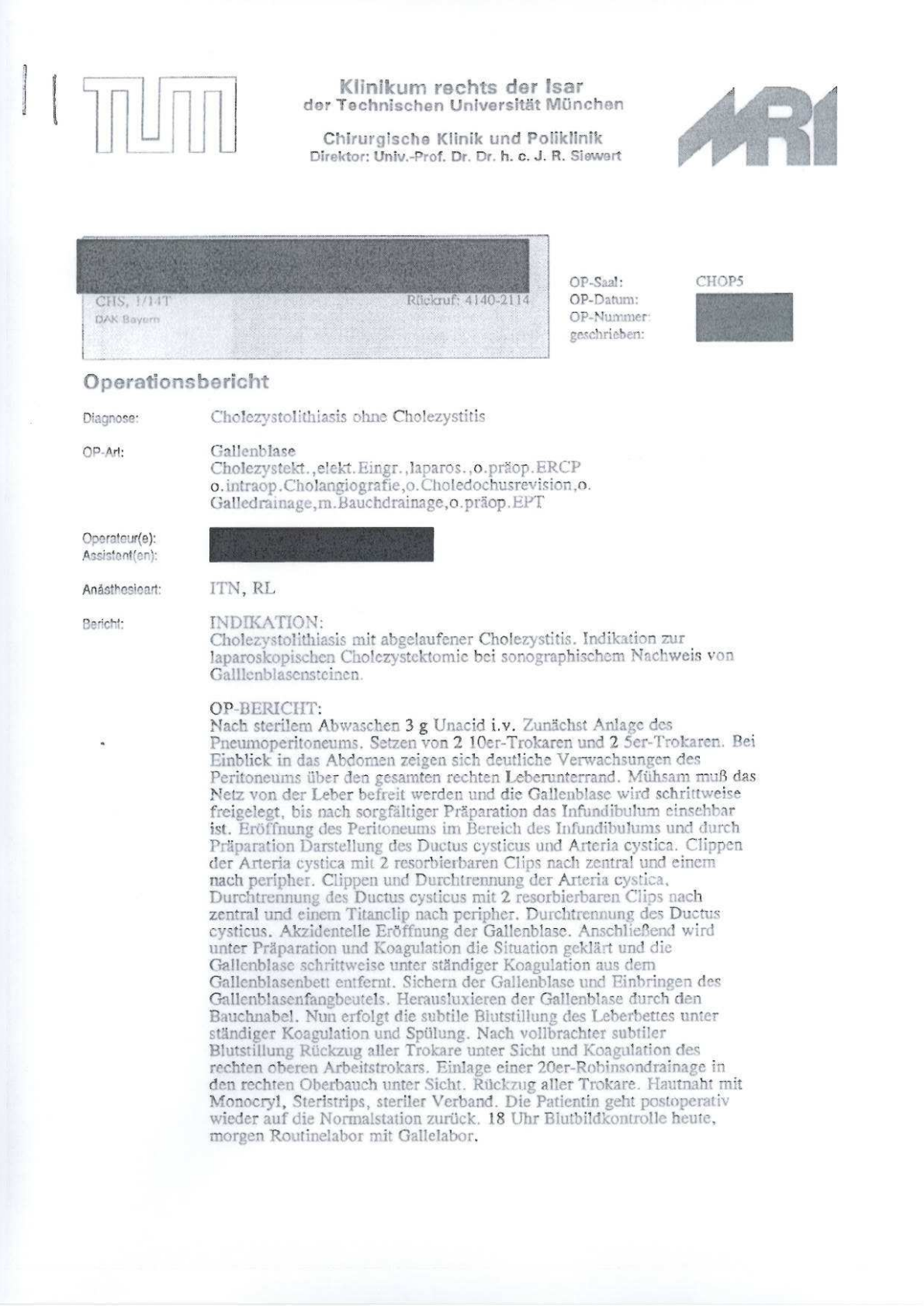
A.3 Conclusion

The definition of an intermediate language of activities is an interesting future direction for the dimensionality reduction and interpretation of unstructured data. The unsupervised generation of the vocabulary can in our case permit the generation of additional discrete signals for recognition. It could also be used for skills assessment [Lin et al., 2006] or database queries [Ikizler and Forsyth, 2008]. While our early experiments are promising in the sense that meaningful patterns can be discovered in the simulated surgical data, a larger dataset of real data is required for further validation. Furthermore, the algorithms are computationally expensive. This is a shortcoming as much data processing is required in order to cope with variations and surgeons' specificities. A practicable approach for vocabulary discovery in the data of complete surgeries needs to be scalable. Other unsupervised clustering techniques for time-series, for instance based on HMMs [Xie et al., 2004] or using embedded spaces, like spatio-temporal-isomaps [Jenkins and Matarić, 2004], could also be investigated.

APPENDIX B

Examples of Surgical Reports for Laparoscopic Cholecystectomy

Figures B.1 and B.2 in the next pages show two surgical reports written by surgeons from Hospital Rechts der Isar, Munich. Using the recognition methods developed in Part II, drafts from such reports could be generated automatically to include objective and quantitative information about the performance of the surgery, such as the beginning of the phases and the usage of the instruments. The surgeon would then only need to verify and to add complementary information to the report. Such information especially include the specific medical characteristics of each case. This is a next step of interesting future work.



Klinikum rechts der Isar
der Technischen Universität München

Chirurgische Klinik und Poliklinik
Direktor: Univ.-Prof. Dr. Dr. h. c. J. R. Siewert

[redacted]
CHS, 1/14T
DAK Bayern
Rückruf: 4140-2114

OP-Saal: CHOP5
OP-Datum: [redacted]
OP-Nummer:
geschrieben:

Operationsbericht

Diagnose: Cholezystolithiasis ohne Cholezystitis

OP-Art: Gallenblase
Cholezystekt.,elekt.Eingr.,laparos.,o.präop.ERCP
o.intraop.Cholangiografie,o.Choledochusrevision,o.
Galledrainage,m.Bauchdrainage,o.präop.EPT

Operateur(e): [redacted]
Assistent(en):

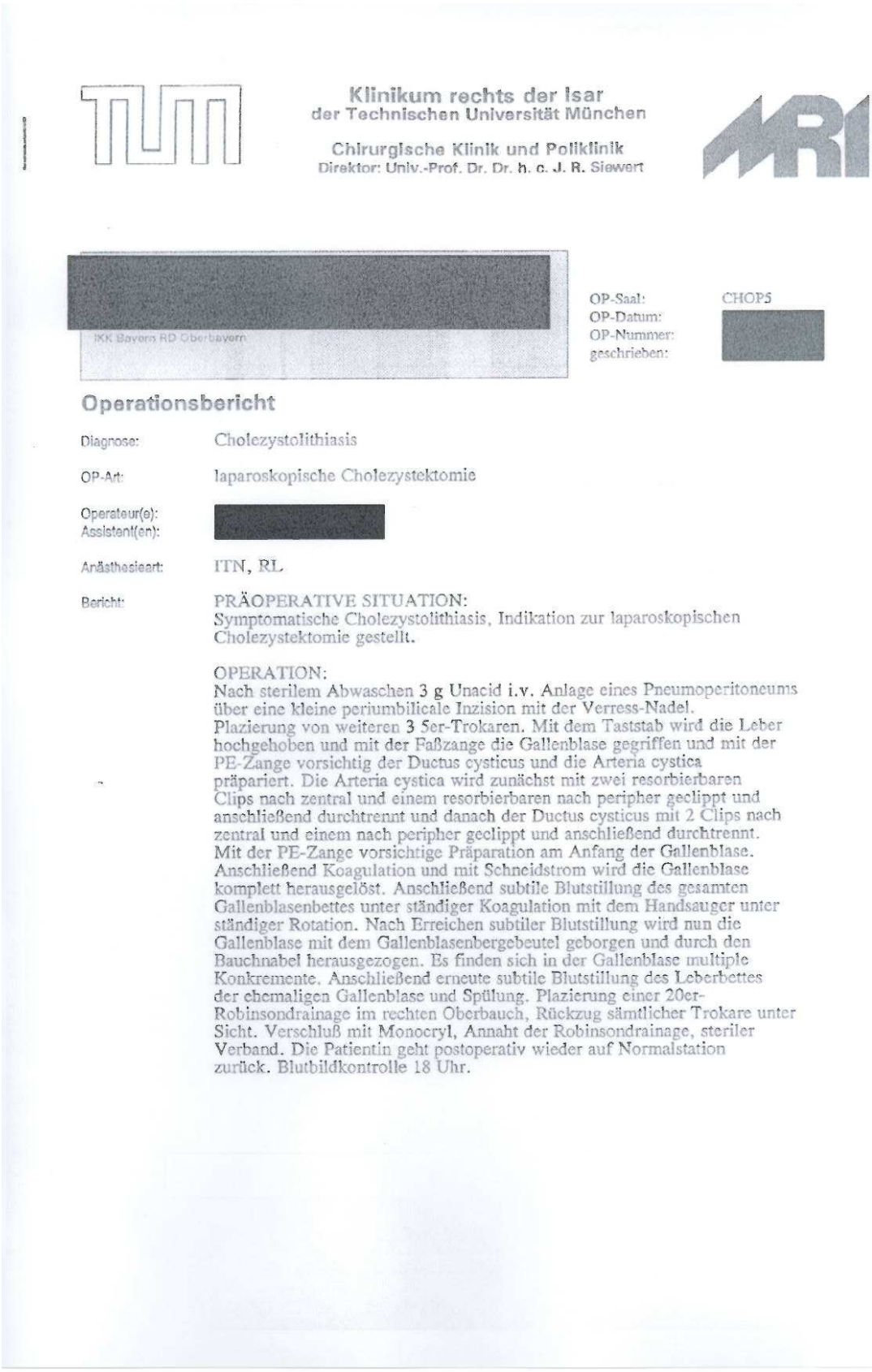
Anästhesieart: ITN, RL

Bericht:

INDIKATION:
Cholezystolithiasis mit abgelaufener Cholezystitis. Indikation zur laparoskopischen Cholezystektomie bei sonographischem Nachweis von Gallenblasensteinen.

OP-BERICHT:
Nach sterilem Abwaschen 3 g Unacid i.v. Zunächst Anlage des Pneumoperitoneums. Setzen von 2 10er-Trokaren und 2 5er-Trokaren. Bei Einblick in das Abdomen zeigen sich deutliche Verwachsungen des Peritoneums über den gesamten rechten Leberunterrand. Mühsam muß das Netz von der Leber befreit werden und die Gallenblase wird schrittweise freigelegt, bis nach sorgfältiger Präparation das Infundibulum einsehbar ist. Eröffnung des Peritoneums im Bereich des Infundibulums und durch Präparation Darstellung des Ductus cysticus und Arteria cystica. Clippen der Arteria cystica mit 2 resorbierbaren Clips nach zentral und einem nach peripher. Clippen und Durchtrennung der Arteria cystica, Durchtrennung des Ductus cysticus mit 2 resorbierbaren Clips nach zentral und einem Titanclip nach peripher. Durchtrennung des Ductus cysticus. Akzidentelle Eröffnung der Gallenblase. Anschließend wird unter Präparation und Koagulation die Situation geklärt und die Gallenblase schrittweise unter ständiger Koagulation aus dem Gallenblasenbett entfernt. Sichern der Gallenblase und Einbringen des Gallenblasenfangbeutels. Herausluxieren der Gallenblase durch den Bauchnabel. Nun erfolgt die subtile Blutstillung des Leberbettes unter ständiger Koagulation und Spülung. Nach vollbrachter subtiler Blutstillung Rückzug aller Trokare unter Sicht und Koagulation des rechten oberen Arbeitstrokars. Einlage einer 20er-Robinsondrainage in den rechten Oberbauch unter Sicht. Rückzug aller Trokare. Hautnaht mit Monocryl, Steristrips, steriler Verband. Die Patientin geht postoperativ wieder auf die Normalstation zurück. 18 Uhr Blutbildkontrolle heute, morgen Routinelabor mit Gallelabor.

Figure B.1: Surgical report of a laparoscopic cholecystectomy.



Klinikum rechts der Isar
der Technischen Universität München

Chirurgische Klinik und Poliklinik
Direktor: Univ.-Prof. Dr. Dr. h. c. J. R. Siewert

[redacted]
IKK Bayern RD Oberbayern

OP-Saal: CHOP5
OP-Datum: [redacted]
OP-Nummer:
geschrieben:

Operationsbericht

Diagnose: Cholezystolithiasis

OP-Art: laparoskopische Cholezystektomie

Operateur(e): [redacted]
Assistent(en):

Anästhesieart: ITN, RL

Bericht:

PRÄOPERATIVE SITUATION:
Symptomatische Cholezystolithiasis, Indikation zur laparoskopischen Cholezystektomie gestellt.

OPERATION:
Nach sterilem Abwaschen 3 g Unacid i.v. Anlage eines Pneumoperitoneums über eine kleine periumbilicale Inzision mit der Verress-Nadel. Plazierung von weiteren 3 5er-Trokaren. Mit dem Taststab wird die Leber hochgehoben und mit der Faßzange die Gallenblase gegriffen und mit der PE-Zange vorsichtig der Ductus cysticus und die Arteria cystica präpariert. Die Arteria cystica wird zunächst mit zwei resorbierbaren Clips nach zentral und einem resorbierbaren nach peripher geclippt und anschließend durchtrennt und danach der Ductus cysticus mit 2 Clips nach zentral und einem nach peripher geclippt und anschließend durchtrennt. Mit der PE-Zange vorsichtige Präparation am Anfang der Gallenblase. Anschließend Koagulation und mit Schneidstrom wird die Gallenblase komplett herausgelöst. Anschließend subtile Blutstillung des gesamten Gallenblasenbettes unter ständiger Koagulation mit dem Handsauger unter ständiger Rotation. Nach Erreichen subtiler Blutstillung wird nun die Gallenblase mit dem Gallenblasenbergebeutel geborgen und durch den Bauchnabel herausgezogen. Es finden sich in der Gallenblase multiple Konkremente. Anschließend erneute subtile Blutstillung des Leberbettes der ehemaligen Gallenblase und Spülung. Plazierung einer 20er-Robinsondrainage im rechten Oberbauch, Rückzug sämtlicher Trokare unter Sicht. Verschluß mit Monocryl, Annaht der Robinsondrainage, steriler Verband. Die Patientin geht postoperativ wieder auf Normalstation zurück. Blutbildkontrolle 18 Uhr.

Figure B.2: Surgical report of a laparoscopic cholecystectomy.

APPENDIX C

List of Abbreviations

ADTW	Adaptive Dynamic Time Warping
AR	Augmented Reality
AVSR	Annotated Virtual Surgery Representation
AWHMM	Annotated Workflow Hidden Markov Model
CT	Computed Tomography
DBN	Dynamic Bayesian Network
DICOM	Digital Imaging and COmmunications in Medicine
DTW	Dynamic Time Warping
EM	Expectation Maximization
EMR	Electronic Medical Record
ESS	Expected Sufficient Statistics
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
MAP	Maximum A Posteriori
MRI	Magnetic Resonance Imaging
OR	Operating Room
ORF	Operating Room of the Future
PACS	Picture Archiving and Communication System
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RFID	Radio Frequency IDentification
TIMMS	Therapy Imaging and Model Management System
UML	Unified Markup Language

APPENDIX D

List of Recurring Mathematical Notations

$\mathbb{O}$	Time series of observations representing a surgery
$\overline{\mathbb{O}}$	Time series representing a virtual surgery
K	Number of instruments used in the observation vectors of cholecystectomies
T	Length of a time-series
Ω	Spatial area reconstructed by the 4D system
$\mathbf{r}$	Visual hull
$\mathbf{v}$	Voxel position
p	Phase label
$\mathcal{L}$	Set of phase labels
$\mathcal{P}$	Labeling function, associating a phase label to each time-step of a time-series
$\#\cdot$	Cardinality operator
l	Number of training sequences
$\mathbb{P}$	Time series representing a phase
$\overline{\mathbb{P}}$	Time series representing a virtual phase
ξ	The annotation of an AVSR
X	Random variable indicating the state of an HMM
x	HMM state
$\mathcal{O}$	Set of observations
O	Observation
λ	Set of parameters of an HMM
N	Number of states of an HMM
A	Transition matrix of an HMM
B	Observation model of an HMM
π	Initial probabilities of an HMM
ϕ	Phase variables
$\text{sync}_{i \leftrightarrow j}$	Warping between two time-series as provided by DTW

APPENDIX E

List of Publications

i. Padoy, N., Mateus, D., Weinland, D., Berger, M.-O., and Navab, N. Workflow monitoring based on 3d motion features. In: Proceedings of the International Conference on Computer Vision (ICCV), Workshop on Video-oriented Object and Event Classification. 2009.

ii. Pauly, O., Padoy, N., Poppert, H., Esposito, L., and Navab, N. Wavelet Energy Map: A Robust Support for Multi-modal Registration of Medical Images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pages 2184-2191. 2009.

iii. Pauly, O., Padoy, N., Poppert, H., Esposito, L., Eckstein, H-H., and Navab, N. Towards Application-specific Multi-modal Similarity Measures: a Regression Approach. In: MICCAI Workshop on Probabilistic Models for Medical Image Analysis (PMMIA). 2009.

iv. Ahmadi, S.-A., Padoy, N., Rybachuk, K., Feussner, H., Heining, S. M., and Navab, N. Motif Discovery in OR Sensor Data with Application to Surgical Workflow Analysis and Activity Detection. In: MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI). 2009.

v. Radrich, H., Padoy, N., Ahmadi, S.-A., Feussner, H., Hager, G., Burschka, D., and Knoll, A. Synchronized Multimodal Recording System for Laparoscopic Minimally Invasive Surgeries. In: MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI). 2009.

vi. Ahmadi, S.-A., Padoy, N., Heining, S. M., Feussner, H., Daumer, M., and Navab, N. Introducing wearable accelerometers in the surgery room for activity detection. In: 7. Jahrestagung der Deutschen Gesellschaft für Computer-und Roboter-Assistierte Chirurgie (CURAC). 2008.

vii. Blum, T., Padoy, N., Feussner, H., and Navab, N. 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. 2008a.

viii. Traub, J., Ahmadi, S.-A., Padoy, N., Wang, L., Heining, S. M., Euler, E., Jannin, P., and Navab, N. Workflow based assessment of the camera augmented mobile c-arm system. In: MICCAI Workshop on Augmented environments for Medical Imaging including Augmented Reality in Computer-aided Surgery (AMIARCS). 2008.

ix. Padoy, N., Blum, T., Feussner, H., Berger, M.-O., and Navab, N. 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. 2008.

x. Blum, T., Padoy, N., Feussner, H., and Navab, N. Workflow mining for visualization and analysis of surgeries. International Journal of Computer Assisted Radiology and Surgery, 3(5):379-386. 2008b.

xi. Klank, U. F., Padoy, N., Feussner, H., and Navab, N. An automatic approach for feature generation in endoscopic images. International Journal of Computer Assisted Radiology and Surgery, 3(3-4):331-339. 2008.

xii. Padoy, N., Blum, T., Essa, I., Feussner, H., Berger, M.-O., and Navab, N. A boosted segmentation method for surgical workflow analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 102-109. 2007a.

xiii. Padoy, N., Horn, M., Feussner, H., Berger, M.-O., and Navab, N. Recovery of surgical workflow: a model-based approach. In: International Journal of Computer Assisted Radiology and Surgery (CARS), Supplement 1. 2:481- 482. 2007b.

xiv. Groher, M., Jakobs, T. F., Padoy, N., and Navab, N. Planning and intraoperative visualization of liver catheterizations: New cta protocol and 2d-3d registration method. Academic Radiology, 14:1324-1339. 2007.

xv. Cohen, J., Jeannot, E., Padoy, N., and Wagner, F. Messages scheduling for parallel data redistribution between clusters. IEEE Trans. Parallel Distrib. Syst., 17(10):1163-1175. 2006.

xvi. Groher, M., Padoy, N., Jakobs, T. F., and Navab, N. New cta protocol and 2d-3d registration method for liver catheterization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). (1):873-881. 2006.

xvii. Groher, M., Jakobs, T. F., Padoy, N., and Navab, N. Towards a feature-based 2d-3d registration method of cta and 2d angiograms for liver tumor chemoembolizations. In: 4. Jahrestagung der Deutschen Gesellschaft für Computerund Roboter-Assistierte Chirurgie (CURAC). 2005.

xviii. Cohen, J., Jeannot, E., and Padoy, N. Messages scheduling for data redistribution between clusters. In: 5th International Conference on Parallel Processing and Applied Mathematics (PPAM). pages 896-906. 2003.

APPENDIX F

Abstract of Major Publications Not Discussed in this Thesis

Towards Application-specific Multi-modal Similarity Measures: a Regression Approach

Olivier Pauly, Nicolas Padoy, Holger Poppert, Lorena Esposito, Hans-Henning Eckstein and Nassir Navab

In multi-modal registration, similarity measures based on intensity statistics are the current standard for aligning medical images acquired with different imaging systems. In fact, the statistical relationship relating the intensities of two multi-modal images is constrained by the application, defined in terms of anatomy and imaging modalities. In this paper, we present the benefits of exploiting application-specific prior information contained in one *single* pair of registered images. By varying the relative transformation parameters of registered images around the ground truth position, we explore the manifold described by their joint intensity distributions. An adapted measure is fitted using support vector regression on the training set formed by points on the manifold and their respective geometric errors. Experiments are conducted on two different pairs of modalities, MR-T1/MR-TOF and MR-T1/SPECT. We compare the results with those obtained using mutual information and Kullback-Leibler distance. Experimental results show that the proposed method presents a promising alternative for multi-modal registration.

Wavelet Energy Map: A Robust Support for Multi-modal Registration of Medical Images

Olivier Pauly, Nicolas Padoy, Holger Poppert, Lorena Esposito, and Nassir Navab

Multi-modal registration is the task of aligning images from an object acquired with different imaging systems, sensors or parameters. The current gold standard for medical images is the maximization of mutual information by computing the joint intensity distribution. However intensities are highly sensitive to various kinds of noise and denoising

is a very challenging task often involving a-priori knowledge and parameter tuning. We propose to perform registration on a novel robust information support: the wavelet energy map, giving a measure of local energy for each pixel. This spatial feature is derived from local spectral components computed with a redundant wavelet transform. The multi-frequential aspect of our method is particularly adapted to robust registration of images showing tissues, complex textures and multiple interfaces. We show that the wavelet energy map approach outperforms the classical framework in rigid registration experiments on synthetic, simulated and real data, whether noise is present or not.

New CTA Protocol and 2D-3D Registration Method for Liver Catheterization

Martin Groher, Nicolas Padoy, Tobias F. Jakobs, and Nassir Navab

2D-3D registration for angiographic liver interventions is an unsolved problem mainly because of two reasons. First, a suitable protocol for Computed Tomography Angiography (CTA) to contrast liver arteries is not used in clinical practice. Second, an adequate registration algorithm which addresses the difficult task of aligning deformed vessel structures has not been developed yet. We address the first issue by introducing an angiographic CT scanning phase and thus create a strong link between radiologists and interventionalists. The scan visualizes arteries similar to the vasculature captured with an intraoperative C-arm acquiring Digitally Subtracted Angiograms (DSAs). Furthermore, we propose a registration algorithm using the new CT phase that aligns arterial structures in two steps: a) Initialization of one corresponding feature using vessel diameter information, b) optimization on three rotational and one translational parameter to register vessel structures that are represented as centerline graphs. We form a space of good features by iteratively creating new graphs from projected centerline images and by restricting the correspondence search only on branching points (the vertices) of the vessel tree. This algorithm shows good convergence and proves to be robust against deformation changes, which is demonstrated through studies on one phantom and three patients.

References

[Agarwal et al., 2007] Agarwal, S., Joshi, A., Finin, T., Yesha, Y., and Ganous, T. 2007. A pervasive computing system for the operating room of the future. Mobile Networks and Applications, 12(2-3):215–228.

[Aggarwal and Cai, 1999] Aggarwal, J. K. and Cai, Q. 1999. Human motion analysis: A review. Computer Vision Image Understanding Journal (CVIU), 73:90–102.

[Ahmadi, 2005] Ahmadi, A. 2005. First steps towards monitoring and recovery of surgical workflow. Bachelor Thesis. Technische Universität München.

[Ahmadi, 2008] Ahmadi, A. 2008. Discovery and detection of surgical activity in percutaneous vertebroplasty. Diploma Thesis. Technische Universität München.

[Ahmadi et al., 2008] Ahmadi, S.-A., Padoy, N., Heining, S. M., Feussner, H., Daumer, M., and Navab, N. 2008. Introducing wearable accelerometers in the surgery room for activity detection. In: 7. Jahrestagung der Deutschen Gesellschaft für Computer-und Roboter-Assistierte Chirurgie (CURAC 2008), Leipzig, Germany.

[Ahmadi et al., 2009] Ahmadi, S.-A., Padoy, N., Rybachuk, K., Heining, S. M., Feussner, H., and Navab, N. 2009. Motif discovery in or sensor data with application to surgical workflow analysis and activity detection. In: MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), London, GB. MICCAI Society.

[Ahmadi et al., 2006] Ahmadi, S.-A., Sielhorst, T., Stauder, R., Horn, M., Feussner, H., and Navab, N. 2006. Recovery of surgical workflow without explicit models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 420–428.

[Bao and Intille, 2004] Bao, L. and Intille, S. S. 2004. Activity recognition from user-annotated acceleration data. Pervasive 2004, pages 1–17.

[Bar-Or et al., 2004] Bar-Or, S. A., Bar-or, A., Goddeau, D., Healey, J., Kontothanassis, L., Logan, B., Nelson, A., and Thong, J. V. 2004. Biostream: A system architecture for real-time processing of physiological. In: Proc. of IEEE Engineering in Medicine and Biology Society. Springer, pages 3101–3104.

[Bardram, 2004] Bardram, J. E. 2004. Applications of context-aware computing in hospital work: examples and design principles. In: SAC '04: Proceedings of the 2004 ACM symposium on Applied computing, New York, NY, USA. ACM, pages 1574–1579.

[Barron, 2004] Barron, J. L. 2004. Experience with 3d optical flow on gated mri cardiac datasets. In: Canadian Conference on Computer and Robot Vision. pages 370–377.

[Barron and Thacker, 2005] Barron, J. L. and Thacker, N. 2005. Tutorial: Computing 2d and 3d optical flow. In: Tina Memo No. 2004-012.

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

[Bhatia et al., 2007] Bhatia, B., Oates, T., Xiao, Y., and Hu, P. 2007. 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.

[Blum, 2007] Blum, T. 2007. Surgical workflow analysis: Representation and application to monitoring. Diploma Thesis. Technische Universität München.

[Blum et al., 2008a] Blum, T., Padoy, N., Feussner, H., and Navab, N. 2008a. Modeling and online recognition of surgical phases using hidden markov models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), New York, USA. pages 627–635.

[Blum et al., 2008b] Blum, T., Padoy, N., Feussner, H., and Navab, N. 2008b. Workflow mining for visualization and analysis of surgeries. International Journal of Computer Assisted Radiology and Surgery, 3(5):379–386.

[Bobick, 1997] Bobick, A. 1997. Movement, activity, and action: The role of knowledge in the perception of motion. Proceedings of the Royal Society, 352:1257–1265.

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. 2001. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):257–267.

[Brand and Kettnaker, 2000] Brand, M. and Kettnaker, V. 2000. Discovery and segmentation of activities in video. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 22(8):844–851.

[Brand et al., 1997] Brand, M., Oliver, N., and Pentland, A. 1997. Coupled hidden markov models for complex action recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA. IEEE Computer Society, page 994.

[Bravo et al., 2008] Bravo, J., Hervás, R., Gallego, R., Casero, G., Vergara, M., Carmona, T., Fuentes, C., Gachet, D., Nava, S., Chavira, G., and Villarreal, V. 2008. Identification technologies to support alzheimer contexts. In: PETRA '08: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments, New York, NY, USA. ACM, pages 1–2.

[Brdiczka et al., 2007] Brdiczka, O., Crowley, J. L., and Reignier, P. 2007. Learning situation models for providing context-aware services. In: Human-Computer Interaction (6). pages 23–32.

[Bremond et al., 2006] Bremond, F., Thonnat, M., and Zuniga, M. 2006. Video understanding framework for automatic behavior recognition. Behavior Research Methods, 3(38):416–426.

[Bucher et al., 2009] Bucher, P., Pugin, F., Buchs, N., Ostermann, S., Charara, F., and Morel, P. 2009. Single port access laparoscopic cholecystectomy (with video). World J. Surg., 33(5):1015–1019.

[Burgert et al., 2007] Burgert, O., Neumuth, T., Gessat, M., Jacobs, S., and Lemke, H. U. 2007. Deriving dicom surgical extensions from surgical workflows. In: SPIE Medical Imaging 2007, PACS and Imaging Informatics. 6516:651604.

[Burgert et al., 2006] Burgert, O., Neumuth, T., Lempp, F., Mudunuri, R., Meixensberger, J., Strauss, G., Dietz, A., Jannin, P., and Lemke, H. U. 2006. Linking top-level ontologies and surgical workflows. In: International Journal of Computer Assisted Radiology and Surgery. 1(1):437–438.

[Cao et al., 1996] Cao, C. G. L., Mackenzie, C. L., and Payandeh, S. 1996. Task and motion analyses in endoscopic surgery. In: 5th Annual Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. pages 583–590.

[Chang et al., 2007] Chang, K.-H., Chen, M. Y., and Canny, J. 2007. Tracking free-weight exercises. In: International Conference on Ubiquitous Computing. pages 19–37.

[Chen et al., 2001] Chen, X., Barron, J. L., Mercer, R. E., and Joe, P. 2001. 3d regularized velocity from 3d doppler radial velocity. In: International Conference on Image Processing (ICIP). pages 664–667.

[Cleary et al., 2005] Cleary, K., Chung, H. Y., and Mun, S. K. 2005. Or 2020: The operating room of the future. Laparoendoscopic and Advanced Surgical Techniques, 15(5):495–500.

[Darrell et al., 1996] Darrell, T., Essa, I. A., and Pentland, A. 1996. Task-specific gesture analysis in real-time using interpolated views. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 18(12):1236–1242.

[Darzi et al., 1999] Darzi, A., Smith, S., and Taffinder, N. 1999. Assessing operative skill: needs to become more objective. British Medical Journal, 318(7188):887–888.

[Davies, 2000] Davies, B. 2000. A review of robotics in surgery. Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine, 214(1):129–140.

[Dubois, 1993] Dubois, S. 1993. Cholécystectomie et exploration de la voie biliaire principale par coelioscopie. Editions Techniques - Encycl. Med. Chir. (Paris, France), Techniques chirurgicales - Généralités-Appareil digestif, 40(950):1–17.

[Efros et al., 2003] Efros, A. A., Berg, A., Mori, G., and Malik, J. 2003. Recognizing action at a distance. In: IEEE International Conference on Computer Vision (ICCV). pages 726–733.

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

[Elgammal et al., 2003] Elgammal, A., Shet, V., Yacoob, Y., and Davis, L. S. 2003. Learning dynamics for exemplar-based gesture recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pages 571–578.

[Favela et al., 2007] Favela, J., Tentori, M., Castro, L. A., Gonzalez, V. M., Moran, E. B., and Martínez-García, A. I. 2007. Activity recognition for context-aware hospital applications: issues and opportunities for the deployment of pervasive networks. Mob. Netw. Appl., 12(2-3):155–171.

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

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

[Fishkin et al., 2004] Fishkin, K., Consolvo, S., Rode, J., Ross, B., Smith, I., , and Souter, K. 2004. Ubiquitous computing support for skills assessment in medical school. In: Ubihealth 2004: The 3rd Int'l Workshop Ubiquitous Computing for Pervasive Healthcare Applications.

[Freund and Schapire, 1995] Freund, Y. and Schapire, R. E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: EuroCOLT. pages 23–37.

[Friedman et al., 2007] Friedman, D. C. W., Dosher, J., Kowalewski, T., Rosen, J., and Hannaford, B. 2007. Automated tool handling for the trauma pod surgical robot. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). pages 1936–1941.

[Friedman et al., 1998] Friedman, N., Murphy, K., and Russell, S. 1998. Learning the structure of dynamic probabilistic networks. In: Uncertainty in Artificial Intelligence (UAI). Morgan Kaufmann, pages 139–147.

[Fukui et al., 2006] Fukui, S., Iwahori, Y., Itoh, H., Kawanaka, H., and R.Woodham 2006. Robust background subtraction for quick illumination changes. In: Proceedings of First Pacific-Rim Symposium on Image and Video Technology. pages 1244–1253.

[Ghahramani et al., 1997] Ghahramani, Z., Jordan, M. I., and Smyth, P. 1997. Factorial hidden markov models. In: Machine Learning. MIT Press.

[Ghodoussi et al., 2002] Ghodoussi, M., Butner, S. E., and Wang, Y. 2002. Robotic surgery - the transatlantic case. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). IEEE, pages 1882–1888.

[Grimson et al., 1998] Grimson, W. E. L., Stauffer, C., Romano, R., and Lee, L. 1998. Using adaptive tracking to classify and monitor activities in a site. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pages 22–29.

[Guthart and Jr., 2000] Guthart, G. and Jr., J. K. S. 2000. The intuitive$^{\text{tm}}$ telesurgery system: Overview and application. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). pages 618–621.

[Hamid et al., 2007] Hamid, R., Maddi, S., Bobick, A. F., and Essa, I. A. 2007. Structure from statistics: Unsupervised activity analysis using suffix trees. In: IEEE International Conference on Computer Vision (ICCV). pages 1–8.

[Hartley and Zisserman, 2004] Hartley, R. I. and Zisserman, A. 2004. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition.

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

[Horn and Schunck, 1980] Horn, B. K. and Schunck, B. G. 1980. Determining optical flow. Technical report, Cambridge, MA, USA.

[Houliston and Parry, 2008] Houliston, B. and Parry, D. 2008. Sensors and insensibility: Monitoring anaesthetic activity with rfid. In: Health Informatics New Zealand, Seventh Annual Conference and Exhibition.

[Hu et al., 2006] Hu, P. F., Xiao, Y., Ho, D., Mackenzie, C. F., Hu, H., Voigt, R., and Martz, D. 2006. Advanced visualization platform for surgical operating room coordination: distributed video board system. Surg Innov, 13(2):129–135.

[Huynh et al., 2007] Huynh, T., Blanke, U., and Schiele, B. 2007. Scalable recognition of daily activities with wearable sensors. In: International Symposium on Location and Context Awareness. pages 50–67.

[Ikizler and Forsyth, 2008] Ikizler, N. and Forsyth, D. A. 2008. Searching for complex human activities with no visual examples. International Journal of Computer Vision, 80(3):337–357.

[Jacques et al., 2007] Jacques, M., Bernard, D., Silvana, P., Arnaud, W., Didier, M., Dimitri, C., and G., H. J. 2007. Surgery without scars : Report of transluminal cholecystectomy in a human being. Archives of surgery, 142(9):823–827.

[James et al., 2007] James, A., Vieira, D., Lo, B. P. L., Darzi, A., and Yang, G.-Z. 2007. Eye-gaze driven surgical workflow segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 110–117.

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

[Jannin et al., 2001] Jannin, P., Raimbault, M., Morandi, X., and Gibaud, B. 2001. Modeling surgical procedures for multimodal image-guided neurosurgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 565–572.

[Jenkins and Matarić, 2004] Jenkins, O. C. and Matarić, M. J. 2004. A spatio-temporal extension to isomap nonlinear dimension reduction. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, New York, NY, USA. ACM, page 56.

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

[Kassidas et al., 1998] Kassidas, A., MacGregor, J. F., and Taylor, P. A. 1998. Synchronization of batch trajectories using dynamic time warping. AIChE Journal, 44(4):864–875.

[Keogh and Lin, 2005] Keogh, E. and Lin, J. 2005. Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and Information Systems, 8(2):154–177.

[Keogh and Ratanamahatana, 2005] Keogh, E. and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. Knowledge and Information Systems, 7(3):358–386.

[Kern et al., 2003] Kern, N., Schiele, B., and Schmidt, A. 2003. Multi-sensor activity context detection for wearable computing. First European Symposium on Ambient Intelligence, pages 220–232.

[Kidd et al., 1999] Kidd, C. D., Orr, R., Abowd, G. D., Atkeson, C. G., Essa, I. A., MacIntyre, B., Mynatt, E. D., Starner, T., and Newstetter, W. 1999. The aware home: A living laboratory for ubiquitous computing research. In: CoBuild '99: Proceedings of the Second International Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture, London, UK. Springer-Verlag, pages 191–198.

[Kläser et al., 2008] Kläser, A., Marszałek, M., and Schmid, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In: British Machine Vision Conference (BMVC).

[Klein and Huesman, 1997] Klein, G. J. and Huesman, R. H. 1997. A 3d optical flow approach to addition of deformable pet volumes. In: NAM '97: Proceedings of the 1997 IEEE Workshop on Motion of Non-Rigid and Articulated Objects (NAM '97), Washington, DC, USA. IEEE Computer Society, page 136.

[Ko et al., 2007] Ko, S.-Y., Kim, J., Lee, W.-J., and Kwon, D.-S. 2007. Surgery task model for intelligent interaction between surgeon and laparoscopic assistant robot. International Journal of Assitive Robotics and Mechatronics, 8(1):38–46.

[Koile et al., 2003] Koile, K., Tollmar, K., Demirdjian, D., Shrobe, H., and Darrell, T. 2003. Activity zones for context-aware computing. In: International Conference on Ubiquitous Computing. Springer-Verlag, pages 90–106.

[Krause et al., 2003] Krause, A., Siewiorek, D. P., Smailagic, A., and Farringdon, J. 2003. Unsupervised, dynamic identification of physiological and activity context in wearable computing. In: ISWC '03: Proceedings of the 7th IEEE International Symposium on Wearable Computers, Washington, DC, USA. IEEE Computer Society.

[Ladikos et al., 2008a] Ladikos, A., Benhimane, S., and Navab, N. 2008a. Efficient visual hull computation for real-time 3d reconstruction using cuda. In: Proceedings of the 2008 Conference on Computer Vision and Pattern Recognition, Workshop on Visual Computer Vision on GPUs (CVGPU).

[Ladikos et al., 2008b] Ladikos, A., Benhimane, S., and Navab, N. 2008b. Real-time 3d reconstruction for collision avoidance in interventional environments. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 526–534.

[Laurentini, 1994] Laurentini, A. 1994. The visual hull concept for silhouette-based image understanding. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 16(2):150–162.

[Lemke, 2007] Lemke, H. U. 2007. Summary of the white paper of dicom wg24 'dicom in surgery'. In: SPIE Medical Imaging 2007 - PACS and Imaging Informatics, Progress in Biomedical Optics and Imaging. 6516.

[Lemke and Berliner, 2007] Lemke, H. U. and Berliner, L. 2007. Specification and design of a therapy imaging and model management system (timms). In: SPIE Medical Imaging 2007 - PACS and Imaging Informatics, Progress in Biomedical Optics and Imaging. 6516:651602.

[Lemke et al., 2005] Lemke, H. U., Ratib, O. M., and Horii, S. C. 2005. Workflow in the operating room: review of Arrowhead 2004 seminar on imaging and informatics (Invited Paper). In: Ratib, O. M. and Horii, S. C. (Editors), Medical Imaging 2005: PACS and Imaging Informatics. Edited by Ratib, Osman M.; Horii, Steven C. Proceedings of the SPIE, Volume 5748, pp. 83-96 (2005). 5748:83–96.

[Lemke and Vannier, 2006] Lemke, H. U. and Vannier, M. W. 2006. The operating room and the need for an it infrastructure and standards. International Journal of Computer Assisted Radiology and Surgery, 1(3):117–121.

[Leong et al., 2007] Leong, J., Nicolaou, M., Atallah, L., Mylonas, G., Darzi, A., and Yang, G.-Z. 2007. HMM Assessment of Quality of Movement Trajectory in Laparoscopic Surgery. Computer Aided Surgery, 12(6):335–346.

[Lester et al., 2006] Lester, J., Choudhury, T., and Borriello, G. 2006. A Practical Approach to Recognizing Physical Activities.

[Lin et al., 2006] Lin, H. C., Shafran, I., Yuh, D., and Hager, G. D. 2006. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery, 11(5):220–230.

[Lo et al., 2003] Lo, B. P. L., Darzi, A., and Yang, G.-Z. 2003. 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.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In: International Joint Conferences on Artificial Intelligence. pages 674–679.

[Lv and Nevatia, 2007] Lv, F. and Nevatia, R. 2007. Single view human action recognition using key pose matching and viterbi path searching. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[Macario et al., 2006] Macario, A., Morris, D., and Morris, S. 2006. Initial Clinical Evaluation of a Handheld Device for Detecting Retained Surgical Gauze Sponges Using Radiofrequency Identification Technology. Arch Surg, 141(7):659–662.

[MacKenzie et al., 2001] MacKenzie, C. L., Ibbotson, J. A., Cao, C. G. L., and Lomax, A. J. 2001. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minimally Invasive Therapy Allied Technologies, 10(3):121:128.

[Maruster et al., 2001] Maruster, L., van der Aalst, W., Weijters, T., van den Bosch, A., and Daelemans, W. 2001. Automatic discovery of workflow models from hospital data. In: Proceedings BNAIC-01. pages 183–194.

[Mayoral et al., 2008] Mayoral, R., Vazquez, A., and Burgert, O. 2008. A general framework for data streaming in the digital operating room. In: Horii, Steven C.; Andriole, K. P. (Editor), SPIE Medical Imaging:PACS and Imaging Informatics. 6919:691933.

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

[Mehta et al., 2001] Mehta, N. Y., Haluck, R. S., Frecker, M. I., and Snyder, A. J. 2001. Sequence and task analysis of instrument use in common laparoscopic procedures. Surgical Endoscopy, 16(2):280–285.

[Meyer et al., 2007] Meyer, M. A., Levine, W. C., Egan, M. T., Cohen, B. J., Spitz, G., Garcia, P., Chueh, H., and Sandberg, W. S. 2007. A computerized perioperative data integration and display system. International Journal of Computer Assisted Radiology and Surgery, 2(3-4):191–202.

[Mihajlovic and Petkovic, 2001] Mihajlovic, V. and Petkovic, M. 2001. Dynamic bayesian networks: A state of the art. CTIT technical reports series, TR-CTIT-34. DMW-project.

[Minnen et al., 2007] Minnen, D., Jr., C. L. I., Essa, I. A., and Starner, T. 2007. Discovering multivariate motifs using subsequence density estimation and greedy mixture learning. In: National Conference on Artificial Intelligence (AAAI). pages 615–620.

[Minnen et al., 2006] Minnen, D., Starner, T., Essa, I. A., and Jr., C. L. I. 2006. Discovering characteristic actions from on-body sensor data. In: Tenth IEEE International Symposium on Wearable Computers (ISWC). pages 11–18.

[Miyawaki et al., 2005] Miyawaki, F., Masamune, K., Suzuki, S., Yoshimitsu, K., and Vain, J. 2005. 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.

[Moore and Essa, 2002] Moore, D. and Essa, I. 2002. Recognizing multitasked activities from video using stochastic context-free grammar. In: National Conference on Artificial Intelligence (AAAI). pages 770–776.

[Moore et al., 1999] Moore, D. J., Essa, I. A., and Hayes, M. H., I. 1999. Exploiting human actions and object context for recognition tasks. In: IEEE International Conference on Computer Vision (ICCV). 1:80–86.

[Morgenstern, 1992] Morgenstern, L. 1992. Carl langenbuch and the first cholecystectomy. Surgical Endoscopy, 6(3):113–114.

[Morris and Paradiso, 2002] Morris, S. J. and Paradiso, J. A. 2002. Shoe-integrated sensor system for wireless gait analysis and real-time feedback. In: Proceedings of the Second Joint EMBS/BMES Conference. pages 2468–2469.

[Murphy, 2002] Murphy, K. P. 2002. Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, UC Berkeley, Computer Science Division.

[Nagy et al., 2006] Nagy, P., George, I., Bernstein, W., Caban, J., Klein, R., Mezrich, R., and Park, A. 2006. Radio frequency identification systems technology in the surgical setting. Surg Innov, 13(1):61–67.

[Navab et al., 2007] Navab, N., Traub, J., Sielhorst, T., Feuerstein, M., and Bichlmeier, C. 2007. Action- and workflow-driven augmented reality for computer-aided medical procedures. IEEE Computer Graphics and Applications, 27(5):10–14.

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

[Neumuth et al., 2008] Neumuth, T., Mansmann, S., Scholl, M. H., and Burgert, O. 2008. Data warehousing technology for surgical workflow analysis. In: IEEE Symposium on Computer-Based Medical Systems (CBMS). pages 230–235.

[Neumuth et al., 2006b] Neumuth, T., Schumann, S., Strauss, G., Jannin, P., Meixensberger, J., Dietz, A., Lemke, H. U., and Burgert, O. 2006b. Visualization options for surgical workflows. International Journal of Computer Assisted Radiology and Surgery, 1(1):438–440.

[Neumuth et al., 2006c] Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H. U., and Burgert, O. 2006c. Acquisition of process descriptions from surgical interventions. In: International Conference on Database and Expert Systems Applications (DEXA). pages 602–611.

[Nguyen et al., 2005] Nguyen, N., Phung, D., Venkatesh, S., and Bui, H. H. 2005. Learning and detecting activities from movement trajectories using the hierarchical HMMs. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pages 955–960.

[Oliver et al., 2004] Oliver, N., Garg, A., and Horvitz, E. 2004. Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision Image Understanding Journal (CVIU), 96(2):163–180.

[Olsen, 2006] Olsen, D. O. 2006. Historical overview and indications for cholecystectomy. Laparoscopic Surgery of the Abdomen, pages 71–72.

[Padoy et al., 2007a] Padoy, N., Blum, T., Essa, I., Feussner, H., Berger, M.-O., and Navab, N. 2007a. A boosted segmentation method for surgical workflow analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pages 102–109.

[Padoy et al., 2008] Padoy, N., Blum, T., Feussner, H., Berger, M.-O., and Navab, N. 2008. 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.

[Padoy et al., 2007b] Padoy, N., Horn, M., Feussner, H., Berger, M.-O., and Navab, N. 2007b. Recovery of surgical workflow: a model-based approach. In: International Journal of Computer Assisted Radiology and Surgery (CARS), Supplement 1. 2:481–482.

[Padoy et al., 2009] Padoy, N., Mateus, D., Weinland, D., Berger, M.-O., and Navab, N. 2009. Workflow monitoring based on 3d motion features. In: Proceedings of the International Conference on Computer Vision (ICCV), Workshop on Video-oriented Object and Event Classification (To Appear).

[Patel et al., 2007] Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J. H., Welsh, M., and Bonato, P. 2007. Analysis of feature space for monitoring persons

with parkinson's disease with application to a wireless wearable sensor system. In: Proc. 29th IEEE EMBS Annual International Conference.

[Paulus et al., 1998] Paulus, D., Csink, L., and Niemann, H. 1998. Color cluster rotation. In: International Conference on Image Processing (ICIP). pages 161–165.

[Peters, 2006] Peters, T. 2006. Image-guidance for surgical procedures. Physics in Medicine and Biology, 51(14):R505–R540.

[Pinhanez and Bobick, 1997] Pinhanez, C. and Bobick, A. 1997. Intelligent studios modeling space and action to control tv cameras. Applied Artificial Intelligence, 11(4):285–305.

[Predey et al., 2002] Predey, T., Sewall, L., and Smith, S. 2002. Percutaneous vertebroplasty: New treatment for vertebral compression fractures. American Family Physician, 66(4):611–615.

[Qi et al., 2006] Qi, J., Jiang, Z., Zhang, G., Miao, R., and Su, Q. 2006. A surgical management information system driven by workflow. In: IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). pages 1014–1018.

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

[Radrich, 2008] Radrich, H. 2008. Vision-based motion monitoring through data fusion from a chirurgical multi-camera recording system. Diploma Thesis. Technische Universität München.

[Riskin et al., 2006] Riskin, D. J., Longaker, M. T., Gertner, M., and Krummel, T. M. 2006. Innovation in surgery, a historical perspective. Annals of Surgery, 244(5):686–693.

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

[Rosen et al., 2006] Rosen, J., Brown, J., Chang, L., Sinanan, M., and Hannaford, B. 2006. 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.

[Rybachuk, 2009] Rybachuk, K. 2009. Motif discovery and refinement for detection and labeling of human activity. Master Thesis. Technische Universität München.

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

[Sanchez et al., 2008] Sanchez, D., Tentori, M., and Favela, J. 2008. Activity recognition for the smart hospital. IEEE Intelligent Systems, 23(2):50–57.

[Sandberg et al., 2005] Sandberg, W. S., Daily, B., Egan, M., Stahl, J. E., Goldman, J. M., Wiklund, R. A., and Rattner, D. 2005. Deliberate perioperative systems design improves operating room throughput. Anesthesiology, 103(2):406–418.

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

[Satava, 2005] Satava, R. M. 2005. Telesurgery, robotics, and the future of telemedicine. European Surgery, 37(5):304–307.

[Seitz et al., 2006] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA. IEEE Computer Society, pages 519–528.

[Shi et al., 2006] Shi, Y., Bobick, A. F., and Essa, I. A. 2006. Learning temporal sequence model from partially labeled data. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pages 1631–1638.

[Shi et al., 2004] Shi, Y., Huang, Y., Minnen, D., Bobick, A., and Essa, I. 2004. Propagation networks for recognition of partially ordered sequential action. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pages 862–869.

[Siddiqi et al., 2007] Siddiqi, S., Gordon, G. J., and Moore, A. W. 2007. Fast state discovery for hmm model selection and learning. In: In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AI-STATS).

[Sielhorst et al., 2005] Sielhorst, T., Blum, T., and Navab, N. 2005. Synchronizing 3d movements for quantitative comparison and simultaneous visualization of actions. In: Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05), Vienna, Austria. pages 38–47.

[Sielhorst et al., 2008] Sielhorst, T., Feuerstein, M., and Navab, N. 2008. Advanced medical displays: A literature review of augmented reality. IEEE/OSA Journal of Display Technology; Special Issue on Medical Displays, 4(4):451–467.

[Simon Baker and Matthews, 2004] Simon Baker, Raju Patil, K. M. C. and Matthews, I. 2004. Lucas-kanade 20 years on: Part 5. Technical Report CMU-RI-TR-04-64, Robotics Institute, Pittsburgh, PA.

[Singer and Ostendorf, 1996] Singer, H. and Ostendorf, M. 1996. Maximum likelihood successive state splitting. In: ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference, Washington, DC, USA. IEEE Computer Society, pages 601–604.

[Smith et al., 2005] Smith, J. R., Fishkin, K. P., Jiang, B., Mamishev, A., Philipose, M., Rea, A. D., Roy, S., and Sundara-Rajan, K. 2005. Rfid-based techniques for human-activity detection. Commun. ACM, 48(9):39–44.

[Speidel et al., 2008] Speidel, S., Sudra, G., Senemaud, J., Drentschew, M., MÃijller-Stich, B. P., Gutt, C., and Dillmann, R. 2008. Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling. In: Med. Imaging. SPIE.

[Stephenson, 2000] Stephenson, T. A. 2000. An introduction to Bayesian network theory and usage. IDIAP-RR 03, IDIAP.

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

[Subramanya et al., 2006] Subramanya, A., Raj, A., Bilmes, J. A., and Fox, D. 2006. Recognizing activities and spatial context using wearable sensors. In: Uncertainty in Artificial Intelligence (UAI).

[Svoboda et al., 2005] Svoboda, T., Martinec, D., and Pajdla, T. 2005. A convenient multicamera self-calibration for virtual environments. Presence: Teleoper. Virtual Environ., 14(4):407–422.

[Szeliski, 1993] Szeliski, R. 1993. Rapid octree construction from image sequences. CVGIP: Image Understanding, 58(1):23–32.

[Tahar et al., 2008] Tahar, K., Korb, W., Burgert, O., Möckel, H., and Neumuth, T. 2008. Ein system zur sensorbasierten erfassung chirurgischer instrumentennutzung für die unterstützung von workflowanalysen. In: 7. Jahrestagung der Deutschen Gesellschaft für Computer-und Roboter-Assistierte Chirurgie (CURAC 2008), Leipzig, Germany.

[Taylor, 2008] Taylor, R. H. 2008. Medical robotics and computer-integrated surgery. Computer Software and Applications Conference, Annual International, 0:1.

[van der Aalst et al., 2003] van der Aalst, W. M. P., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., and Weijters, A. J. M. M. 2003. Workflow mining: a survey of issues and approaches. Data Knowl. Eng., 47(2):237–267.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. 2004. Robust real-time face detection. International Journal of Computer Vision (IJCV), 57(2):137–154.

[Vu et al., 2003] Vu, V.-T., Brémond, F., and Thonnat, M. 2003. Automatic video interpretation: A novel algorithm for temporal scenario recognition. In: International Joint Conferences on Artificial Intelligence. pages 1295–1302.

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

[Wang et al., 2007] Wang, X., Ma, X., and Grimson, E. 2007. Unsupervised activity perception by hierarchical bayesian models. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pages 1–8.

[Ward et al., 2006] Ward, J., Lukowicz, P., Tröster, G., and Starner, T. 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans. Pattern Analysis and Machine Intelligence, 28(10):1553–1567.

[Weinland, 2008] Weinland, D. 2008. Action Representation and Recognition. PhD thesis, Institut National Polytechnique de Grenoble.

[Weinland et al., 2007] Weinland, D., Boyer, E., and Ronfard, R. 2007. Action recognition from arbitrary views using 3d exemplars. In: IEEE International Conference on Computer Vision (ICCV). pages 1–7.

[Weinland et al., 2006] Weinland, D., Ronfard, R., and Boyer, E. 2006. Free viewpoint action recognition using motion history volumes. Computer Vision Image Understanding Journal (CVIU), 104(2-3):249–257.

[Won et al., 2006] Won, K.-J., Prugel-Bennett, A., and Krogh, A. 2006. Evolving the structure of hidden markov models. IEEE Transactions on Evolutionary Computation, 10(1):39–49.

[Wyatt, 2005] Wyatt, D. 2005. Unsupervised activity recognition using automatically mined common sense. In: National Conference on Artificial Intelligence (AAAI). pages 21–27.

[Xiang and Gong, 2008] Xiang, T. and Gong, S. 2008. Optimising dynamic graphical models for video content analysis. Computer Vision Image Understanding Journal (CVIU), 112(3):310–323.

[Xiao et al., 2005] Xiao, Y., Hu, P., Hu, H., Ho, D., Dexter, F., Mackenzie, C. F., Seagull, F. J., and Dutton, R. P. 2005. An algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time. Anesth Analg, 101:823–829.

[Xie et al., 2003] Xie, L., Chang, S.-F., Divakaran, A., and Sun, H. 2003. Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models. In: ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03), Washington, DC, USA. IEEE Computer Society, pages 29–32.

[Xie et al., 2004] Xie, L., Xu, P., Chang, S.-F., Divakaran, A., and Sun, H. 2004. Structure analysis of soccer video with domain knowledge and hidden markov models. Pattern Recognition Letters, 25(7):767–775.

[Yan et al., 2008] Yan, P., Khan, S. M., and Shah, M. 2008. Learning 4d action feature models for arbitrary view action recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[Yang, 2006] Yang, G. Z. 2006. Body Sensor Networks. Springer.

[Yaniv and Burshtein, 2003] Yaniv, R. and Burshtein, D. 2003. An enhanced dynamic time warping model for improved estimation of dtw parameters. IEEE Transactions on Speech and Audio Processing, 11(3):216–228.

[Yaniv and Cleary, 2006] Yaniv, Z. and Cleary, K. 2006. Image-guided procedures: A review. Technical report, Imaging Science and Information Systems, Georgetown Univ.

[Yoshimitsu et al., 2007] Yoshimitsu, K., Miyawaki, F., Sadahiro, T., Ohnuma, K., Fukui, Y., Hashimoto, D., and Masamune, K. 2007. 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.

[Zhang et al., 2008] Zhang, G. G., Huang, T.-C., Guerrero, T., Lin, K.-P., Stevens, C., Starkschall, G., and Forster, K. 2008. The use of 3d optical flow method in mapping of 3d anatomical structure and tumor contours across 4d ct data. Journal of Applied Clinical Medical Physics, 9(1):59–69.

[Zornig et al., 2008] Zornig, C., Mofid, H., Emmermann, A., Alm, M., von Waldenfels, H.-A., and Felixmüller, C. 2008. Scarless cholecystectomy with combined transvaginal and transumbilical approach in a series of 20 patients. Surgical Endoscopy, 22(6):1427–1429.