Knowledge-based support for surgical workflow analysis and recognition

Olga Dergachyova

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du signal et télécommunications

École doctorale MathSTIC

présentée par

Olga DERGACHYOVA

Préparée à l'unité de recherche U1099 LTSI Inserm
Modélisation des compétences chirurgicales et interventionnelles
Informatique Electronique (ISTIC)

Knowledge-based support for surgical workflow analysis and recognition

Soutenue à Rennes
le 28 novembre 2017

devant le jury composé de :

Arya NABAVI
PU-PH, INI, Hanovre, Allemagne / président

Sandrine VOROS
CR Inserm, Université Grenoble Alpes / rapporteur

Paolo FIORINI
PU, University of Verona, Italie / rapporteur

Luc SOLER
PU, IRCAD, Strasbourg / examinateur

Xavier MORANDI
PU-PH, CHU Rennes, MediCIS / directeur de thèse

Pierre JANNIN
DR Inserm, MediCIS / co-directeur de thèse

Посвящается папе, бабушке Оле, тёте Марине и Кеше - любимым, которых я потеряла в этом нелёгком пути...

Благодарности

Первым делом я хочу поблагодарить многоуважаемых членов диссертационной комиссии за предоставленную мне возможность и честь: профессора Арью Набави, доктора Сандрин Ворос, профессора Паоло Фиорини и профессора Люка Солера. Также выражаю благодарность своим научным руководителям - профессору Ксавье Моранди и доктору Пьеру Жаннану. Отдельное спасибо Пьеру, который в меня поверил и продолжал верить на протяжении всего пути, даже когда я сама эту веру теряла. У тебя всегда получалось найти правильные слова, подбодрить и дать мотивацию в нужный момент. Спасибо за отеческую заботу и превосходное руководство.

Хотелось бы отдать дань двум годам магистратуры проведенным в Страсбургском университете, которые открыли двери в мир исследований. Я никогда не забуду моих одногруппников, в частности Криса, Макса, Жака и Арно, которые оказали огромную помощь в учебе и были незаменимыми партнерами по веселью. В моей памяти также останутся мои университетские профессора и коллеги из группы IGG: Паскаль Шрек, Сильван Тери, Пьер Кремер, Паскаль Матис, Базиль Соваж, Василий Михальчук, Александр Урстель, Лионель Антеренер и другие. Отдельное спасибо хочется сказать Фредерику Ляру, Кеннету Ванхую, Нуре Хамзе и Пьеру Бутри, с которыми у меня завязались теплые дружеские отношения. Хотелось бы отдельно поблагодарить руководительницу моего исследовательского стажа Каролин Эссерт. Именно Каролин пробудила во мне желание продолжить обучение. Спасибо, что однажды дала мне шанс!

Разумеется, я не могу не отметить моих коллег из университета Ренна и группы MediCIS. Огромное спасибо Давиду, который передал мне ценные знания и опыт. Спасибо за понимание и нескончаемое терпение (не сосчитать количество раз, когда я бегала за помощью в его кабинет). Спасибо моему коллеге и другу Хавьеру. Несмотря на наши небольшие столкновения в начале, я очень рада, что в нашу ко-

манду пришел такой человек как ты. Я всегда обожала наши разговоры на самые различные темы и знала, что могу тебе доверять. Спасибо также нашему секретарю и моей подруге Ноэми за помощь по всем административным вопросам, а также кучу смеха и приколов. Думаю все будут еще долго помнить наше "sinon soumission!". Спасибо Себастьяну, появившемуся в нужный момент, за интересные разговоры и мотивацию. Также благодарю моих коллег Фреда, Юлонга, Дарко, Пьер-Луи, Клемана, Арно, Мари-Стефани, Шанталь и остальных за отличную атмосферу и море позитива. Я буду скучать по нашим совместным обедам, играм, посиделкам в барах и, да-да, даже по картингу (только не смеяться!).

Конечно же хочется поблагодарить моих самых близких друзей: Женечку, Сашку "блинчика-пыхту"и Танюшку. Мне с вами нереально повезло! Спасибо за вашу преданность и безмерную поддержку. Я по вам очень скучаю. Спасибо также Физкультурникам, в особенности Даниилу и Айдосу, которые привили мне любовь к спорту и путешествиям, которые, в свою очередь, помогли бороться со стрессом в течении этого трудного периода. Спасибо Роме за возможность поехать во Францию в самый первый раз. Спасибо моим страсбургским подругам Ане, Малике, Лере и Зарине. Большое спасибо Инессе и Марку, с которыми я познакомилась во время летней школы на Сицилии. Также спасибо всем тем, кто повлиял на меня и сделал меня такой, какая я есть, что позволило мне перенести все трудности и реализовать данную работу.

Огромное спасибо хочется сказать моей французской семье, которая меня так тепло приняла: бабушке Сюзанне, дедушке Рафаэлю, Франсуаз и Рафаэлю, Сюзанн, Сабрине, Янику, Александру, Ниле и Эльосу. Конечно же, отдельное спасибо Селии и Жану за всю ту неоценимую поддержку и помощь!

Я безусловно не могу не поблагодарить моего дорогого Фабьяна, который, несмотря ни на что, навсегда останется в моём сердце. Огромное спасибо за эти годы и всё то прекрасное, что между нами было. Спасибо за нашего любимого масеньку Орвелла (Люлюшку). Ты столько для меня сделал, ты всегда был рядом, и особенно в самые трудные моменты. Я бы не смогла закончить эту работу без тебя и твоей поддержки!

Я невыразимо благодарна своей собственной семье, всем родным и близким за моральную поддержку и неустанную веру в мой успех. Особое спасибо моим бабушкам и дедушкам, крестной Людмиле, сестренке Виктории, ее мужу Ринату, моей сладкой племяшке Никушке и моим котикам Кешке и Муське.

Я никогда не смогу выразить словами благодарность моим родителям. Мамочка и папочка, спасибо вам за всё, что вы для меня сделали! Спасибо за замечательное воспитание, бесконечную помощь, неоценимую поддержку, многочисленные жертвы, великую гордость за мои успехи, море ласки и необъятную любовь. Без вас я бы ничего не добилась. Я так вас люблю!

Последние слова хочется посвятить лучику света, недавно озарившему мою жизнь. Спасибо, что ты есть...

Acknowledgements

First of all, I would like to thank my thesis committee: Pr. Arya Nabavi, Dr. Sandrine Voros, Pr. Paolo Fiorini, and Pr. Luc Soler for the honor and opportunity to defend my work. I would also like to express gratitude to Pr. Xavier Morandi and Dr. Pierre Jannin, my thesis directors. Special thanks to Pierre who believed in me and continued to believe all the way, even when I was losing faith myself. You always managed to find the right words, cheer and motivate me when I needed it. Thank you for fatherly care and excellent supervision.

I would like to praise two years of my Master's program spent at the University of Strasbourg, which created a gateway to the world of research. I will never forget my classmates, in particular Christophe, Maxime, Jacques and Arnaud, who helped me a lot in my studies and were irreplaceable partners for fun. I will also remember my university professors and colleagues from the IGG team: Pascal Schreck, Sylvain Thery, Pierre Kraemer, Pascal Mathis, Basile Sauvage, Vasyl Mykhalchuk, Alexandre Hurstel, Lionel Untereiner and others. I would like to say a special thank you to Frédéric Larue, Kenneth Vanhoey, Noura Hamze and Pierre Boutry with whom I developed warm friendship. I would like to separately thank my internship supervisor Caroline Essert. It was Caroline who aroused in me the desire to continue my studies. Thank you for giving me a chance!

Of course, I can not fail to mention my colleagues from the University of Rennes and the MediCIS team. Many thanks to David who transmitted me his valuable knowledge and experience. Thank you for your understanding and endless patience (it is hard to count the number of times I ran for help in his office). I wish to thank my colleague and friend Javier. Despite our small clashes at the beginning, I am very glad that a person like you came to our team. I always adored our conversations on various topics and knew that I could trust you. Thanks also to our administrative assistant and my friend Noémie for her help and a bunch of laughs and jokes. I think everyone will remember our "sinon soumission!" for a long time. Thanks to Sebastien, who appeared at the right time, for interesting conversations and motivation. I also thank my colleagues Frédéric, Yulong, Darko, Pierre-Louis, Clément, Arnaud, Marie-Stéphanie, Chantal and others for the excellent atmosphere and

many positive vibes. I will miss our joint lunches, games, gatherings in bars and even (do not laugh!) karting.

I want to thank my closest friends Zhenechka, Sashka “pancake”, and Tanyushka. I am unbelievably lucky to have you! Thank you for your loyalty and immense support. I really miss you. Great thanks to “Physculturniki”, especially to Daniil and Aidos, who have instilled in me a love for sport and travel, which in turn helped to deal with stress during this difficult period. Thanks to Roma for giving the opportunity to go to France for the very first time. Thanks to my Strasbourg’s friends Anna, Malika, Lera, and Zarina. Many thanks to Inessa and Marc whom I met during the summer school in Sicily. Also thanks to all those who influenced me and made me what I am, which allowed me to overcome all the difficulties and perform this work.

Many thanks to my French family who received me so warmly: grandma Suzanna, grandpa Raphaël, Françoise and Raphaël, Suzanne, Sabrina, Yanick, Alexandre, Nyla, and Hélios. Of course, very special thanks to Célia and Jean for all the invaluable support and help!

I certainly can not thank enough my dear Fabien who despite everything will remain in my heart forever. Thank you so much for these years and all the wonderful things we had. Thank you for our beloved little Orwell (Lyulyushka). You have done so much, you have always been next to me and especially at the most difficult moments. I could not finish this work without you and your support!

I greatly thank my own family and all my relatives for the moral support and tireless faith in my success. Special thanks to my grandmothers and grandfathers, my aunt Lyudmila, my sister Viktoria, my brother-in-law Rinat, my sweetest little niece Nikushka, and my darling cats Keshka and Muska.

I will never be able to express gratitude to my parents with words. Mom and Dad, thank you for everything you have done for me! Thank you for the wonderful upbringing, endless help, invaluable support, numerous sacrifices, great pride in my successes, plenty of tenderness and infinite love. Without you I would not have achieved anything. I love you so much!

I want to devote the last words to the ray of light that recently lit up my life. Thank you for being a part of it...

Publications

International journal

O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, P. Jannin. *Automatic data-driven real-time segmentation and recognition of surgical workflow.* International Journal of Computer Assisted Radiology and Surgery, vol. 11(6), pages 1081-1089, 2016.

International conference

O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, P. Jannin. *Automatic data-driven real-time segmentation and recognition of surgical workflow.* International Conference on Information Processing in Computer-Assisted Interventions, Heidelberg, 21-22 June 2016.
The paper won the "Bench to Bedside" award.

National conference

O. Dergachyova, X. Morandi, P. Jannin. *Sensors evaluation for low-level surgical activity recognition.* Surgertica, Strasbourg, 20-22 November 2017.

ArXiv e-Prints

O. Dergachyova, X. Morandi, P. Jannin. *Knowledge transfer for surgical activity prediction.* arXiv: 1711.05848v1 [cs.LG], 15 November 2017.

O. Dergachyova, X. Morandi, P. Jannin. *Analyzing before solving: which parameters influence low-level surgical activity recognition.* arXiv: 1711.06259v1 [cs.HC], 15 November 2017.

Challenge

O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, P. Jannin. *Data-Driven Surgical Workflow Detection: Technical Report for M2CAI 2016 Surgical Workflow Challenge.* M2CAI, Medical Image Computing and Computer-Assisted Intervention, Athens, 21 October 2016. **The proposed method took third place.**

Software

D. Bouget, O. Dergachyova, F. Despinoy, P. Jannin. *ADCAS Framework: Automatic Detection and Classification of Surgeries.* APP (Agence pour la Protection des Programmes) deposit, 2016.

Contents

List of Figures

List of Tables

Part I

Introduction and related work

CHAPTER

1 Introduction

Preamble

The face of surgery has greatly changed through the centuries. A significant breakthrough happened during the industrial revolution largely extending previously limited surgical practices. People learnt to overcome three main obstacles: pain, bleeding and infection. The invention of anaesthesia and antiseptics encouraged more surgeries provoking a further progress. Notable advances made in 20th century brought new technologies. Minimalist environments with basic instruments were replaced by much more complex operating suites putting into practice new imaging devices, laparoscopic, robotic and computer-assisted surgery. Today, surgical field evolves with an unprecedented speed offering a tremendous amount of modern technologies and operative techniques that open doors to a totally new epoch.

Contents

1.1 Computer-assisted interventions

Computer-assisted surgery (CAS), otherwise called computer-aided surgery, emerged in 20th century, represents a concept in which computer technology is used to guide and assist surgeons during the entire surgical process. Computers contribute to the surgical process since the earliest stages of learning and training. Rapid development of surgical skills and their practice are enabled via various simulators and trainers [Satava 2001, Gallagher 2005]. Computers are also involved in automatic evaluation of gained skills [Reiley 2011]. In the pre-operative period, decision-making support and planning tools are provided [Garg 2005]. Intra-operative computer assistance includes but is not limited to robotic surgical systems [Lanfranco 2004], image guidance and navigation [Peters 2006], augmented reality and visualization [Shuhaiber 2004, Kersten-Oertel 2013]. Post-operative assistance offers tools for analysis of performed procedures and outcomes [Schumann 2015], as well as for their improvement and optimization [Hübler 2014].

A successful integration and cooperative functioning of multiple systems and devices are essential for enhancement of surgical procedures. Unfortunately, despite all the advances and valuable assistance, a seamless integration of computer aids into the operating room (OR) and surgical process has not yet been achieved. Existent ORs contain a stack of unrelated independent systems and devices mostly present in an isolated form disabling proper communication and interaction. Present computer-aids facilitate some individual surgical tasks but the absence of their synchronization with with surgical process impedes the work of the surgical team and the process of resource management. It results in higher levels of stress [Arora 2010], in frequent misunderstandings between the surgical staff members causing risks and delays, as well as in low efficiency of surgical suites that generates excessive expenses for the hospital [Macario 2010].

1.2 Operating room of the future

At the beginning of the new millennium, a concept of intelligent OR called operating room of the future, was proposed to overcome the problems mentioned above [Cleary 2005, Feussner 2003, Satava 2003]. In this vision of ultra-modern OR, surgery is motivated by safety and efficiency [Bharathan 2013]. New technologies deeply integrated in the operating theatre and synergistically working together are synchronized with the surgical procedure. Concordance in functioning of all systems simplifies surgical process; procedures take less time making them safer for the patient and less expensive for the hospital. Through thorough organization, the surgical non-technical skills as

Figure 1.1: Operating room of the future
Source: http://neurocirugiaferrer.com

team work and communication are also improved, which contributes to a better technical performance [Hull 2012]. Computer assistance becomes surgeon-, patient- and procedure-specific. Yet, realization of such an OR of the future requires a shift of existing paradigms, design of new infrastructures, information processing models and standards [Lemke 2006, Maier-Hein 2017]. Multiple research projects of advanced ORs were initiated in several countries, for example OR.NET, SCOT, and CONDOR described in following sections.

1.2.1 OR.NET

The OR.NET (Secure Dynamic Networking in the Operating Room and Clinic) is a project (2012-2016) of Heidelberg University Hospital and RWTH Aachen University funded by the German Federal Ministry of Education and Research. The project was developed in tight connection with providers of integrated operating rooms, manufacturers of medical devices, IT service providers, and software vendors. The goal of the project was the development of certifiable, dynamic, multi-vendor networking opportunities for existing and future devices and software solutions in the medical environment. The project also resulted in creation of a non-profit organization OR.NET eV evaluating concepts of secure dynamic networking of OR components and transforming them into standardization activities.

1.2.2 SCOT

The SCOT (Smart Cyber Operating Theater) project started in 2011 by the initiative of the Institute of Advanced Biomedical Engineering & Science (Tokyo Women's Medical University). The project was funded by the state-backed Japan Agency for Medical Research and Development (AMED). The SCOT was conceived as advanced medical information

analyser for guidance of surgical procedures, brain tumour surgeries in particular, to improve treatment safety and efficiency using high performance computing and networking. The information from all medical devices is consolidated and shared in real-time. The basic version of SCOT (Figure 1.2), essentially displaying patient's information in real-time, was lunched at Hiroshima University hospital in May 2016. The "hyper" version with additional functions such as decision making, navigation system, and robot remote control will start functioning at Tokyo Woman's Medical University by summer 2019.

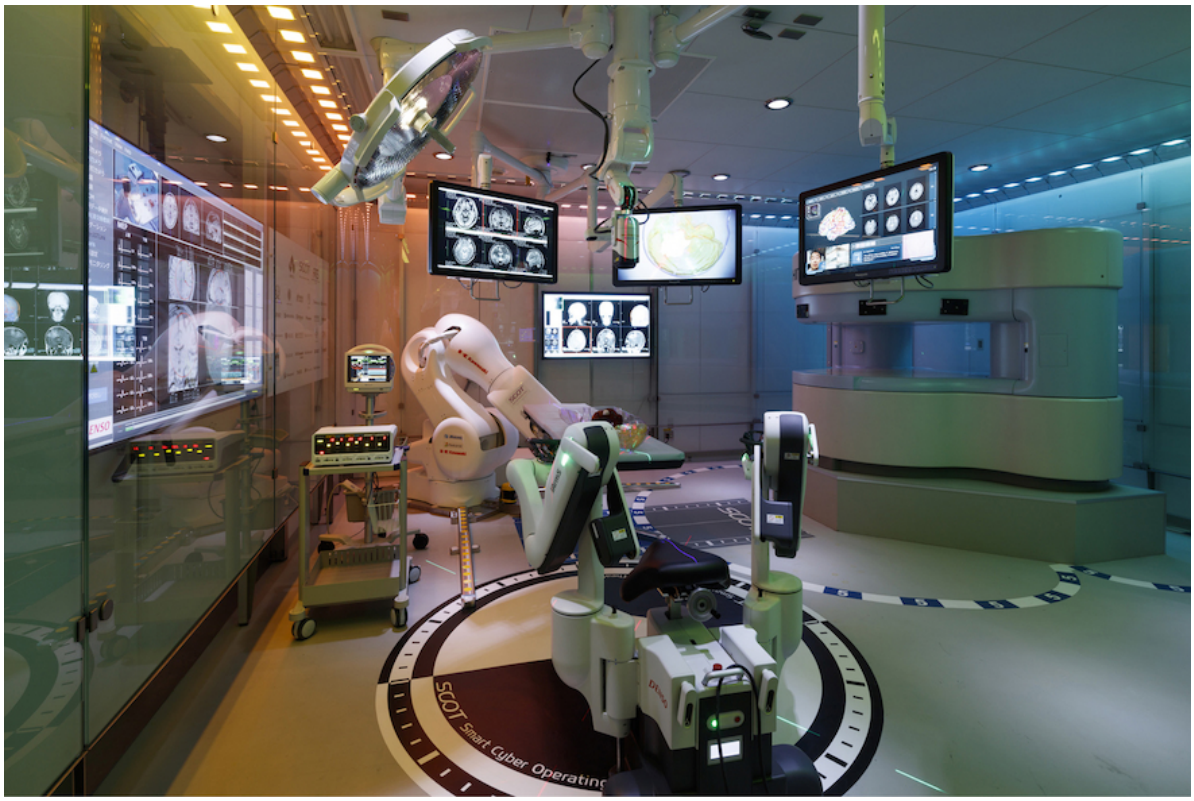

Figure 1.2: Smart Cyber Operating Theater
Source: http://www.g-mark.org

1.2.3 CONDOR

The french 14 million euros project CONDOR, which name stands for Connected Optimized Network & Data in Operating Rooms, started in 2016. It is funded with the Investissement d'Avenir program and managed by the Institute of Research and Technology b<>com in association with private institutes as IHU of Strasbourg and IRCAD, public laboratories LTSI-Inserm and ICube, and industrial partners Thomson Video Networks and

Medtronic IHS. The project, inspired by aeronautics, seeks to develop a "black box" recording all the parameters of a surgical intervention, as well as a "control center" enabling the optimization of patient's care, in ambulatory surgery in particular. The results of the CONDOR project will be integrated in the innovative platforms of the Strasbourg's IHU.

1.3 Surgical Data Science

Multiple devices and treatment techniques introduced in the surgical process generate large amounts of complex versatile data which are the main focus of a recently emerged research direction called *Surgical Data Science* (SDS). The term SDS was first introduced in [Maier-Hein 2017] - a collaborative paper written by the members of the namesake workshop held in Heidelberg in 2016. [1] The term originates from the notion of *Data science* interested in extraction of knowledge from data. SDS has the following definition in the paper: "Surgical Data Science is an emerging scientific field with the objective of improving the quality of interventional healthcare and its value through capturing, organization, analysis, and modelling of data. (...) Data may pertain to any part of the patient care process (from initial presentation to long-term outcomes), may concern the patient, caregivers, and/or technology used to deliver care, and is analysed in the context of generic domain-specific knowledge derived from existing evidence, clinical guidelines, current practice patterns, caregiver experience, and patient preferences". This field is tightly related to the concept of the OR of the future, as it seeks to improve interventional medicine trough data analysis leading to objective decision-making and personalized care. Colossal progress in computing solutions and machine-learning enables information processing beyond human capacities. Figure 1.3 demonstrates how SDS progressively interconnects a multitude of data sources for enhanced computer assistance.

1.4 Situation-awareness

As it was said before, one of the characteristics distinguishing the OR of the future from the present one is a flawless connection of all systems and their cooperation with the surgical team. A control center playing a role of a kapellmeister directing the work of all units is then necessary. This center has to possess a powerful artificial intelligence understanding situation in the OR and sensing clinicians' needs. By keeping track of the procedure and by constantly observing the surgical scene with its actors, it should always be aware of happening events, performed actions and the current state. In the literature, this is called *situation- or context-aware computer-assisted surgery* (CA-CAS). The relevance and applications of context-aware systems are exposed in following sections.

1. http://www.surgical-data-science.org/

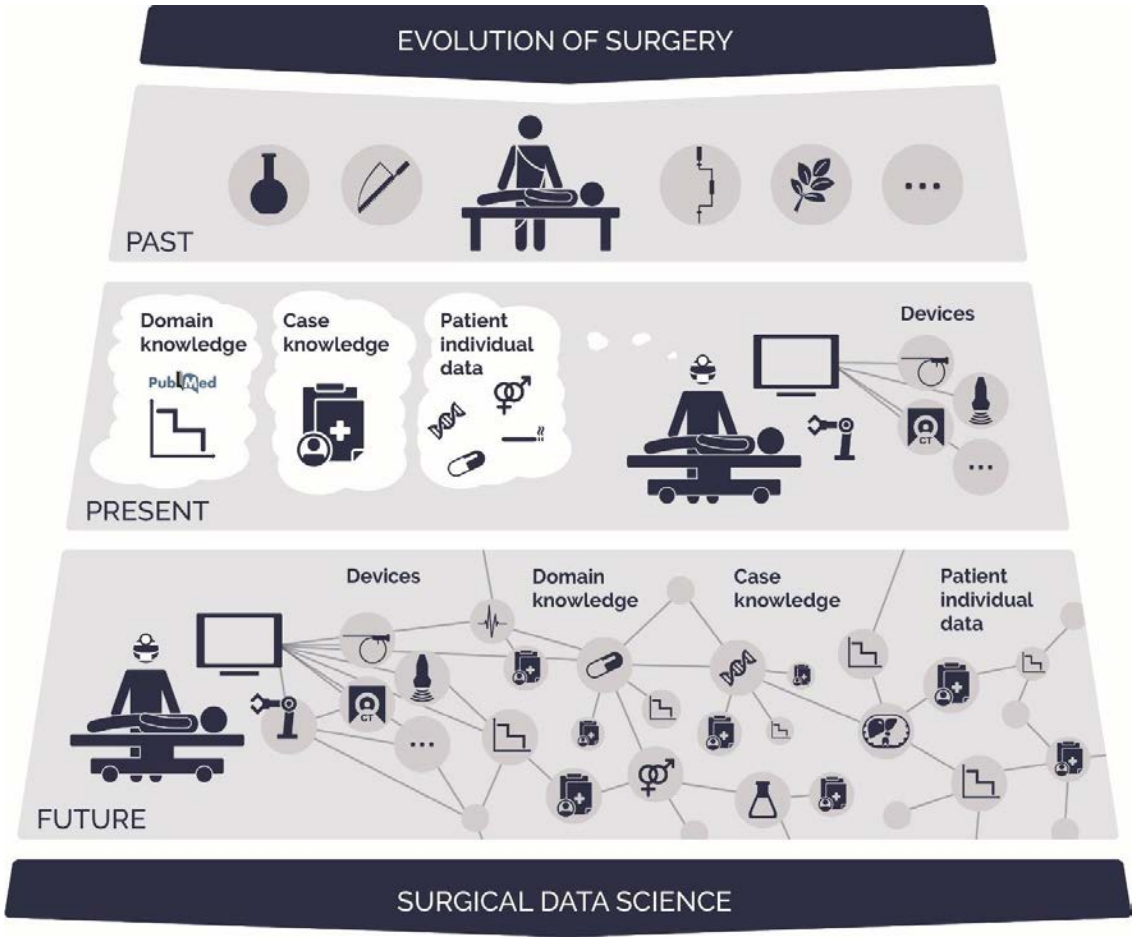

Figure 1.3: Surgical Data Science
Source: [Maier-Hein 2017]

1.4.1 Optimization and management

Hospital expenses are tightly connected to the efficiency of operating suites. Even for highly standardized procedures, the operating time can vary greatly depending on patient's particularities, stage of the disease and unexpected complications occurring during surgery. In most hospitals, the OR schedule is based on average durations of planned surgeries. It is not rare to have long delays or holes in the schedule when OR time is simply wasted. In this context, real-time automatic awareness of situation has several applications.

— *Optimization of surgical process.* Reducing operating time through optimization of surgical process is highly desired to decrease the cost of treatment and take care of more patients.

— *Scheduling.* Predicting remaining time is useful to adapt the OR schedule, start preparing the next patient at the right moment, and thus enable quick transitions between operations [Franke 2013, Guédon 2016]. Moreover, displaying real-time information about the OR state outside the room can be useful for nurses as it takes away the necessity of manual check [Bhatia 2007].

— *Resource anticipation and management.* Operating suites share expensive equipment which is constantly relocated according to the needs of every procedure. In-time detection or anticipation of these needs can improve the organization of equipment use [Maktabi 2017].

1.4.2 Assistance and ergonomics

One of the purposes of the context-aware system is providing the right assistance at the right moment. In addition, it can improve interactions between clinicians and machines providing the assistance and make the surgical process more efficient.

— *Intra-operative assistance.* Different kinds of assistance (e.g., robotic, visual, haptic) can support the surgeon in the technical aspect of the procedure. A shared control of surgical instruments between the surgeon and a context-aware robot enables effortless automatic execution of certain difficult tasks [Nessi 2015, Fard 2016]. Miyawaki et al. [Miyawaki 2005] created a robotic scrub nurse to fill the lack of qualified surgical personnel. Katić et al. [Katić 2010] offered visual assistance in a form of augmented reality to automatically display planned positions of the implants during dental surgery.

— *Human-Machine Interface.* Analysis of interactions with new devices can reveal integration and usability problems, as well as improve the interface [Hübler 2014].

1.4.3 Decision support

Every patient and every surgery are unique. Although technical skills, conceptual and procedural knowledge are very important, the outcome of the surgery also depends on the surgeon's ability of making correct intra-operative decisions. Decision-making is one of the non-technical skills that a surgeon must have [Flin 2013]. Automatic awareness of surgical situation enables different types of decision support.

— *Information display.* The surgery-related information (e.g., treatment plan, patient's history, vital signs, data coming from countless devices) is much more useful if relevant to the surgical task performed at the moment.

— *Decision-making.* The context-aware system can also support the decision making process by advising a set of actions to undertake in a given state of the procedure to achieve the best results. This support would be especially valuable in emergency situations when the surgeon has no time for consideration.

— *Adverse events detection.* The on-going surgery can also be analysed to detect deviations from the usual surgical process in order to prevent adverse events and complications [Bouarfa 2012, Huaulmé 2017a].

1.4.4 Training and assessment

— *Analysis.* Observation and analysis of surgical process is essential for understanding surgical practice, defining the right educational standards, and improving teaching and learning.

— *Training.* Training of junior surgeons is a time-consuming process requiring supervision by a senior surgeon. Automatic real-time workflow recognition can help to provide a feedback and indicate precisely committed workflow errors just after or even during the surgery.

— *Assessment.* Traditional approaches for evaluation of surgeons are qualitative and subjective since based on human observations and questionnaires. Automatic recognition enables quantitative objective assessment of surgical skills.

1.4.5 Task automation

— *Device control.* The surgical equipment and devices can be automatically (dis)activated, moved or prepared for use depending on the moment of surgery and the surgeon's needs [Franke 2015b].

— *Event triggering.* Rodas et al. in [Rodas 2017] tracked body positions to automatically inform clinicians when they get too much effected by the radiation during X-ray guided minimally invasive procedures.

— *Documentation.* The context-aware system can automatically create pre-filled surgical records and edit post-operative reports to economize surgeon's time [Agarwal 2006].

— *Procedure annotation.* Manually annotated clinical data was used for the applications as remaining time estimation [Franke 2013], next surgical task prediction [Forestier 2017]) and analysis of surgical behaviours [Greenhalgh 2001]. Automatic annotation can replace tedious manual process.

The research on this topic started about a decade ago. Despite great advancements, it remains a relatively recent field inspiring minds of scientists and clinicians. The next chapter presents a a review of the research on recognition of surgical workflow for situation awareness.

CHAPTER

2 Related work in recognition of surgical workflow

Preamble

This chapter presents a review of published in the literature methods for recognition of surgical workflow. After describing the context, it defines the scope of interest and determines the criteria for selection of publications for the review. It then explains the methodology of the current reviewing process providing a detailed diagram containing the main examined aspects. The review of each aspect is presented in a text form as well as in a structured table. Finally, we discuss the problems blocking the domain addressed in this thesis.

Contents

2.1 Context

As introduced in the previous chapter, operating room has significantly evolved becoming a high-tech complex environment. Computer assistance rapidly took an important place in modern surgery inciting further methodological and technological advances. Researches from all over the world started working towards a new generation of intelligent ORs setting ambitious goals and taking arduous challenges. One of these challenges is a design of an integrated situation-aware system understanding and following the procedure. Although, situation awareness in terms of happening events or other patient or device related factors is important, automatic recognition of surgical workflow is perhaps the most clinically useful feature in CA-CAS. It is thus the main focus of this chapter.

Mackenzie et al. were among the first who proposed to create a procedural model. In [MacKenzie 2001] it was based on a structured multi-level decomposition describing surgical actions performed during surgery. The same year, Jannin et al. [Jannin 2001] also proposed to model a neurological procedure using a decomposition based on Unified Mark-up Language. Later, a notion of *surgical workflow* was introduced. Neumuth et al. in [Neumuth 2006b] presented it as "general methodological concept of the acquisition of process descriptions from surgical interventions, the clinical and technical analysis of them, and the automated processing into workflow schemes that are able to drive a workflow management system as a meta process control for the operating room of the future". Meanwhile, Jannin and Morandi in [Jannin 2007] defined surgical workflow as "the automation of a business process in the surgical management of patients, in whole or part, during which documents, information, images or tasks are passed from one participant to another for action, according to a set of procedural rules". In this thesis, we employ the notion of surgical workflow to describe a sequence of surgical tasks that are accomplished to perform a procedure following a repetitive schema. This definition is close to the one that was given in [Padoy 2010].

The workflow of every surgical procedure is unique due patient- and surgeon-specific features (e.g., anatomy, disease stage, patient's reaction to drugs, surgeon's experience and habits). Despite this uniqueness, it is possible to create an abstracted model representing a set of surgeries called *Surgical Process Model* (SPM). Two types of SPM models are distinguished: individual and generic [Neumuth 2011]. The individual SPM (iSPM) describes the workflow of one particular intervention which is obtained by data acquisition. The generic SPM (gSPM) represents the set of theoretically possible ways to perform the procedure. Unlike iSPM, the gSPM does not represent a real intervention. To build a gSPM, first an iSPM for each intervention has to be created from raw data (i.e., video or sensor signals). These iSPMs are then aggregated together to form the gSPM as in Figure 2.1 [Huaulmé 2017b]. The SPM is often used in the process of workflow recognition.

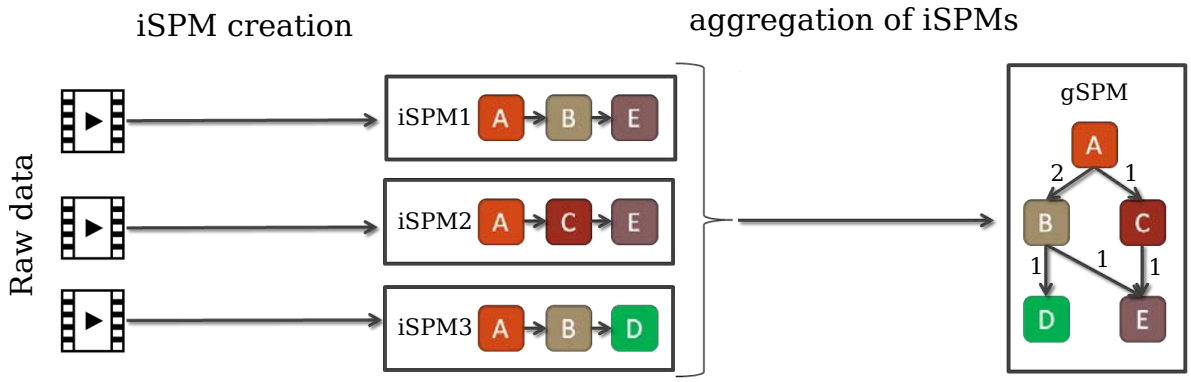

Figure 2.1: Process of gSPM construction. Here the gSPM is represented as a graph where several paths are possible from each state. *Source: [Huaulmé 2017b]*

2.2 Scope of interest

The current review gathered scientific papers proposing approaches and algorithms for off-line or on-line recognition of surgical workflow at different granularity levels (see Section 2.5.3) to be one day applied within the situation awareness context. Here, the term recognition is employed in a broad sense, it includes discovering the workflow from raw signals (e.g., videos, data from sensors), predicting the following surgical tasks(s) from the previous ones, and inferring the tasks of higher granularity level from the lower. The following inclusion/exclusion criteria define the scope of interest more accurately:

+ Only the methods designed for scheduled surgeries conducted in real sterile surgical conditions were considered.
- The works placed in the context of anaesthesia [Houliston 2011], trauma [Chakraborty 2013], emergency [Parlak 2011] or intensive care [Lea 2012] were not taken into account. At the exception of using some sensing devices not portable into sterile conditions, methodologically the recognition problem is close to the surgical context. However, a major difference in workflow, clinical needs and applications exists.
- The papers as [Béjar Haro 2012, DiPietro 2016, Gao 2016a, Nakawala 2017] placing the work in the pre-operative training context were excluded. Training environments presuming execution of separate tasks in simplified conditions are much more constrained than actual surgeries.
+ The papers targeting training and assessment applications for real surgeries were, however, included.
+ The review encircled only the methods dealing with surgical workflow as a sequence of tasks to realize a surgical objective.
- Therefore, the methods recognizing occasional or incidental events were excluded.

- The papers as [Glaser 2015, Neumuth 2012] proposing stand-alone methods for recognition of surgical objects related to situation understanding were excluded unless the recognized objects were explicitly correlated with the workflow.
- The papers proposing pure modelling approaches without any recognition compound were not taken into account.

2.3 Search methodology

The search of literature for the present review was done by means of Google Scholar. Three categories of key words were used for the search: 1) “surgery”, “surgical”, “operation”, and “operating”, 2) “recognition”, “classification”, “segmentation”, “analysis” and “monitoring”, and 3) “process”, “workflow”, “procedure”, “state”, “task”, “phase”, “step”, “activity”, “gesture”. Queried references had to contain at least one word from each category in their title and be published from 2007 to June 2017. In the next step, we only selected the queried references with titles relevant to the subject, which then formed a subset of publications for a closer consideration. In the same step, only the articles and papers from peer-reviewed international journals, conferences and workshops published in English were kept. Additionally, relevant citations found in the selected papers were also added to the list. They were then browsed to select only those corresponding to the scope of interest. Finally, only major publications were chosen. When a work on the subject was published multiple times in incremental manner, either more recent or impacting journal was kept. A total of 34 publications were finally chosen for a full-text review. The flow chart from Figure 2.2 shows the process of selection with the number of references at each step.

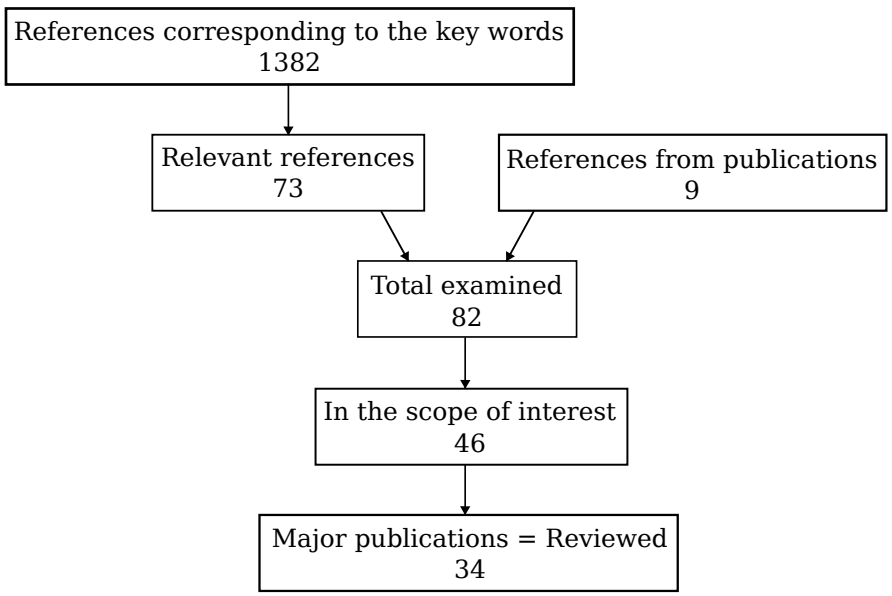

Figure 2.2: Process of selection of publications for a full-text review

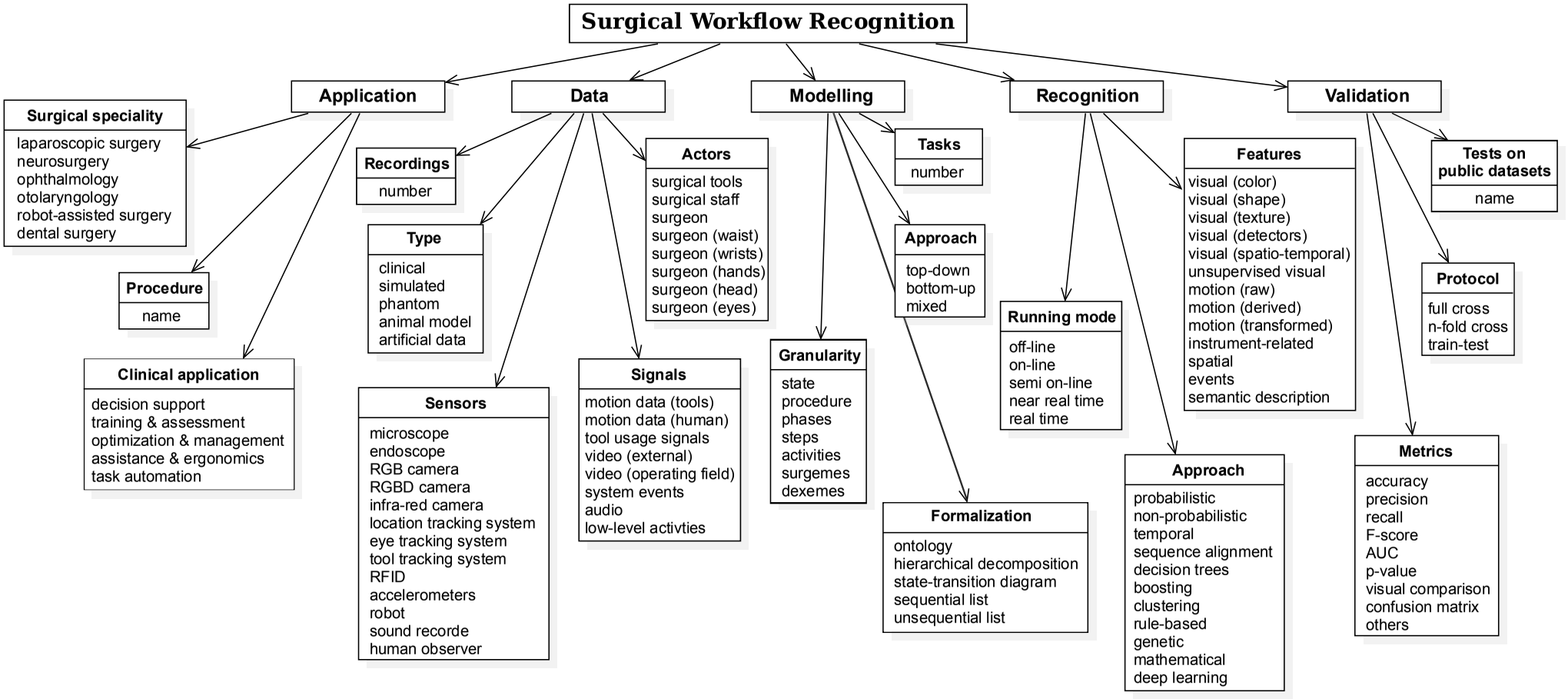

Figure 2.3: Review organization

Table 2.1: Classification of reviewed publications

N°	Reference	Application			Data				
		Surgical speciality	Procedure	Clinical application	Type	Recordings	Sensors	Signals	Actors
1	Ahmadi 2009	neurosurgery, laparoscopic surgery	percutaneous vertebroplasty, cholecystectomy	training & assessment	phantom	15, 1	accelerometers	motion data (humain)	surgeon (waist and wrists)
2	Bardram 2011	laparoscopic surgery	appendectomy	decision support, task automation	simulated	4	location tracking system, RFID	motions data (humain), tool usage signals	surgical staff, surgical tools
3	Bhatia 2007	n/a	n/a	optimization & management	clinical	24 hours	RGB camera	video (external)	no
4	Blum 2010	laparoscopic surgery	cholecystectomy	training & assessment	clinical	10	endoscope, human observer	video (operating field), tool usage signals	surgical tools
5	Bouarfa 2011	laparoscopic surgery	cholecystectomy	n/a	clinical	10	human observer	tool usage signals	surgical tools
6	Dergachyova 2016	laparoscopic surgery	cholecystectomy	decision support, task automation	clinical	7	endoscope, human observer	video (operating field), tool usage signals	surgical tools
7	Droueche 2014	ophthalmology	epiretinal membrane surgery, cataract surgery	decision support	clinical	23, 250	microscope	video (operating field)	surgical tools
8	Forestier 2015	neurosurgery	lumbar disc herniation	training & assessment	clinical	22	human observer	low-level activities	surgeon
9	Forestier 2017	neurosurgery	anterior cervical discectomy and fusion, lumbar disc herniation	training & assessment, decision support	clinical	18, 24	human observer	low-level activities	surgeon
10	Franke 2015a	neurosurgery	brain tumor removal	task automation	clinical	60	human observer	low-level activities	surgeon
11	Katić 2010	dental surgery	dental implant surgery	assistance & ergonomics	phantom, artificial data	n/a	tool tracking system	motion data (tools)	surgical tools
12	Katić 2015	laparoscopic surgery	pancreas resection, cholecystecomy, adrenalectomy	decision support	clinical	11, 3, 5	human observer	low-level activities	surgeon
13	Katić 2016	laparoscopic surgery	pancreas resection, adrenalectomy	decision support	clinical	11, 5	human observer	low-level activities	surgeon
14	Klank 2008	laparoscopic surgery	cholecystectomy	training & assessment, decision support	clinical	6	endoscope	video (operating field)	surgical tools
15	Lalys 2011	neurosurgery	pituitary surgery	optimization & management, assistance & ergonomics	clinical	16	microscope	video (operating field)	surgical tools
16	Lalys 2012	ophthalmology	cataract	optimization & management, assistance & ergonomics	clinical	20	microscope	video (operating field)	surgical tools

17	Lalys 2013	ophthalmology	cataract	decision support, assistance & ergonomics	clinical	20	microscope	video (operating field)	surgical tools
18	Malpani 2016	robot-assisted surgery	hysterectomy	training & assessment	clinical	24	robot	motions data (tools), system events	surgical tools, surgeon (head)
19	Meißner 2014	otolaryngology	endoscopic sinus surgery	optimization & management	phantom	24	accelerometers, RFID	motions data (humain), tool usage signals	surgeon (hands), surgical tools
20	Nara 2017	neurosurgery	craniotomy	n/a	clinical	10	location tracking system, RGB camera	motions data (humain), video (external)	surgical staff
21	Padoy 2009	n/a	n/a	training & assessment, decision support	simulated	22	RGBD cameras	videos (external)	surgical staff
22	Padoy 2012	laparoscopic surgery	cholecystectomy	task automation	clinical	16	human observer	tool usage signals	surgical tools
23	Primus 2016	laparoscopic surgery	cholecystectomy	decision support	clinical	6	endoscope	video (operating field)	surgical tools
24	Quellec 2014a	ophthalmology	epiretinal membrane surgery, cataract	decision support	clinical	23, 100	microscope	video (operating field)	surgical tools
25	Quellec 2014b	ophthalmology	cataract	decision support	clinical	186	microscope	video (operating field)	surgical tools
26	Quellec 2015	ophthalmology	cataract	decision support	clinical	186	microscope	video (operating field)	surgical tools
27	Schreiter 2016	n/a	n/a	assistance & ergonomics	simulated	12	tool tracking system, RGBD cameras	motions data (tools), videos (external)	surgical staff, surgical tools
28	Thiemjarus 2012	laparoscopic surgery	cholecystectomy	training & assessment	animal model	15	eye and tool tracking systems	motion data (humain and tools)	surgeon (eyes), surgical tools
29	Tran 2017	laparoscopic surgery	cholecystectomy	n/a	simulated	28	RGB camera	video (external)	surgeon
30	Twinanda 2015	neurosurgery	vertebroplasty	optimization & management, task automation	clinical	11 days	RGBD cameras	videos (external)	surgical staff
31	Twinanda 2017	laparoscopic surgery	cholecystectomy	optimization & management, decision support	clinical	80 + 7	endoscope	video (operating field)	surgical tools
32	Unger 2014	otolaryngology	endoscopic sinus surgery	optimization & management, assistance & ergonomics	simulated	9	infra-red camera	video (external)	surgeon (hands)
33	Weede 2012	laparoscopic surgery	single-port sigma resection	optimization & management, task automation	animal model, artificial data	3, 8	tool tracking system, endoscope, sound recorder	motions data (tools), video (operating field), audio	surgical tools
34	Zia 2017	robot-assisted surgery	multiple surgeries	optimization & management, training & assessment	animal model	9	robot	motion data (tools), system events	surgeon tools

N°	Reference	Modelling				Recognition			Validation and Evaluation		
		Granularity	Tasks	Approach	Formalization	Features	Approach	Running mode	Protocol	Metrics	Tests on public datasets
1	Ahmadi 2009	steps	9, 2	bottom-up	unsequential list	motion (raw)	non-probabilistic, temporal	n/a	n/a	TPR, FPR, TNR, FNR motif-related	no
2	Bardram 2011	state and steps (steps)	7 + 21	mixed	state-transition diagram, hierarchical decomposition	spatial, instrument-related	decision trees	on-line	full cross	confusion matrix	no
3	Bhatia 2007	state	4	bottom-up	state-transition diagram	visual (color)	non-probabilistic, temporal	real-time	full cross	accuracy	no
4	Blum 2010	phases	14	bottom-up	state-transition diagram	visual (color), instrument-related	temporal, sequence alignment	on-line off-line	full cross	accuracy, p-value, confusion matrix	no
5	Bouarfa 2011	phases	5	bottom-up	state-transition diagram	instrument-related	temporal	n/a	full cross	standard (3), AUC	no
6	Dergachyova 2016	phases	7	bottom-up	state-transition diagram	visual (color, texture, shape), instrument-related	boosting, temporal	on-line	full cross	standard (3), standard with delay, temporal delay, confusion matrix, visual comparison	EndoVis2015
7	Droueche 2014	phases	3, 11	top-down	sequantial list	visual (spatio-temporal)	sequence alignment	semi on-line	2-fold cross, train-test	precision	HOLLYWOOD2
8	Forestier 2015	phases and activities (phases)	4	bottom-up	hierarchical decomposition	semantic description	probabilistic	n/a	full cross	standard (3), F-score, confusion matrix	no
9	Forestier 2017	activities	82, 108	bottom-up	sequential list	semantic description	sequence alignment	off-line	full cross	standard (2), F-score, prediction confidence, no decision count, confusion matrix, visual comparison	no
10	Franke 2015a	phases, steps and activities (steps)	4 + 17 + 168	top-down	state-transition diagram, hierarchical decomposition	semantic description	temporal	real-time	full cross	accuracy	no
11	Katić 2010	phases	4	top-down	ontology	spatial	probabilistic, rule-based	real-time	n/a	accuracy	no
12	Katić 2015	phases and activities (phases)	12 + 501, n/a, 9 + 221	top-down	ontology	semantic description	rule-based	near real-time	n/a	accuracy	no
13	Katić 2016	phases and activities (phases)	12 + 501, 9 + 221	mixed	ontology	semantic description	decision trees, genetic	real-time	full cross	accuracy, variance	no
14	Klank 2008	phases	6	bottom-up	sequantial list	visual (color, texture, shape)	non-probabilistic, genetic	near real-time	train-test	accuracy	no
15	Lalys 2011	phases	6	bottom-up	state-transition diagram	visual (color, texture, shape)	non-probabilistic, temporal	n/a	10-fold cross	standard with delay	no

16	Lalys 2012	phases	12	bottom-up	state-transition diagram	visual (color, texture, shape, detectors)	non-probabilistic, temporal, boosting, sequence alignment	on-line off-line	10-fold cross	accuracy	no
17	Lalys 2013	phases and activities (activities)	8 + 18	bottom-up	hierarchical decomposition	visual (color, detectors)	non-probabilistic, boosting, sequence alignment	off-line	full cross	accuracy	no
18	Malpani 2016	phases	5	bottom-up	sequantial list	instrument-related, events	(non) probabilistic, decision trees, deep learning	n/a	n/a	standard (3), Levenshtein distance	no
19	Meißner 2014	activities	10	bottom-up	state-transition diagram	motion (transformed), instrument-related	temporal	n/a	full cross	standard (3)	no
20	Nara 2017	phases	6	bottom-up	sequantial list	motion (transformed), visual (spatio-temporal)	non-probabilistic, clustering	real-time	full cross	accuracy	no
21	Padoy 2009	state	10	bottom-up	state-transition diagram	visual (spatio-temporal)	temporal	on-line off-line	full cross	accuracy	no
22	Padoy 2012	phases	14	bottom-up	state-transition diagram	instrument-related	temporal, sequence alignment	on-line off-line	full cross	standard (3), failure rate, negative predictive value	no
23	Primus 2016	phases	6	bottom-up	sequantial list	visual (color, detectors)	non-probabilistic	n/a	n/a	nb of transitions, temporal delay	no
24	Quellec 2014a	phases	3, 9	bottom-up	sequantial list	visual (color, texture, spatio-temporal)	decision trees, mathematical	semi on-line	train-test	AUC	HOLLYWOOD2
25	Quellec 2014b	phases	10	bottom-up	sequantial list	visual (spatio-temporal)	probabilistic	semi on-line	train-test	AUC	no
26	Quellec 2015	phases	8	bottom-up	sequantial list	visual (detectors, spatio-temporal)	mathematical	semi on-line	train-test	standard (2), AUC, FPR, p-value	no
27	Schreiter 2016	phases	7	bottom-up	sequantial list	instrument-related, spatial	decision trees	n/a	full cross	confusion matrix	no
28	Thiemjarus 2012	phases	5	top-down	hierarchical decomposition	motion (derived, transformed)	probabilistic	n/a	5-fold cross	accuracy	no
29	Tran 2017	phases	12	bottom-up	state-transition diagram	visual (spatio-temporal)	probabilistic	off-line	full cross	accuracy	no
30	Twinanda 2015	steps	15	bottom-up	unsequential list	visual (spatio-temporal)	non-probabilistic, clustering	n/a	n/a	accuracy	no
31	Twinanda 2017	phases	7	bottom-up	state-transition diagram	unsupervised visual	non-probabilistic, temporal, deep learning	n/a	train-test, full cross	standard (3), nb of tasks in window, confusion matrix, visual comparison	Cholec80, EndoVis2015
32	Unger 2014	phases	12	bottom-up	unsequential list	visual (color)	probabilistic, genetic	n/a	full cross	standard (3), confusion matrix	no
33	Weede 2012	phases	8	bottom-up	sequantial list	instrument-related, spatial, visual (detectors), events	probabilistic	on-line	train-test	accuracy, F-score, failure rate	no
34	Zia 2017	phases	5	bottom-up	unsequential list	motion (derived), events	clustering	n/a	n/a	accuracy, visual comparison	no

2.4 Review methodology

Figure 2.3 displays the diagram according to which the selected publications were reviewed. Every category and its corresponding subcategories addresses a major aspect of context-aware system conception: application, data, modelling, recognition and validation. The *application* indicates the targeted surgery and clinical purposes of the system. The *data* describes the nature, form and acquisition process of the data used to train and test the system. The *modelling* describes the formalization of surgical process model or knowledge representing the chosen procedure. The *recognition* explains the process of "teaching" the system to recognize the surgical workflow from the acquired data. The *validation* defines the methodology used to assess the created recognition system. The results of the review are exposed in Table 2.1 and in Section 2.5. The references in brackets in the text refer to the entries of the table. The current review was motivated by the analysis of the methodology applied to recognition problem solving. It sought to reveal and exhibit issues blocking further development of the field.

2.5 Analysis

2.5.1 Application

Speciality. This subcategory indicates the surgical speciality of the procedure targeted for recognition. Every speciality is characterized by its own surgical objectives, operative techniques, OR settings (e.g., tools, devices, conditions) and required surgical skills. But more importantly, each of them has its own specific needs for computer assistance. Situation-awareness in minimally-invasive laparoscopic surgery seems to attract the greatest part of researchers' attention [1, 2, 4, 5, 6, 12, 13, 14, 22, 23, 28, 29, 31, 33], fol-

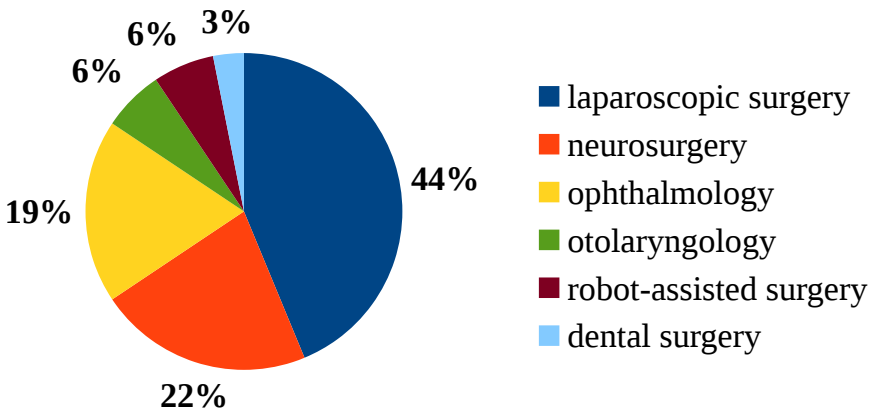

Figure 2.4: Repartition of studies by surgical speciality

lowed by neurosurgery [1, 8, 9, 10, 15, 20, 30] and ophthalmology [7, 16, 17, 24, 25, 26]. Other impacted specialities were dental surgery [11], otolaryngology [19, 32] and robot-assisted surgeries (RAS) [18, 34]. In fact, a large number of publications in RAS domain exist: numerous contributions were made by the team of John Hopkins University [Béjar Haro 2012, DiPietro 2016, Gao 2016b, Gao 2016a, Tao 2013]. However, they mostly place recognition in the training context and thus were excluded from the review as explained in Section 2.2.

Procedure. This subcategory determines the type of surgery the workflow of which has to be recognized. The most targeted procedures were cholecystectomy [1, 4, 5, 6, 12, 14, 22, 23, 28, 29, 31] and cataract surgery [7, 16, 17, 24, 25, 26]. A particular choice of procedure may be explained by simpler conditions or facility to leverage informative input signals without installing too much of new sensors. For example, during laparoscopic surgery the whole procedure is performed through endoscopic video with a limited number of instruments and field of view. Cholecystectomy and cataract are highly standardized procedures with few deviations. Such procedures are also conducted frequently which facilitates data collection. Other targeted procedures were vertebroplasty [1, 30], sinus surgery [19, 32], epiretinal membrane surgery [7, 24], pancreas resection [12, 13], adrenalectomy [12, 13], appendectomy [2], hysterectomy [18], sigma resection [33], cervical discectomy [9], lumbar disc herniation [9], pituitary surgery [15], craniotomy [20], brain tumour removal [10] and dental implant surgery [11].

Clinical application. Clinical application defines the purpose of the recognition system. Authors often give an insight on clinical use of the proposed methods, though most of the time multiple possible applications are cited. The final application is mostly constrained by the origin of the sensors and the working mode of the method (e.g., on-line or off-line) that will be discussed later in this chapter. Decision support was the most targeted application [2, 6, 7, 9, 12, 13, 14, 17, 21, 23, 24, 25, 26, 31], followed by training & assessment [1, 4,

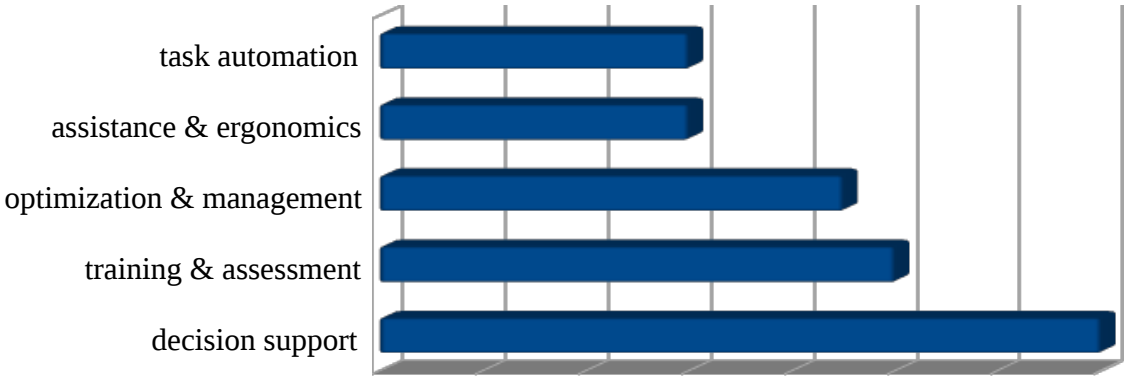

Figure 2.5: Targeted clinical applications

8, 9, 14, 18, 21, 28, 34, 18] and optimization & management [3, 15, 16, 19, 30, 31, 32, 33, 34]. The applications as assistance & ergonomics [11, 15, 16, 17, 27, 32] and task automation [2, 6, 10, 22, 30, 33] were targeted slightly less.

2.5.2 Data

Type. This subcategory defines the nature of the acquired data. Many medical researchers get inspired by recognition approaches from other domains. Unfortunately, some of them are hard to translate into the surgical context mostly due to sensor-related issues. Some sensors do not respect sterile conditions, disturb the surgeon or they are simply too burdensome or expensive to install in the OR (e.g., installation works, approval, certification). These constraints can be gotten around over time. However, researchers often prefer to make preliminary tests of the adopted approaches to see if they suit their objectives. That is why some published work was made in simulated conditions [2, 21, 27, 29, 32], using animal models [28, 34] or phantoms [1, 11, 19]. In some cases additional artificial data was generated to support the learning process [11, 33]. Luckily, certain devices already present in the OR or some other easily integratable sensors allowed to assess the methods on real clinical data.

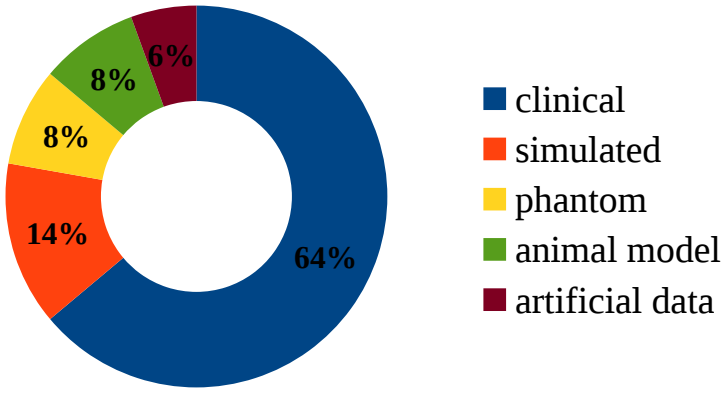

Figure 2.6: Types of data used in the analysed studies

Number of recordings. This subcategory indicates how many interventions (i.e., surgical cases) were recorded or simulated for each study. Some works trained and assessed their methods on several procedures; in this case, the table provides a separate number for each procedure. Three publications [3, 11, 30] did not indicate the number of observed interventions, but two of them cited the amount of recorded hours or days [3, 30].

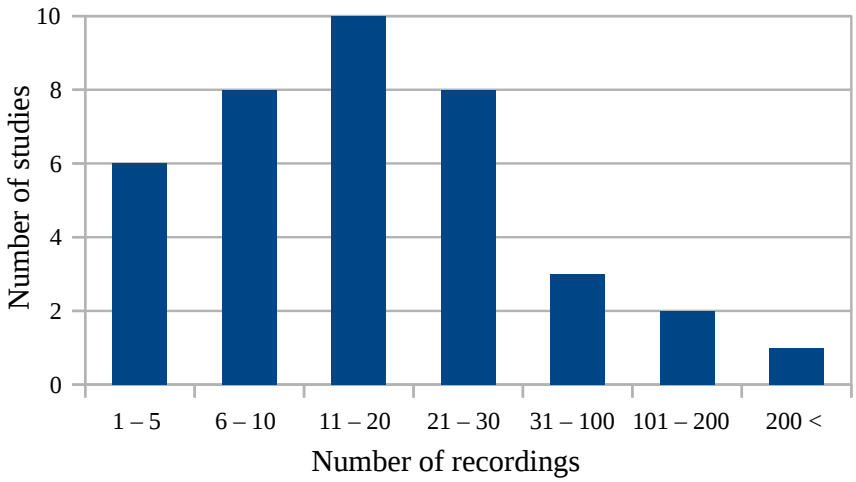

Figure 2.7: Number of recorded interventions per study

Sensors. This subcategory designates the types of devices acquiring the data. Numerous sensors were used to record information about the surgical worflow: a robot providing versatile information about the instruments and system events [18, 34], a location tracking system detecting people positions inside the room and tracing their trajectories [2, 20], accelerometers placed on surgeon's body parts [1, 19], eye [28] and tool [11, 27, 28, 33] tracking systems recording trajectories, radio frequency identification (RFID) technology [2, 19], RBG/RGBD [21, 27, 29, 30] and infra-red [32] cameras, a sound recorder [33], as well as standard devices already present in the OR as endoscope [4, 6, 14, 23, 31, 33], microscope [7, 15, 16, 17, 24, 25, 26] and rgb surveillance camera [3, 20]. In some cases of

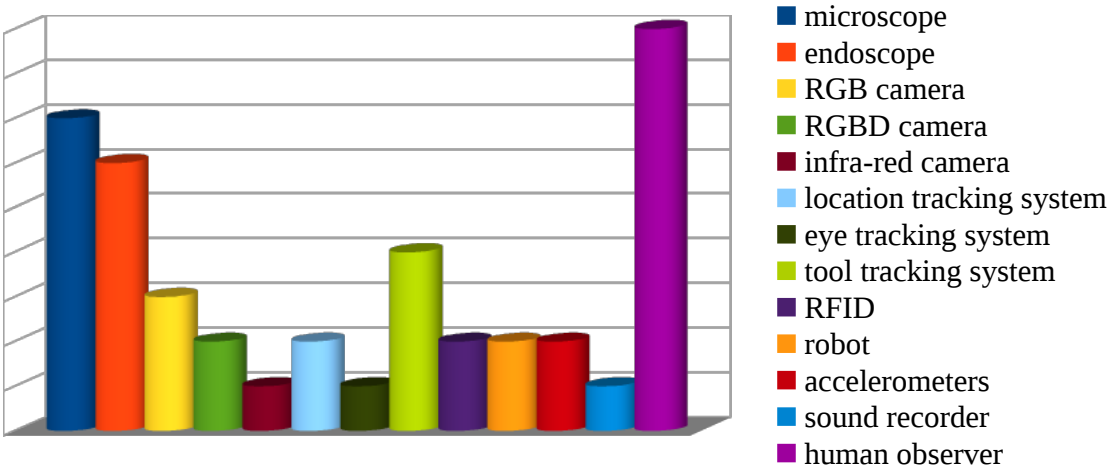

Figure 2.8: Used sensors

real surgeries, human observers manually annotated procedural workflow by following the surgery in real-time or watching its video recordings later on [4, 5, 6, 8, 9, 10, 12, 13, 22]. This approach supposes that the necessary sensors will be placed in the OR sometime in the future.

Signals. This subcategory indicates the nature and form of the acquired data. The signals from the sensors used in the reviewed publications were divided into following groups: videos, motion data, tool usage signals, audio [Weede 2012] and other specific signals as robotic system events [18, 32]. The videos recorded in RBG, RGBD or infra-red format captured the operating field [4, 6, 7, 14, 15, 16, 17, 23, 24, 25, 26, 31, 33] or external views [3 , 20, 21, 27, 29, 30, 32]. The motion data characterized instrument [11, 27, 28, 33, 34] or human motions (body [2, 20], hands [19], eyes [28], wrists and waist [1]) in terms of position (i.e. trajectory) [2, 11, 20, 27, 28, 33, 24] and/or acceleration [1, 19, 34]. The tool usage signals indicating surgical instruments currently in use were recorded automatically [2, 18, 19] or manually [4, 5, 6, 22]. The low-level surgical activities (see Section 2.5.3) performed by the surgeon were recorded manually in all cases [8, 9, 10, 12, 13].

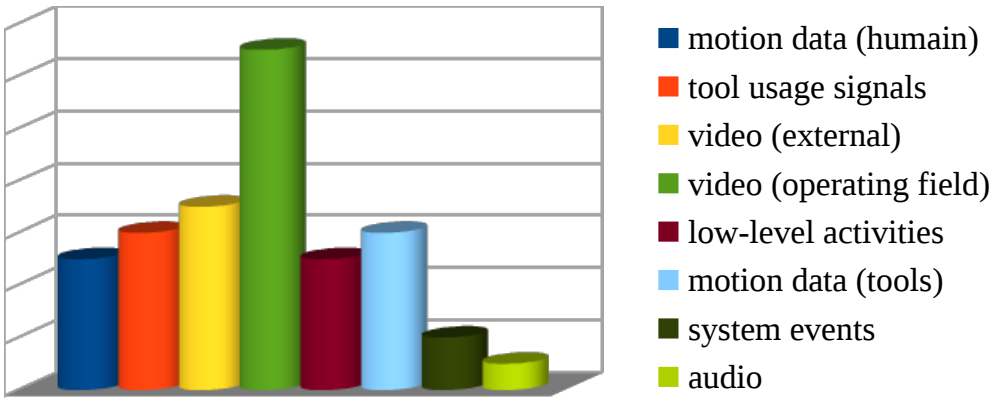

Figure 2.9: Acquired signals

Actors. Actors define persons and manipulators (i.e., instruments or body parts) executing the actions that were recorded by the sensors. The sensors can perceive the procedure from various angles and follow different key actors. In the reviewed publications, most of the time the surgeon (in whole [8, 9, 10, 12, 17, 29], hands [19, 32], wrists [1], waist [1], head [18] or eyes [28]) and/or manipulated surgical tools [2, 4, 5, 6, 7, 11, 14, 15, 16, 17, 22, 23, 24, 25, 26, 27, 28, 31, 33, 34] were observed. An effort in recording actions and movements of the surgical staff including anaesthesist, scrub, and assistant nurses was also made [2, 20, 21, 27, 30].

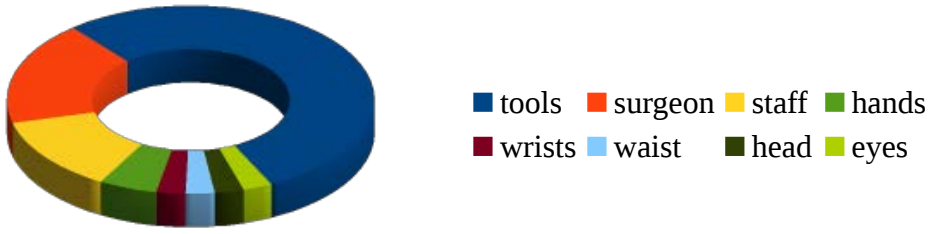

Figure 2.10: Observed actors and manipulators

2.5.3 Modelling

Granularity. Granularity is defined as a level of abstraction for describing the surgical process; it determines in which details the surgical procedure is modelled. Unfortunately, no universal taxonomy or vocabulary defining the difference between the granularity levels exist. Researchers often use the same term when speaking about different notions, or use some other uncommon terms to describe some of the concepts presented here. The taxonomy presented here was inspired from [Lalys 2014] that presented a complete review on surgical process modelling. It was further enriched by [Despinoy 2016, Huaulmé 2017a] who added more precision for low granularity levels concerning surgical gestures. The surgical workflow can be modelled at different levels: dexemes, surgemes, activities, steps, phases, procedure and state. The most detailed level of workflow observation is *dexemes*. They describe short gestures performed by one hand of the surgeon devoid of medical sense, e.g. “turn left”. A sequence of dexemes allows to accomplish a *surgeme*, which represents a surgical gesture made with a precise purpose and having an explicit semantic meaning, e.g. “make a knot”. An *activity* is a physical action described by the anatomical structure operated by the surgeon at the moment, the used surgical tool and the performed action. A *step* is defined as a set of activities towards a surgical objective. A *phase* represents a longer interval of time including several steps and may involve interactions with other members of the surgical team. A *procedure* corresponds to the entire surgery, which starts with the first incision and ends with the last stitch closing the patient. In [Lalys 2014, Despinoy 2016, Huaulmé 2017a], phase was the highest granularity level. After reviewing the publications, we, however, felt the urge to add a higher modelling level - state. A *state* represents an extensive period of the surgical process standing out of the surgery itself and representing the state of the operating room, e.g. patient arrival, preparation, surgery, cleaning. In this thesis, an instance of any granularity level will be called a *surgical task*. Table 2.1 indicates a granularity level for each examined study. If several levels were modelled, the value inside parentheses indicates the one used for recognition. The border between steps and phases still remains fuzzy. Here, the separation was essen-

tially made based on the number of modelled tasks and their descriptions. The dexemes and surgemes were not studied in the reviewed publications, as they are mostly considered in the context of pre-operative dexterity training. The procedure level was not covered in the reviewed publications neither, as it is mostly relevant for modelling purposes. It represents no particular interest for recognition except automatically distinguishing between different types of surgery.

Figure 2.11: Granularity levels

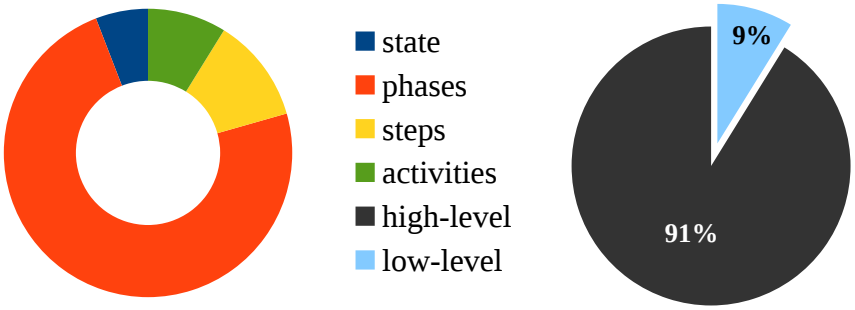

Figure 2.12: Repartition of methods by granularity levels

Number of tasks. This subcategory defines the number of distinct surgical tasks that were modelled and recognized. This number is correlated with the complexity of the recognition problem. A higher number is typically connected to a lower granularity level, hence, a greater challenge. The workflows of the same procedure varied in different studies resulting in different number of tasks. In a large measure, this depends on the surgical style or taxonomy accepted at each particular hospital. In some cases, certain tasks that actually took place during surgery got excluded from the study. This came from inability to record the data in certain moments of surgery, or was caused by a short duration of tasks or even their irrelevance to the situation understanding.

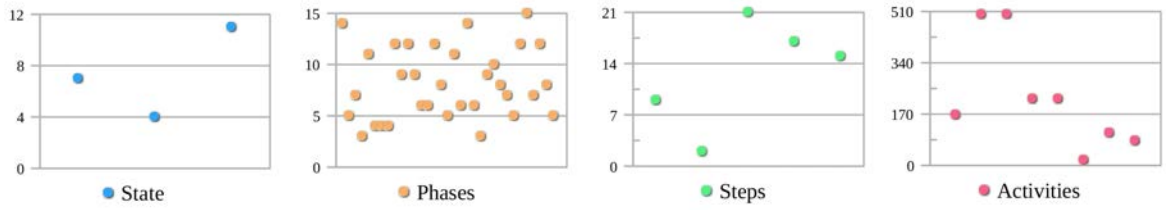

Figure 2.13: Number of distinct surgical tasks modelled per study

Approach. This subcategory defines the formalization approach used to model the surgical process of the chosen procedure. Two main approaches for SPM formalization exist: bottom-up and top-down. A bottom-up approach means inference of an SPM (up) describing the process through analysis of observed procedures (bottom). Despite the fact that such a model suits (but not necessarily explains) all the observations the best, it is not complete. It is possible to observe later on something that will not suit the constructed model any more. This approach is generally easier and takes less time, though the recognition of a procedure with a workflow deviating from the observed ones can be difficult. The top-down approach, on the other hand, assumes a collection of fundamental knowledge (top) about the domain first. That way, the knowledge contained in the model should explain the observations (down). Most of the time, it implies an advanced analysis of medical literature, detailed interviews with surgeons and an appropriate structure for knowledge representation. These two approaches can also be combined. Please refer to Table 2.1 to see which approach was used in every reviewed publication.

Formalization. Formalization determines the type of knowledge representation used to model the surgical process. According to [Lalys 2014], different modelling approaches implicate different formats of knowledge representation. The best example of a top-down approach implementation is ontology [11, 12, 13] defining concepts and entities existing in the domain, as well as relations between them. Other formats generally more suitable for bottom-up and mixed approaches are hierarchical decompositions, state-transition diagrams, sequential and unsequential lists. A hierarchical decomposition model is a representation where hierarchical relations between different granularity levels are strictly defined [2, 8, 10, 17, 28]. State-transition diagrams are 2D graphs determining the transitions between the surgical tasks. These transitions can be defined explicitly as in gSPM [6, 10] or implicitly as in finite-state machines like Markov models [3, 4, 5, 15, 16, 19, 22, 29, 31]. Sequential lists are appropriate for modelling linear surgical workflows [7, 14, 18, 20, 23, 24, 25, 26, 27, 33]. Some works also used unsequential lists with no fixed temporal order between the surgical tasks [1, 30, 32, 34].

Figure 2.14: Modelling formalizations

2.5.4 Recognition

Features. Features are the relevant characteristics of the data extracted to describe the surgical tasks and enable their distinction. In classical approaches, extracted features are hand-crafted; it means that they should be defined by human in a supervised way. Let us first speak about low-level features that can be directly extracted from the acquired data by means of standard image analysis or signal processing. The video, one of the most used signals, can be described using visual features. Static features characterize one frame only in terms of color [3, 4, 6, 14, 15, 16, 17, 23, 24, 32], shape [6, 14, 15, 16] and texture [6, 14, 15, 16, 24] in a global or local manner. More sophisticated detectors [16, 17, 23, 26, 33] describe objects and forms. Spatio-temporal features capture the difference between frames caused by object or background motion [7, 20, 21, 24, 25, 26, 29, 30]. The motion data from tracking and locating devices may be used in its raw [1], derived [28, 33, 34] or transformed forms [19, 20, 28, 33]. Deep-learning approaches, unlike classical ones, automatically extract low-level features in an unsupervised manner without human involvement [31].

The second category is high-level features that are obtained using some specific reasoning. For instance, the data from tracking or locating systems can be processed to semantically describe spatial relations between objects (e.g., near or far) [11, 33] or uniquely identify the position of someone or something in space (e.g., zone in the OR) [2, 27]. The instrument-related data can be described in form of binary vectors [4, 5, 22] indicating for each instrument its state (1 - in use, 0 - not) or with some other semantic [2, 16, 19] or quantitative representations [18, 27, 33] containing various useful information about the surgical tools. The high-level category also includes features as duration, count or occurrence of certain events (e.g., time that the surgeon passed looking into the console, number of robotic clutches, number of coagulations, etc.) [18, 33, 34]. Finally, manually annotated activities in turn can be transformed in semantic descriptions [8, 9, 10, 12, 13] that is a sequence of tuples containing words for actions, instruments and anatomical structures.

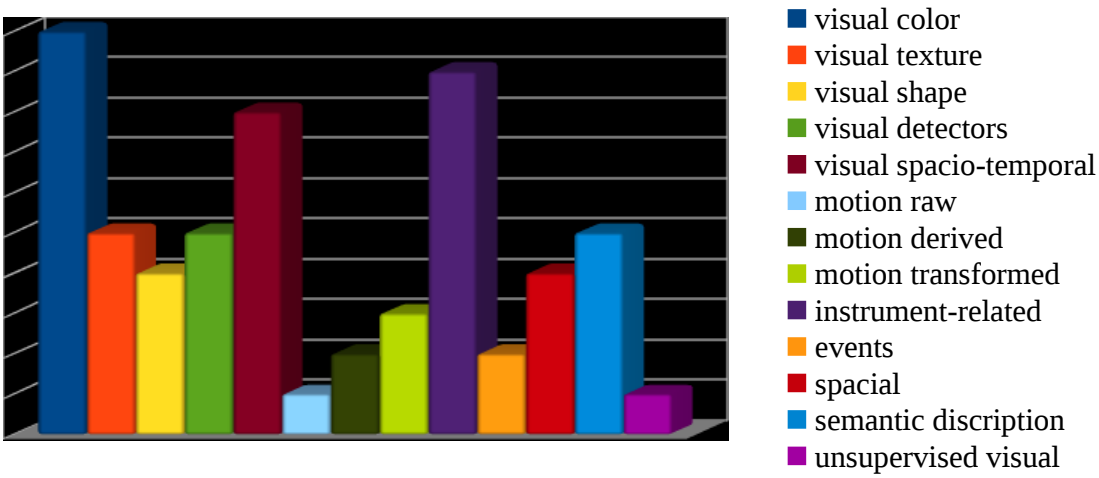

Figure 2.15: Features used to extract relevant information from the signal

Approach. This subcategory describes the methodological approaches and algorithms used to "teach" the system to recognize the surgical workflow from the provided data. This part is probably the most intricate and difficult aspect of CA-CAS creation. Most of researchers' attention is devoted to the process of development and adaptation of new recognition algorithms. All machine learning approaches are typically divided in three categories: supervised requiring full annotation of input data (i.e., each training sample should have a label), unsupervised finding relevant information in the input by its own (i.e., no labelling required) and semi-supervised when only a part of input data is labelled. The overall approach may be simple, meaning that only one machine learning algorithm is used to learn from the provided features, and complex consisting of several learning lev-

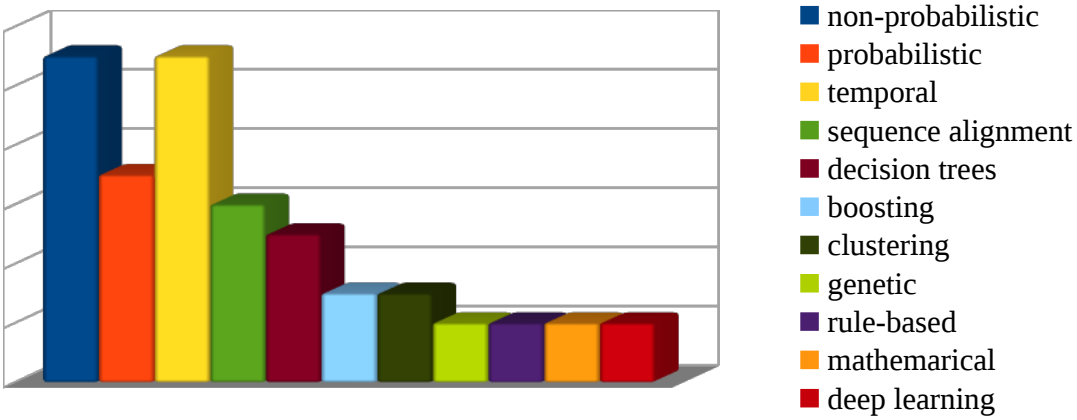

Figure 2.16: Distribution of methods used for recognition

els. In complex approaches output from one level is often used as input for the next one. A Random Forest performing task classification directly from the extracted features [27] is an example of a simple approach. The approach in [15] may be considered as complex: it first uses an SVM classifier to automatically derive semantic information from low-level features (e.g., presence-absence of certain objects, type of view, etc.), and then constructs a HMM to learn temporal relations between surgical tasks. Most of the reviewed works used much more complex multi-level approaches.

All presented approaches were divided into the following groups: probabilistic models, non-probabilistic classifiers, temporal models, sequence alignment algorithms, decisional trees, rule-based approaches, neural networks (including deep learning), clustering methods, boosting approaches, genetic algorithms and others. *Probabilistic models* (e.g., Naive Bayes, Bayesian network, Conditional Random Fields, etc.) apply Bayes' Theorem to expresses the conditional dependences between a set of random variables [8, 11, 18, 25, 28, 29, 32, 33]. *Non-probabilistic classifiers* include classical classifiers (e.g., Support Vector Machine and K-Nearest Neighbors) that do not involve probability theory and do not belong to any other group [1, 3, 14, 15, 16, 17, 18, 20, 23, 30, 31]. *Temporal models* (e.g., Hidden Markov Models and its derivatives) are used to model time-series representing stochastic processes [1, 3, 4, 5, 6, 10, 15, 16, 19, 21, 22, 31]. Most of the temporal models are technically probabilistic as well, but they were put in a separate category in order to highlight the difference between static and dynamic systems. *Sequence alignment algorithms* (e.g. Dynamic Time Warping) measure the similarity between two temporal sequences varying in length and calculate an optimal match between them [4, 7, 9, 16, 17, 22]. *Decision trees* (e.g. Random Forest, Decision Stamp, etc.) is a family of methods that use tree-like structures to create a model of decisions based on actual values of the features [2, 13, 18, 24, 27]. A *rule-based approach* is a general term for methods that identify, learn and apply rules that explain the observed relationships between the labels and the signal [11, 12]. *Neural networks* are the methods inspired by the structure and functioning of human brain. They include classic artificial neural networks and *deep learning* approaches (e.g., Convolutional Neural Networks and Recurrent Neural Networks) [18, 31]. *Clustering methods* (e.g. k-means, Aligned Cluster Analysis, etc.) divide input data on subsets called clusters in a way that the samples inside one cluster are similar and the samples from distinct clusters are dissimilar [20, 30, 34]. A *boosting approach* uses a great number of weak learners assembled into a cascade to form one strong classifier [6, 16, 17]. *Genetic algorithms* (e.g., cultural optimization and evolutionary reinforcement learning) represent search heuristics that mimic the process of natural selection, and use methods such as mutation and crossover to generate new genotypes in order to find good solutions for solving the problem [13, 14, 32]. Other algorithms mostly operate with mathematical concepts (e.g., polynomials) [24, 26].

Running mode. This subcategory defines the system's period of functioning and recognition speed. Depending on the mode and targeted application, the presented methods can be divided into on-line and off-line ones. The on-line methods have a potential to be used in the OR during surgery. Their recognition process is based on the information about the present moment and (optionally) the moments from the past but not from the future. The off-line methods, on the other hand, require to have a complete data sequence from the start till the the end of the surgery or at least the entire task to perform their analysis. Of course, the on-line methods have an advantage to be used for any application (e.g., assistance, decision support, device control and documentation), whereas the off-line methods suit post-treatment applications only (e.g., documentation, video indexing, etc.). However, generally, the off-line approaches perform better. Nevertheless, if a method is theoretically on-line, technically it does not mean that its implementation runs in real time. Some complex methods without optimization take too much time to process, and can not be applied in the OR as they are. There were also papers that proposed implementations working in near real time. This means that a small but often tolerable delay in delivering recognition decision still exists. Table 2.1 indicates a running mode for the publications that reported it. The methods [7, 24, 25, 26], marked as semi on-line enabled a recognition compatible with real-time computing but the decisions were provided only at the end of each task. It worth noticing that a real-time implementation is an important trump. The papers that did not make an appropriate statement have "n/a" in this field but should rather be considered as non real-time or off-line methods.

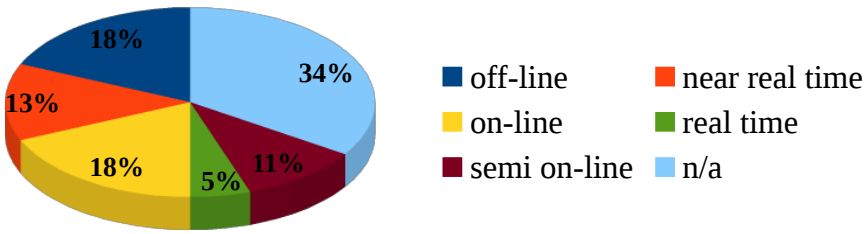

Figure 2.17: Repartition of methods by running mode

2.5.5 Validation

Protocol. Protocol defines the organization of the validation study performed after the training process. Two main validation protocols are train-test and cross-validation. A train-test is a protocol in which all available data is divided into two distinct groups and any given intervention belongs to one group only: a training set - the sequences that are used to train the system, and a testing set - the sequences used to assess the recognition ability of the system. The testing set allows to evaluate the generalization ability of the method and to see how well it performs on new data which has not been seen during training. This protocol is substantially applied for validation when the overall dataset is quite large.

A cross-validation is a protocol in which a complete dataset is first split on n equal parts, called folds. Then, the interventions from all folds except one are used for training and one fold left for validation. At the next round, the fold earlier used for validation is now used for training along with the others, and an another fold is chosen for validation. This process is then repeated n times in a way that every fold is selected for validation only once. In almost all the publications n equalled the number of interventions in the dataset; this is called a full cross or leave-one-out validation. This protocol is more often applied to small datasets since a train-test partitioning would leave the method with few data to learn from, and not enough to objectively test its performance.

Both of these protocols, at least in the way they were used in the publications, have hidden bias. Almost every method has parameters that have be tuned to provide better results. Unfortunately, all the data, including the sequences used for validation, in one way or another, were involved in the process of tuning. It means that the provided results probably overestimated the real performance. If the method, in its final form, was tested on some completely new data (not involved in the overall process of system creation at all), it would show lower results. The most objective protocol is train-validation-test: a train set is for training, a validation set for finding the best model or the best parameters, and a test set for making the final evaluation. However, none of the published methods used it, presumably, due to small dataset sizes. This is one of the validation biases.

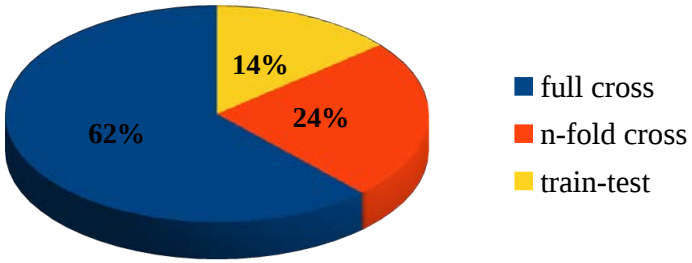

Figure 2.18: Validation protocols

Metrics. Metrics define the measurements used to assess the performance of the recognition system during validation. The most commonly used metrics were overall accuracy and per-class/average recall and precision. Accuracy, also called recognition rate or correct classification rate, is the percentage of correctly recognized samples in a sequence or a dataset. Recall, otherwise known as sensitivity, correct positives or true positive rate (TPR), is computed as the number of true positives divided by the total number of actual positives. Precision is the number of true positives divided by the sum of true and false positives. F-score representing a harmonic mean of precision and recall was also used in some papers [8, 9, 33].

Another frequent way to validate recognition methods was using a ROC curve and computing its AUC (area under the curve) [5, 24, 25, 26]. The ROC curve is a plot of true positive rate (TPR) against false positive rate (FPR) at various thresholds. FPR, also known as fall-out, is the number of false positives divided by the sum of false positives and true negatives, or 1 - specificity [1, 26]. Specificity or true negative rate (TNR) is the number of true negatives divided by the sum of true negatives and false positives [1]. False negative rate (FNR) or miss rate, computed as 1 - sensitivity, was also used [1].

Some more rare yet standard measures include failure rate (i.e., 1 - accuracy [22, 33]), negative predictive value (i.e., the number of true negatives by the total number of actual negatives) [22], variance of recognition rate [13] and Levenshtein distance [18]. Two papers [4, 26] performed statistical tests and computed p-values. The paper [1], performing pattern mining used some interesting metrics as motif seed diversity (MSD), motif seed hit rate (MSHR) and motif seed purity (MSP). MSD indicates how many of the surgical tasks were recognized. MSHR shows how many of the relevant tasks (i.e., important for CA-CAS) were recognized. Finally, MSP measures how much each learned motif is based on the data from one single task rather then their mixture.

Among uncommon metrics were the number of correctly identified tasks within a temporal window [31], the number of detected transitions between tasks [23], and the average transitional delay (i.e., a temporal delay between the moment when a transition is detected and actually occurred) [6, 23]. Forestier et al. in [9] also measured a prediction confidence of their method and counted the occasions when no conclusive decision has been made. The paper [6] recomputed accuracy, precision and recall allowing a small tolerant delay in recognition.

A confusion matrix, which is not an actual metric, was often used to display the distributions of correct and false recognitions between the surgical tasks [2, 4, 6, 8, 27, 30, 32]. Some publications also proposed a visual comparison of recognized sequences and the ground truth [6, 9, 31, 34].

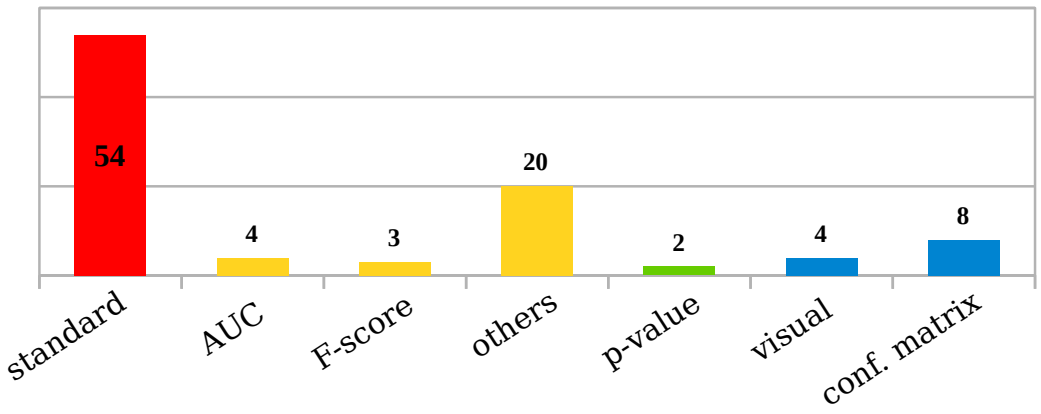

Figure 2.19: Distribution of metrics used for validation

Tests on public datasets. This subcategory indicates whether the recognition method was also tested on other types or sources of data that has not been initially considered as targeted. The comparison of the results is a very biased procedure if the methods were designed and validated on distinct datasets having different levels of complexity. The results may only be compared when they were obtained on common data. Unfortunately, only few authors validated their approaches on publicly available datasets. Droueche et al. [7] and Quellec et al. [24] tested their methods initially designed for eye surgeries on HOLLYWOOD2 human action dataset [Marszałek 2009] in order to demonstrate the generalization ability of the proposed approaches. This shows that the same method is able to suit other contexts, possibly different surgeries. Only three surgical datasets used in the reviewed publications were publicly released: set of cataract surgeries CATARACTS used in [7, 24, 25, 26] and two datasets containing laparoscopic cholecystectomies Cholec80 [Twinanda 2017] used in [31] and EndoVis2015 [Stauder 2016, Twinanda 2017] in [6, 31].

Evaluation. Evaluation is assessing the acceptance of the system by the end-users and its value [Jannin 2006a]. It is the last stage when the developed system is finally applied and assessed in real conditions. This helps to receive clinicians' feedback after its use and to measure its impact on the surgical process. The integration and tests require a significant effort and may take time. That is why, only Katić et al. in [11] actually got to this stage by making preliminary tests with clinicians yet still in simulated conditions. They evaluated the medical usability of the developed tool for augmented reality visualization and guidance by proposing questionnaires to the surgeons.

2.6 Discussion, conclusion and thesis positioning

The performed analysis of the works published in the domain allowed to shed light on some serious issues slowing down appearance and integration of context-aware systems.

First of all, it is obvious that despite all the advancements made through the years, there is still too few research on surgical activities (see Section 2.5.3). Less than 10% of the reviewed studies focused on low-level clinical information. Yet, this detailed level of workflow observation allows things that others do not (e.g. in-depth understanding of surgical process and adverse events detection). On the other hand, a delay in development of this direction may be explained by the fact that among other granularity levels the activities represent the biggest challenge for the recognition. Unlike states, phases or steps, they generally have much shorter durations, higher diversity resulting in hundreds of distinct activities to recognize, and more complex workflows represented by a multitude of possible ways to order and perform the activities. It is also harder for a machine to make a difference between activities based on a raw signal only. An abstraction from the physical signal should be made in favour of a semantic description of the scene that would give sense to the situation (e.g., the name of action that the surgeon is performing at the mo-

ment, the operated anatomical structure or the instrument used to accomplish the action). However, no study has yet been done to analyse the relevance of different semantic information to the activity recognition process. Yet, a good choice of description would help to improve and facilitate the recognition. This knowledge would also provide an insight on sensing devices that should be installed in the OR, as well as guidelines for researchers.

The second problem that has to be resolved is a lack of training data. Today, deep learning greatly overcomes classical machine learning methods in many domains. Although only two reviewed papers applied it in their research, it rapidly conquers medical field. It is now widely used for brain segmentation [Havaei 2017, Zhang 2015] and cancer detection [Cruz-Roa 2013, Wang 2016]. Almost all the participants of the MICCAI 2016 M2CAI challenge on workflow recognition presented deep learning approaches. For a successful functioning, however, deep learning requires enormous amounts of training data. Yet, on average for the reviewed studies on activity recognition, the training dataset contained 17 recorded interventions only. Generally, the following solutions can be applied for getting more data: using data available on the internet, acquisition of new data, data generation or using alternative methods. Several datasets for phase (e.g., Cholec80, EndoVis2015, CATARACTS) and gesture (e.g., JIGSAWS, MISTIC) recognition exist. Yet, no dataset containing surgical activities is freely available. The acquisition of clinical data is an intricate process requiring an ethical approval, patient consent, sensors installation and time. One third of publications was made using simulated interventions (e.g., mock-up OR, phantom or artificial generation). This simulated data, however, can be used for testing new recognition methods only but is not enough to realistically assess the method's performance in real clinical conditions. Thus, exploring alternative techniques, knowledge transfer for example, is highly relevant.

Finally, the third issue relates to the validation process. Currently, there is no common validation standard in the domain. Many studies neglected this process by applying not-suitable metrics and paying poor attention to exploration of strengths and weaknesses of their methods. Standard machine learning metrics as accuracy, recall and precision, most frequently used by the authors, are not sufficient to give an exhaustive in-depth prospect of the obtained results. These scores conform to problems as data retrieval or image classification but not to surgical workflow recognition. To begin with, they were designed for non-sequential data to report the amount and type of occurred errors but not to tell much about how the errors are distributed in the sequence or why they occur. Yet, this is a highly important information for a CA-CAS application. The consistency of detection and robustness of the method, rarely measured in the papers, are also essential. Another problem with the current validation process is its disconnection from the actual application. The requirements of performance vary depending on clinical use. For certain applications, a correct recognition of only a portion of tasks is necessary. Some applications may easily tolerate a delay in detection. This ignorance of the clinical objective leads to insufficient validation process and slows down the integration and use of the system.

In conclusion, this chapter presented a review of existing methods for surgical workflow recognition and provided an analysis of the obstacles creating a blockage in the field. As a result, three main problems were emphasized: 1) difficulties in recognition of semantic surgical activities, 2) deficiency of clinical data for training, and 3) inappropriate validation process. The purpose of this thesis was to propose solutions to these problems. Therefore,

— Chapter 3 presents the clinical data used in this work.

— Chapter 4 presents an approach for analysis of semantic components of the activity. It also discusses pertinent results of the performed analysis suggesting essential elements necessary for a high-quality recognition and provides recommendations for a wise choice of sensors for the OR.

— Chapter 5 examines methods of knowledge transfer to enlarge learning dataset and proposes techniques allowing to improve recognition results.

— Chapter 6 offers a fresh vision on the validation process along with strategies and new metrics adapted for surgical workflow recognition.

— Chapter 7 concludes the work and discusses its perspectives.

Part II

Knowledge support for context-awareness

CHAPTER

3 Data presentation

Preamble

In this chapter, we present a detailed description and analysis of clinical data used in the studies conducted within this work. First, we explain the purpose and main steps of four available surgical procedures. Then, we describe datasets composed of these procedures giving a description and examples of surgical phases, activities and activity elements (i.e., verbs, instruments, and structures). Finally, we discuss the variability, particular features and interconnections of the data.

Contents

Various types of data (e.g., simulated or artificially generated) can be used for solving problems of workflow recognition. Clinical data from real surgical procedures performed in the OR is, however, eventually required for a more realistic assessment of the recognition method. The present work was based on the previously recorded surgeries of anterior cervical discectomy and fusion (ACDF) [Forestier 2013], lumbar disc herniation (LDH) [Riffaud 2010], pituitary adenoma (PA) [Lalys 2010] and cataract surgery (CS) [Lalys 2013]. This data represents workflow diversity, variety of practices and expertise levels. Its use contributes to a better understanding and generalization of the approaches and discoveries from this work. A description of the procedures and detailed information about the composed datasets are presented below.

3.1 Surgical procedures

3.1.1 Anterior Cervical Discectomy and Fusion

Anterior cervical discectomy and fusion (ACDF), also called anterior cervical decompression, is a neck surgery during which a damaged intervertebral disc pinched between two backbones is removed to relive the pain caused by an excessive pressure on the spinal cord or nerve root. ACDF may be done for one or several discs of the cervical spine at once. The procedure is conducted as follows. First, during the first phase - approach to the spine, a three to five centimetres skin incision is made on one side of the neck. The incision is usually horizontal unless a multilevel operation is performed. Muscles between the skin and the pre-vertebral fascia are then split to get access to the spine. The fascia, fibrous tissue covering the spine, is then dissected away from the disc space. Next, a discectomy (i.e., disc removal) is performed by cutting a fibrous ring around the disc and extracting its soft inner core. The posterior longitudinal ligament and portions of the lower vertebral

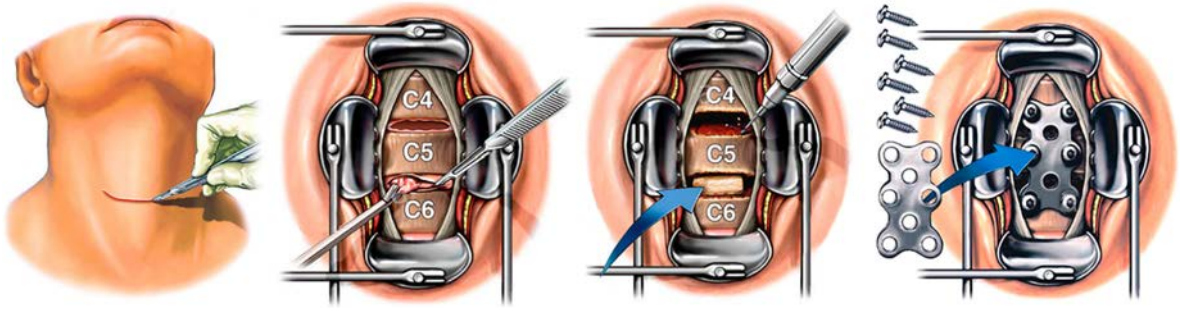

Figure 3.1: Anterior cervical discectomy and fusion
Source: http://www.alamy.com Author: Nucleus Medical Art Inc.

bone are also typically partially removed to evacuate any leftovers of the disc material and to relieve the spinal cord or nerve root compression. A cervical fusion, called arthrodesis, is done after the discectomy. A bone graft is inserted into the liberated disc space in order to prevent its collapse and to allow two vertebrae to grow together into a single unit. This helps to maintain the decompression by giving enough room for the nerve roots and spinal cord. To provide additional stability across the disc space, a small metallic plate is attached with screws to each vertebra. Finally, the surgical incision is closed.

3.1.2 Lumbar Disc Herniation

Posterior lumbar discectomy is the most common surgery used to remove a herniated portion of a disc in the lumbar region of the spine pressing on the nerve root and the spinal cord. It helps to relieve the back and leg pain, called sciatica, caused by pinched spinal nerves by giving more space to the nerve root. A small incision is made posterior down the middle of the back over the affected vertebrae to start the procedure. Then, back muscles are lifted off the bony arch of the spine and moved to one side in order to expose the lamina - a bone at the back side of the spinal canal. During surgery, the back muscles are held to the side with retractors with no need to be cut. Next, a small opening of the affected lamina, above and below the spinal nerve, is made with a drill or bone-biting tools to access the herniated disc. This is called laminectomy or laminotomy. It is performed on one or both sides, or on multiple vertebrae levels. Then, a protective sac of the nerve root is gently retracted and moved to the side. Using small instruments, the surgeon goes under the nerve root and removes a ruptured portion of the disc to decompress the spinal cord. Finally, the muscles are moved back into place by removing retractors, and the muscle and skin incisions are sewn together.

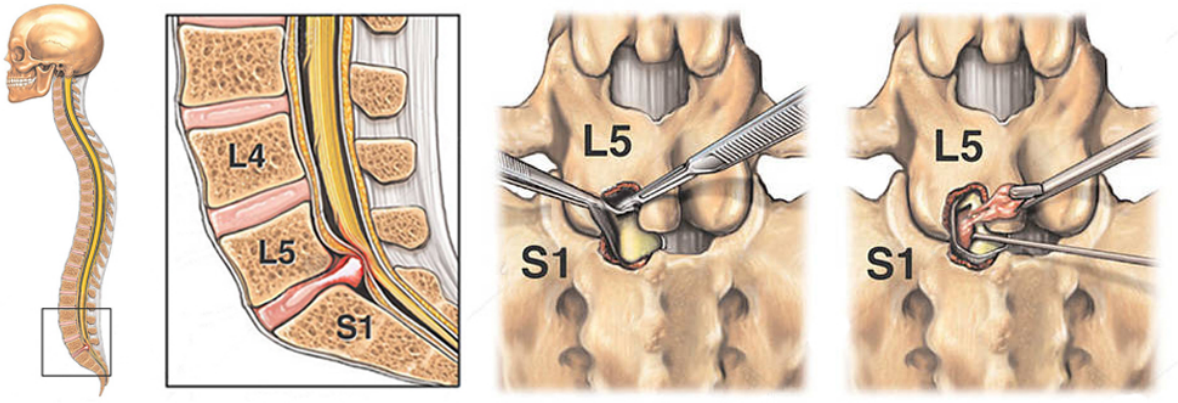

Figure 3.2: Lumbar disc herniation
Source: http://www.alamy.com Author: Nucleus Medical Art Inc.

3.1.3 Pituitary Adenoma

Pituitary adenoma is a benign tumour which arises from the pituitary gland and can cause vision loss and hormone problems. Pituitary tumours are typically removed transsphenoidally. A transsphenoidal surgery is a type of surgery in which an endoscope and long surgical instruments are inserted into the sphenoidal sinus cavity (i.e., air space behind the nose) by going through the nose and the sphenoid bone in order to remove tumours from the pituitary gland and the skull base. To begin with, the surgeon passes an endoscopic camera through the nostril. A small portion of the nasal septum is removed, and the front wall of the spheniod sinus is opened using bone-biting instruments. After that, the bone of the sella (i.e., a bony cavity in the skull base where the pituitary gland is located) is removed to expose the dura - a thin lining of the skull. The opening of the dura enables access to the tumour and the gland. The tumour is usually cut into small pieces and removed with special surgical instruments called curettes. When the visible tumour is extracted, the surgeon advances the endoscope into the sella to control that no hidden tumour is left. The surgeon may also use an x-ray to control the position and removed amount of the tumour. If needed, a fat graft (from the abdomen) is used to fill the empty space left by the tumour. At the end of the surgery, the holes in the sella and sphenoid sinus are sealed with biologic glue and bone grafts from the septum. The glue prevents the cerebrospinal fluid from leaking into the sinus and nasal cavity.

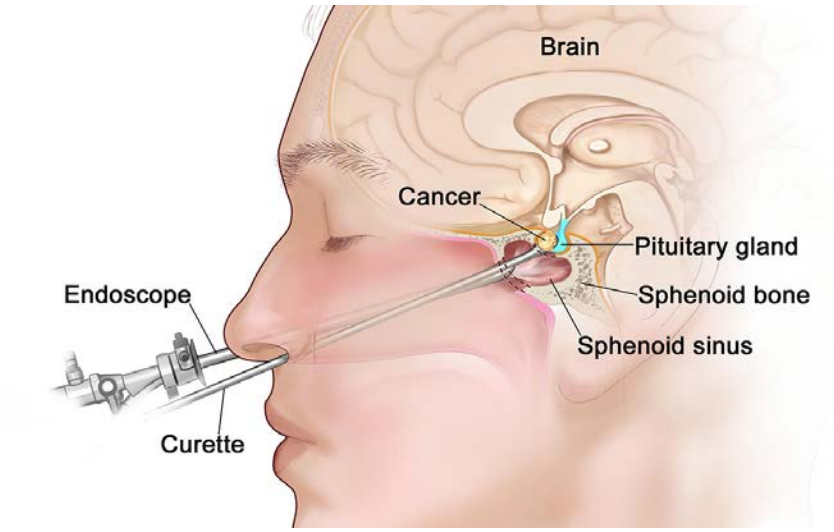

Figure 3.3: Pituitary Adenoma
Source: Terese Winslow LLC http://www.teresewinslow.com

3.1.4 Cataract surgery

Cataract surgery serves to replace a cloudy natural lens inside the eye with an artificial lens to improve blurry vision. During the procedure, a tiny incision is made to access the lens. A high-frequency ultrasound device is then used to fraction the lens into small pieces, which are then gently removed from the eye with suction. This procedure is called phacoemulsification. After all residues of the natural lens have been removed, the surgeon inserts in the eye a folded clear intra-ocular lens (IOL). It is then securely positioned and adjusted behind the iris and pupil, occupying the location of the natural lens. The cataract removal and IOL implantation is finally completed by closing the incision. A protective shield is placed over the eye to keep it safe in the early stages of recovery.

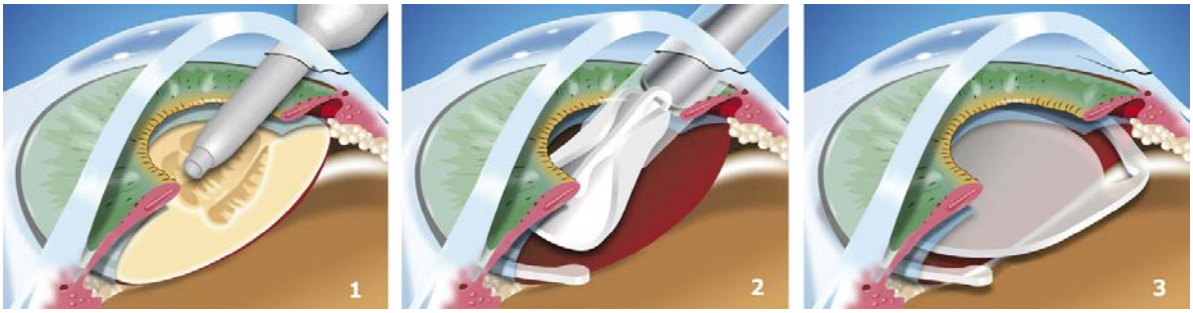

Figure 3.4: Cataract surgery
Source: http://www.ranelle.com/cataract-surgery

3.2 Datasets

All the data used in this thesis was organized in seven datasets by procedure and hospital. The data from neurosurgeries (i.e., ACDF, LDH and PA) was collected at two university hospitals: Rennes (France) and Leipzig (Germany). The ophthalmological surgery of cataract was recorded at the University hospital of Munich (Germany). A total of 154 interventions were acquired. Table 3.1 presents information about the number and average duration of conducted procedures, as well as the number of involved leading surgeons. All the procedures were performed by both senior and junior surgeons. The senior neurosurgeons were considered those who had performed more than 100 interventions of the given type. The juniors were residents who had completed more than two years of their residence program. All surgeons were right-handed.

For all studies presented in this thesis, only manual annotations of workflow were used, i.e. surgical phases and surgical activities. Sections 3.2.1, 3.2.2 and 3.2.3 describe existing phases, activities and activity elements (i.e., verbs, instruments and structures) respectively. The data was annotated using ICCAS Surgical Workflow Editor software developed in Leipzig [Neumuth 2006a, Neumuth 2007]. The program enables recording of individual

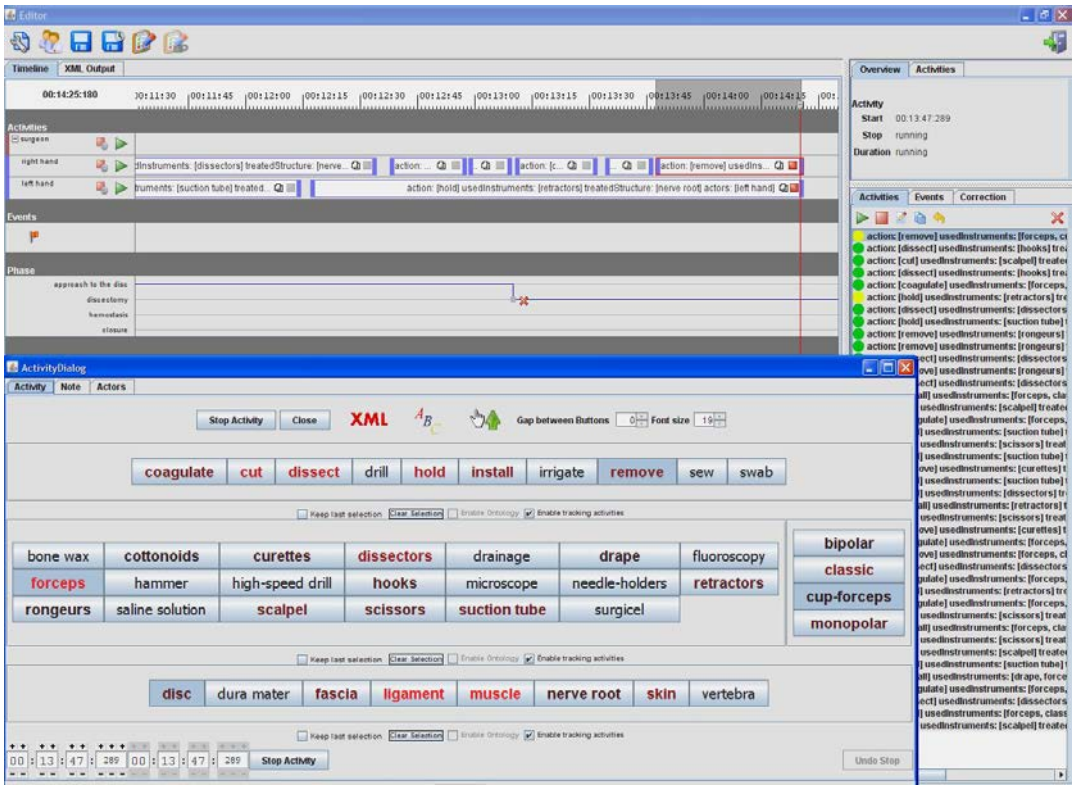

Figure 3.5: ICCAS Surgical Workflow Editor
Source: [Riffaud 2010]

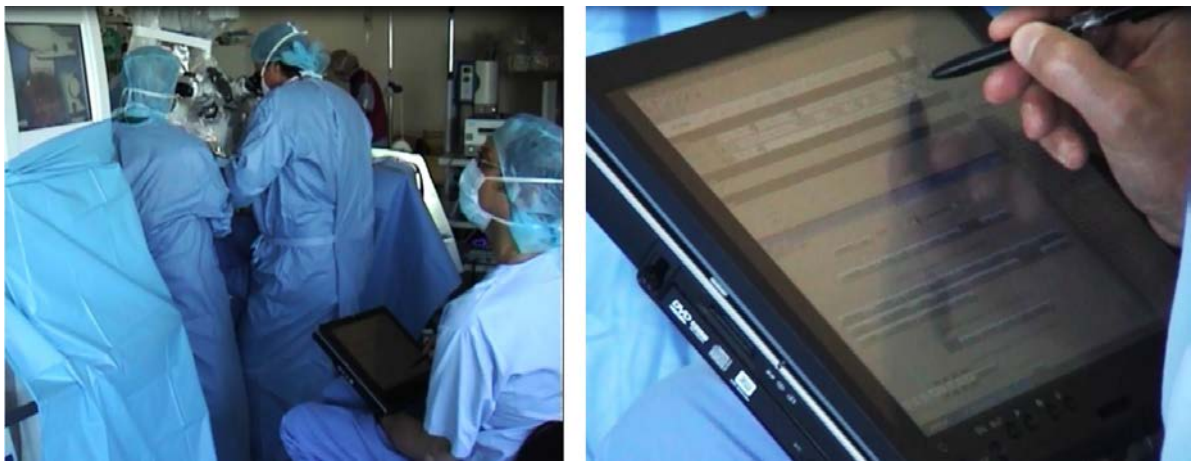

Figure 3.6: Real-time recording of surgical procedures
Source: [Riffaud 2010]

Table 3.1: General information about the datasets

Surgery	ACDF		LDH		PA		CS
Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
Nb. of procedures	16	48	25	20	15	11	19
Nb. of surgeons	4	5	6	5	2	1	2
Duration (min)	156±61	85±26	80±26	38±14	78±21	58±22	12±3

surgical process models containing several operating actors and manipulators (see Figure 3.5). The software was validated on neurosurgeries, ophthalmology, hear, ear, noise and throat surgeries performed in Leipzig. It also offers instant statistics about the surgical process such as number and duration of use of each instrument, number of repeated gestures and others. The neurosurgeries were annotated by the same senior surgeon in real time as shown in Figure 3.6. The cataract surgeries were annotated by one PhD student from video recordings. Both annotators were carefully trained on the annotation software beforehand. Five first annotations of each procedure were considered as tests and were not taken into account. All the annotations were carefully reviewed with the same software afterwards.

3.2.1 Phases

As defined in 2.5.3, a surgical phase is a period of time during which several associated surgical objectives are attained by the surgeon and other members of the team. Each procedure has its own set of m possible phases $P = \{p_1, p_2, ..., p_m\}$. Each intervention is annotated as a sequence of n phases $Seq = (a_1, a_2, ..., a_n)$, where a phase annotation $a_i = < p, t_{start}, t_{end} >$ is 3-tuple containing its name $p \in P$, start time $t_{start} \in \mathbb{N}$ and end time $t_{end} \in \mathbb{N}$ in seconds, e.g. *(discectomy, 1525, 2576)*. The phase annotations also satisfy the following conditions: $t_{start_i} < t_{end_i}$, $t_{end_i} < t_{start_{i+1}}$, $t_{start_1} = 0$, and t_{end_n} equals the duration of a given intervention.

The phases of every surgery are listed in Table 3.2. Their average duration, number and order vary from one intervention to another. Cataract surgery, for example, consists of more but much shorter phases that occur in a linear order. Neurosurgical procedures, on the other hand, may have come-backs and repetitions in the phase order. This is often the case for multi-level ACDF and LDH surgeries. The hemostasis may also happen several times between other phases if necessary, or can be skipped when the patient has poor bleeding. The neurosurgical phases are common for both operating sites (i.e., Leipzig and Rennes), except the PA phase “fat graft" that took place at the hospital of Rennes only.

Table 3.2: Surgical phases

ACDF	LDH	PA	CS
1. Approach to the spine 2. Discectomy 3. Arthrodesis 4. Hemostasis 5. Closure	1. Approach to the disc 2. Discectomy 3. Hemostasis 4. Closure	1. Nose preparation 2. Approach to the dura 3. Tumour removal 4. Fat graft 5. Hemostasis 6. Reconstruction and closure	1. Betaisodona injection 2. Corneal incision 3. Visco-elastic injection and capsulorhexis 4. Phaco-emulsification 5. Cortical aspiration 6. IOL implantation 7. IOL adjustment

3.2.2 Activities

A surgical activity is defined as a physical activity performed by the surgeon expressed in an action verb, a surgical instrument and an anatomical structure. Each intervention is annotated as a sequence of l activities $Seq = (a_1, a_2, ..., a_l)$, where an activity annotation $a_i = < acr, bp, V, I, S, t_{start}, t_{end} >$ is a 7-tuple consisting of an actor act (here "surgeon", "assistant" or "nurse"), a body part bp (here "left hand" or "right hand"), a verb V, an instrument I, a structure S, start time $t_{start} \in \mathbb{N}$ and end time $t_{end} \in \mathbb{N}$, e.g. *(surgeon, left hand, hold, classic forceps, muscle, 1020, 1041)*. The activity annotations also satisfy the conditions: $t_{start_i} < t_{end_i}$ and $t_{start_i} \leqslant t_{start_{i+1}}$, as well as $t_{end_i} < t_{start_j}$ for $\forall\, i < j$ if $s_i = s_j$ and $h_i = h_j$.

In this thesis, the workflow was studied in terms of surgeon's activities only. Knowing the information about both surgeon's hands simultaneously was also important for a more complete understanding of the situation. Thus, for the purposes of this thesis, the definition of a surgical activity was changed to a tuple composed of six items: a verb, an instrument and a structure distinctly defined for both left and right hands of the surgeon at the same time. Such wise $a_i = < L(V, I, S), R(V, I, S) >$, where L and R correspond to the left and right hands respectively, e.g. *(hold, classic forceps, muscle, dissect, scissors, fascia, 1025, 1034)*. Using the start and end timestamps, each annotated intervention was transformed to a temporally ordered sequence of n 6-tuple activities $Seq = (a_1, a_2, ..., a_n)$. Finally, a set $\alpha = \{A_1, A_2, ..., A_m\}$ of all possible activities was constructed for each surgery from its annotations.

Table 3.3 displays the number of unique activity tuples for each dataset. The SPMs of given datasets differ greatly in terms of activities. Cataract, being the shortest and the most standardized surgery, has the smallest number of activities that occur in the same order each time. The number and sequencing of neurosurgical activities, contrariwise, notably vary between the procedures, sites and patients. As displaying here a synthesis of all surgical scenarios is infeasible due to their great variability, table 3.4 gives examples of activities for each procedure.

Table 3.3: Number of unique activity tuples for each dataset

Surgery	ACDF		LDH		PA		CS
Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
Both hands	377	379	413	243	282	255	45
Left hand	49	33	47	20	16	31	17
Right hand	93	111	116	74	103	92	16

Table 3.4: Examples of activities

	left hand			right hand		
	verb	instrument	structure	verb	instrument	structure
ACDF	suck blood	suction tube	disc	drill	high-speed drill	disc
	remove	rongeurs	disc	suck blood	suction tube	disc
	hold	classic forceps	muscle	sew	needle-holders	muscle
	hold	retractors	muscle	dissect	dissectors	fascia
	suck blood	suction tube	ligament	install	arthrodesis	disc
LDH	suck blood	suction tube	ligament	dissect	hooks	nerve root
	suck blood	suction tube	muscle	coagulate	bipolar forceps	muscle
	hold	retractors	nerve root	remove	cup-forceps	disc
	hold	classic forceps	skin	sew	needle-holders	skin
	suck blood	suction tube	disc	cut	hammer	vertebra
PA	suck blood	suction tube	tumour	remove	curettes	tumour
	hold	retractors	mucosa	dissect	dissectors	nasal septum
	suck blood	suction tube	sphenoid	remove	rongeurs	sphenoid
	suck blood	suction tube	sphenoid	cut	scissors	dura mater
	hold	retractors	mucosa	install	cottonoids	nose
CS	hold	micro spatula	bulbus oculi	aspirate	aspiration cannula	lens
	none	none	none	inject	methocel	anterior chamber
	incise	1.4mm knife	cornea	hold	colibri tweezers	bulbus oculi
	place	reposition hooklet	lens	none	none	none
	none	none	none	phaco	chopper	lens

3.2.3 Activity elements

Activity elements are the main compounds of the surgical activity: the verb, the instrument and the structure. A particular value of an element, for example “cut" for the verb, is called an *instance*. Table 3.5 indicates the number of unique instances of each activity element for every dataset. The verb is generally represented by a single physical action of one hand. Although the instrument and the structure can be instantiated as a union of two physical objects simultaneously manipulated by one hand, e.g. classic forceps + cot-

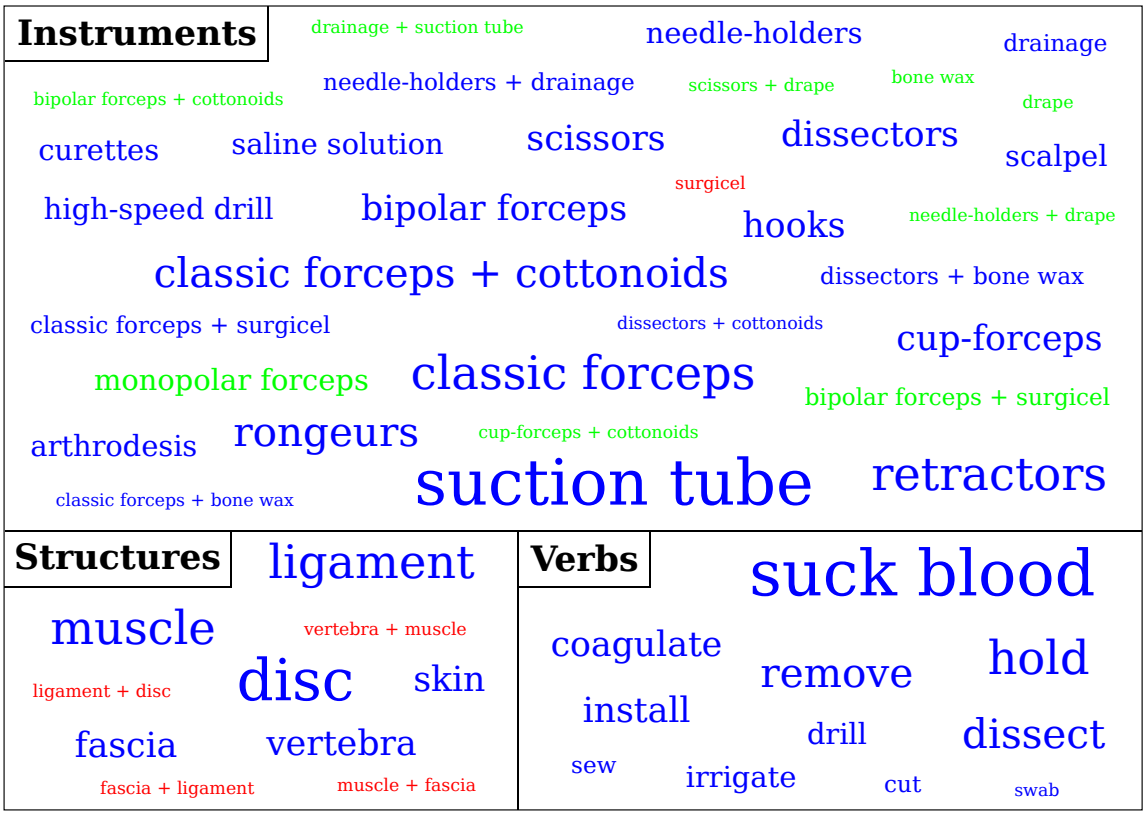

Figure 3.7: Instances of ACDF activity elements

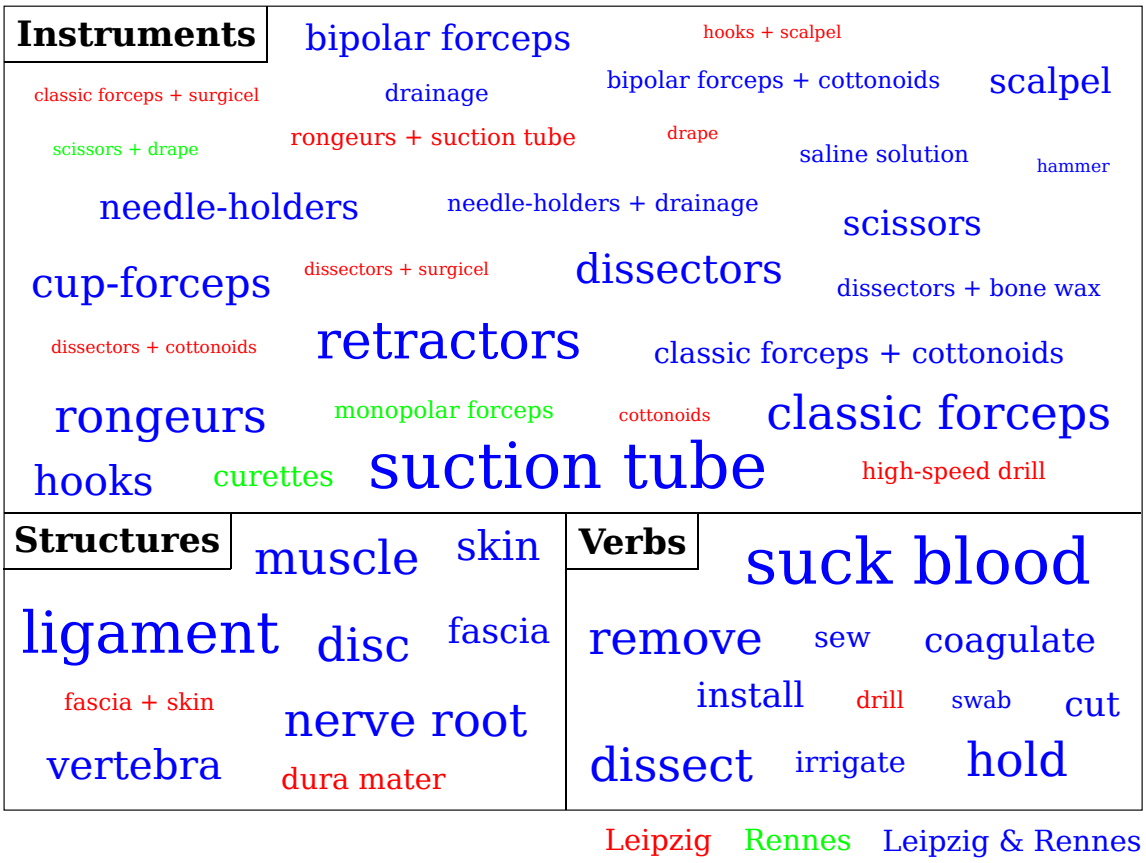

Figure 3.8: Instances of LDH activity elements

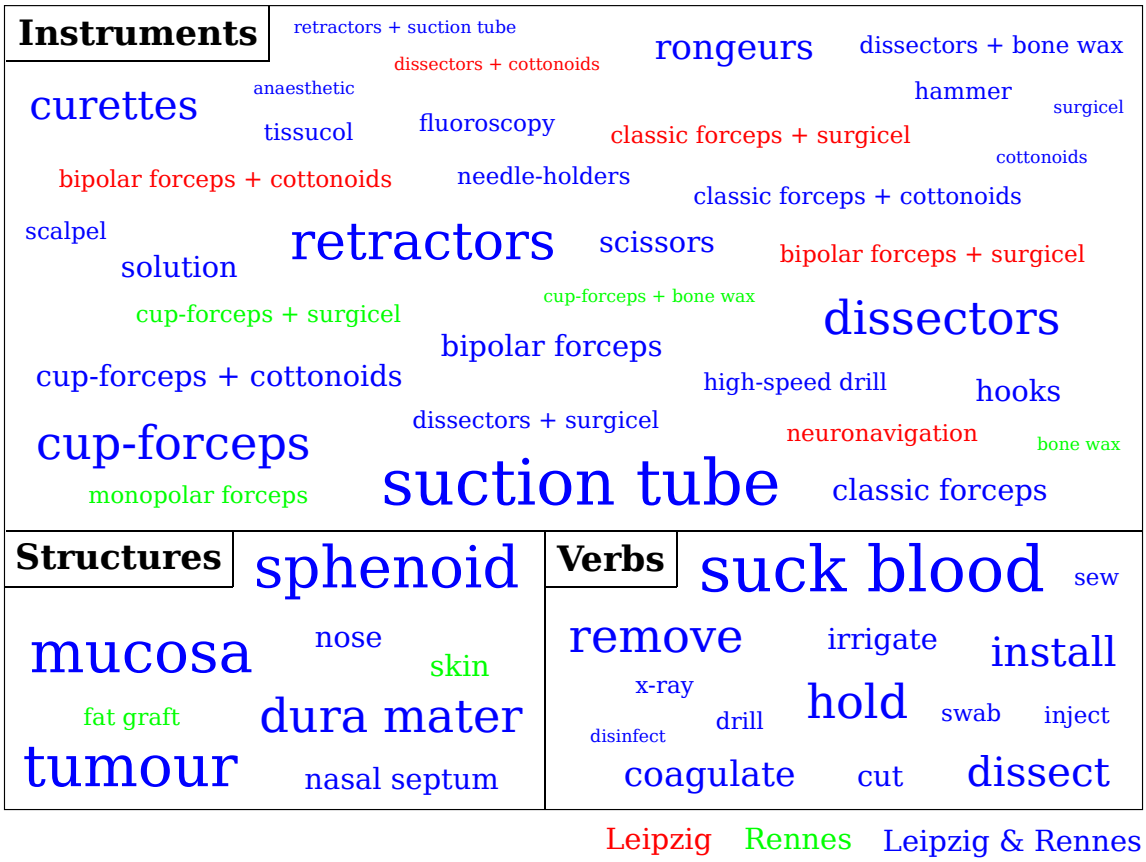

Figure 3.9: Instances of PA activity elements

verbs	instruments	structures
hold	irrigation cannula	bulbus oculi
aspirate	colibri tweezers	anterior chamber
irrigate	micro spatula	lens
incise	aspiration cannula	cornea and conjunctiva
wash	syringe	cornea
place	1.4mm knife	conjunctiva
inject	methocel	
phacoemulsification	reposition hooklet	
cut	1.1mm knife	
implant	iol	
desinfect	chopper	
	wecker scissors	
	betaisodona	
	sauter cannula	

Figure 3.10: Instances of CS activity elements

tonoids as instrument, or cornea + conjunctiva as anatomical structure. The union of two objects, counted as a unique instance, helps more clearly describe the activity and its surgical meaning (e.g., additional use of cottonoids indicates a higher blood loss). One or several elements may also be absent from the tuple, for example, if only one hand is involved in the activity or an action is performed with a bare hand using no particular instrument. In this case, the absent element is instantiated with the word “none". Figures 3.7, 3.8, 3.9 and 3.10 display all possible instances of each element (except the “none" word) for ACDF, LDH, PA and CS respectively.

Table 3.5: Number of unique element instances in each dataset

Surgery	ACDF		LDH		PA		CS
Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
Verbs	11	11	11	10	14	14	11
Instruments	24	31	26	22	30	29	14
Structures	10	6	9	7	6	8	6

3.3 Analysis of data variability

The previous section gave formal definitions of phases, activities, and activity elements, the number of unique items in each category and their examples. This section will present the analysis of data variability, as well as intra- and inter-procedural relationships. Section 3.3.1 discusses the phases and the relationships between the surgical sites in terms of phase workflow in the context of each procedure. Section 3.3.2 dedicated to the activities first analyses their frequency within the entire dataset, then analyses them within an intervention, and finally discusses the relationships between the sites and procedures. Section 3.3.3 discusses the activity elements: first in terms of hands, then frequency and finally intra- and inter-procedural similarities.

3.3.1 Phases

In the ACDF procedure, 5 different phase workflows (i.e., different sequencing of phases) were observed in Leipzig, and 8 in Rennes. Four of these workflows, occurring in 90.1% of all operating cases, were observed in both sites. Table 3.6 shows other statistics about the ACDF procedure, such as the number of cases in which the hemostasis or arthrodesis phases were skipped, and multiple discectomies or arthrodesis were performed. The table also shows the number of cases in which the hemostasis phase followed the discectomy and arthrodesis phases. According to these statistics, the patients in

Table 3.6: Statistics on ACDF phases

	Leipzig		Rennes	
No hemostasis phase	6 cases	37.5%	17 cases	35.4%
No arthrodesis phase	1 case	2.1%	-	
Multiple discectomies	2 cases	12.5%	1 case	2.1%
Multiple arthrodesis	2 cases	12.5%	2 cases	4.2%
Discectomy -> Hemostasis	4 cases	25.0%	21 case	43.8%
Arthrodesis -> Hemostasis	7 cases	43.8%	12 cases	25.0%

Leipzig and Rennes had similar levels of bleeding, but the hemostatis phase was performed in different moments: mostly after the arthrodesis in Leipzig, and after the discectomy in Rennes. These numbers also make ACDF the most heterogeneous procedure in terms of phase workflow. This heterogeneity, however, can partially be explained by a higher number of cases in the dataset.

In the LDH procedure, 3 unique phase workflows were found in Leipzig dataset, and only 2 in Rennes dataset, which were also shared with Leipzig in 95.6% of cases. The hemostasis phase was absent in only 4 cases (16%) in Leipzig but in 16 cases (80%) in Rennes. The frequent absence of this phase in Rennes can explain a big difference in average duration of the procedure, which was twice shorter in Rennes. It also indicates, that the patients operated in Rennes tended to bleed less during this procedure. As for the discectomy phase, it was performed twice only in 2 cases (8%) in Leipzig, and never in Rennes.

In the PA procedure, all the interventions in Leipzig except one were performed within the same phase workflow, while 3 different workflows were observed in Rennes. The sites shared only one common workflow which was followed in 80.8% of all cases. The hemostasis phase was skipped in only one case (6.7%) in Leipzig and one (9.1%) in Rennes. In Leipzig the fat graft was never placed, whereas in Rennes this was done in 4 cases (36.4%).

In fact, the cataract surgery performed in Munich always followed the same workflow. Its recordings, however, had different workflows due to the fact that the interventions were annotated from videos afterwards. The actual first phase “preparation” was never recorded as no microscope had been used when performing it. The next phase “betaisodona injection” was also entirely recorded and annotated in 5 cases only (26.3%).

3.3.2 Activities

Table 3.3 displays the total number of unique activity tuples within each dataset. The frequency of occurrence, however, varies greatly for different tuples. This table thus does not reflect their frequency distribution. To have an idea of how much different activity tuples were actually observed most of the time, we computed the number of unique tuples covering 50% of all performed activities. To calculate this number, we first counted

all the activities performed within a dataset, and then multiplied their amount by 0.5 to obtain a rate r. Then, using a greedy algorithm, we sorted the tuples by their frequency and progressively counted them starting from the most frequent one until the sum of their frequencies attained r. The results of the computation are exposed in Table 3.7. It shows that in neurosurgeries, only 7 to 14% of all unique 6-tuples composed 50% of all performed activities. For the ACDF procedure, in percentage terms, this number was twice bigger in Leipzig than in Rennes. For the LDH and PA procedures, conversely, it was 1.3 times bigger in Rennes than in Leipzig. The activity tuples of cataract surgery had a distribution close to uniform since they all had similar frequencies (approximately 45-50% of unique tuples covered 50% of all performed activities).

Tables 3.3 and 3.7 show the statistics computed within the entire dataset mixing the activities from all interventions. Table 3.8, on the other hand, displays the statistics computed with an intervention. It shows the average number of performed activities per intervention, average number of unique tuples and the average number of repetitions of one tuple. For the neurosurgical procedures, the number of two-hand activities performed

Table 3.7: Number of unique activities that compose 50% of all occurrences

Surgery	ACDF		LDH		PA		CS
Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
Both hands	55 (14.6%)	27 (7.1%)	48 (11.6%)	36 (14.8%)	31 (11.0%)	36 (14.1%)	20 (44.4%)
Left hand	9 (18.4%)	5 (15.2%)	9 (19.1%)	8 (40.0%)	5 (31.3%)	10 (32.3%)	8 (47.1%)
Right hand	18 (19.4%)	21 (18.9%)	21 (18.1%)	20 (27.0%)	23 (22.3%)	25 (27.2%)	8 (50.0%)

Table 3.8: Average number of activities occurring during one intervention

	Surgery	ACDF		LDH		PA		CS
	Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
Both	In total	367±149	244±76	242±72	148±49	266±77	213±46	29±5
	Unique	79±17	63±8	62±9	52±10	56±8	66±9	19±2
	Repeats	5±10	4±9	4±9	3±6	5±12	3±6	1±1
Left	In total	38±25	23±9	21±8	15±6	13±4	19±5	13±3
	Unique	13±5	8±2	10±2	9±2	5±2	10±2	9±3
	Repeats	3±4	3±3	2±2	2±1	2±2	2±1	1±1
Right	In total	162±63	106±32	105±31	62±22	121±37	94±21	14±3
	Unique	36±5	38±4	36±5	29±5	38±4	38±5	9±3
	Repeats	4±6	3±3	3±4	2±2	3±4	2±2	1±1

within one intervention was bigger in Leipzig than in Rennes (1.5 times bigger for ACDF, 1.6 for LDH and 1.3 for PA). The number of unique 6-tuples was however 3 to 5 times smaller for neurosurgeries and twice smaller for CS. One 6-tuple repeated on average 2 to 5 times per intervention during neurosurgeries, whereas only one time during cataract surgery. It means that the CS activity workflow had almost no loops.

We also computed the number of shared 6-tuples for the neurosurgical datasets. The left part of Table 3.9, describing intra-procedural relationships, shows the number of tuples common for both sites of one procedure. Its right part, related to inter-procedural similarity, shows the number of shared tuples for three pairs of procedures (both sites mixed). Leipzig and Rennes shared the biggest number of activities within the LDH procedure (almost 38%), while the least within ACDF (31.5%). For the inter-procedural part, the interventions from both sites were first put together to compute the number of shared tuples between the procedures. There were tuples that were shared between the procedures from one site, but not between the sites within one procedure. This explains the number of common tuples for the ACDF x LDH pair which was bigger than the intra-procedural numbers (177 vs. 119 and 124). The PA procedure, being different from ACDF and LDH, shared with them only about 2% of tuples.

Table 3.9: Number of activity 6-tuples common for surgical procedures and sites

intra-procedural			inter-procedural		
ACDF	LDH	PA	ACDF x LDH	ACDF x PA	LDH x PA
119 (31.5%)	124 (37.8%)	94 (35.0%)	177 (30.3%)	8 (1.5%)	13 (2.7%)

3.3.3 Activity elements

Table 3.5 shows the number of unique instances per element together for the left and right hands. For the neurosurgeries, however, the right hand could make 1.5 times more actions (i.e., verbs) than the left hand, it could also manipulate 2.6 times more instruments. As for the anatomical structures, both hands could operate the same ones except two different structures in RCVA Leipzig dataset, and one in PA Leipzig. In CS, at the exception of one verb and one instrument, the left hand elements can have the same instances as the right hand.

For each element we computed the number of instances covering 50% of all occurrences as well. For the neurosurgeries, the half of the verbs and instruments of the left hand were represented by only one or two instances (see Table 3.10). The two most frequent verbs for the left hand were “suck blood” and “hold”. The most frequent instrument was thus “suction tube”, and the second position was shared between “retractors” (for PA Leipzig, PA Rennes and LDH Rennes datasets) and “classic forceps” (for RCVA Leipzig,

Table 3.10: Number of unique element instances that compose 50% of all element occurrences. V, I and S correspond to the verb, instrument and structure respectively, whereas L and R to the left and right hands

Surgery	ACDF		LDH		PA		CS
Site	Leipzig	Rennes	Leipzig	Rennes	Leipzig	Rennes	Munich
V (L)	2 (22.2%)	2 (28.6%)	1 (11.1%)	2 (40.0%)	1 (25.5%)	2 (33.3%)	5 (50.0%)
V (R)	4 (36.4%)	4 (36.4%)	5 (45.5%)	4 (40.0%)	4 (28.6%)	5 (35.7%)	5 (45.5%)
I (L)	2 (13.3%)	2 (18.2%)	1 (7.1%)	2 (33.3%)	1 (14.3%)	2 (18.2%)	6 (46.2%)
I (R)	8 (53.3%)	10 (32.3%)	8 (30.8%)	8 (36.4%)	9 (30.0%)	9 (32.1%)	6 (42.9%)
S (L)	3 (37.5%)	2 (33.3%)	3 (33.3%)	3 (42.9%)	2 (40.0%)	3 (37.5%)	3 (50.0%)
S (R)	4 (50.0%)	4 (66.7%)	4 (44.4%)	4 (57.1%)	3 (50.0%)	3 (37.5%)	3 (50.0%)

RCVA Rennes and LDH Leipzig datasets). Only 2 to 3 structures were operated by the left hand in 50% of time. Meanwhile, the right hand verbs had 4 to 5 instances on average, instruments 8 to 10, and structures 3 to 4. In CS, as with the activity tuples, each element instance occurred similar number of times for both hands, thus a half of instances covered 50% of time.

Figures 3.7, 3.8, 3.9 exhibit the intra-procedural relationships between the element instances, meaning the difference between two sites within one neurosurgical procedure. Shared and site-specific instances have distinct colours. Higher font size of the font signifies higher frequency of occurrence. The figures show that the sites shared the most part of the instances, especially the most frequent ones. Both sites shared all verbs, except “drill” present in LDH Leipzig but not in LDH Rennes. Within the ACDF procedure 68.8% of instrument instances were shared between the sites, 65.1% within LDH, and 73.5% within PA. The relatively frequent not shared instrument instances included “monopolar forceps” (present in ACDF Rennes but never used in Leipzig), “curettes” (used only in LDH Rennes) and “high-speed drill” (used in LDH Leipzig but not in LDH Rennes). As for the structures, the ACDF surgeries from Leipzig and Rennes shared 60%, LDH 77.8%, and PA all of them. Among the frequent ones could be found “dura mater” which was operated only in Leipzig during LDH.

The inter-procedural similarities and differences in terms of element instances are displayed via a colourful word cloud in Figure 3.11 (for neurosurgeries only). As before, the frequency of occurrence manifests through the font size. The procedures shared mostly the verb instances: all of them were common to ACDF and LDH, and 78.6% are were shared with PA. The PA verbs that did not exist in ACDF and LDH (i.e., disinfect, inject and x-ray) had much lower frequency than others. In terms of instrument instances, ACDF and LDH shared 65.7%, ACDF and PA 60.5%, and LDH and PA 57.9%. As well as with the verbs, all

Figure 3.11: Common instances of activity elements for neurosurgical procedures

the most frequent instrument instances were shared. As for the structure instances, ACDF and LDH share almost a half of the instances (which also happened to be the most frequent ones) but both had only “skin” in common with PA (which did not exceed 15% of all instances).

3.4 Conclusion

The surgeries described in this chapter are clinical procedures belonging to two different domains: neurosurgery and ophthalmology. They all vary in terms of duration, surgical workflow and hospital-specific practices. Despite their differences, the procedures from the same domain still have similarities in phases, activities and activity elements. Some surgeries resemble more than others (e.g., anterior cervical discectomy and fusion is much closer to lumbar disc herniation than to pituitary adenoma surgery). As part of this work, the presented data has three main advantages. First of all, representing real surgeries with complex worflows, it actually suits the goal of a context-aware system creation for the OR of the future. Secondly, the diversity of data (different domains and procedures) allows performing a more objective analysis of importance of the OR sensors to the activity recognition and providing generalizable results. Finally, the shared properties of the procedures enable transfer of knowledge that can help to resolve the problem of training data deficiency.

CHAPTER

4 Sensors for automatic surgical activity recognition

Preamble

In this chapter, we introduce the studies performed to discover which signals and sensors could facilitate automatic recognition of low-level surgical activities. We made the following hypothesis: the activity recognition does not require sensors for all three activity elements. A large multi-scale study on four different surgeries confirmed the hypothesis and revealed relevant sensors to place in the OR. A deeper analysis examining the influence of noise in data on the recognition was also conducted. Finally, several relevant observations about surgical practices were also made. These observations, discussed at the end of the chapter, help better understand the surgical process. The findings of this work provide cues for designing a new generation of operating rooms.

Contents

4.1 Introduction

Today, an overwhelming flow of new technologies and equipment threatens to overrun operating rooms, adding even more complexity to the surgical process. A large amount of research is focused on smart and situation-aware intra-operative assistance to help alleviate the surgeon's stress and facilitate procedures. Automatic recognition of surgical workflow represents a substantial part of it. As it was shown in 2.5.3, most research groups studied the recognition of surgical phases [Forestier 2015, Katić 2016, Twinanda 2017, Bodenstedt 2017] and steps [Bardram 2011, Twinanda 2015]. A large amount of research has also been devoted to recognition of surgical gestures [Haro 2012, Gao 2016a, DiPietro 2016] from pre-operating training sessions (e.g., on JIGSAWS or MISTIC datasets), which offer much lower granularity. Yet, few works exist studying the automatic recognition of low-level activities from true complex clinical procedures [Lalys 2013, Meißner 2014]. Their automatic recognition is not only useful for in-depth situation awareness, analysing semantic activities also enables better understanding, learning and teaching of surgical procedures [Forestier 2012]. Surgical skills can be objectively evaluated based on a sequence of performed actions [Riffaud 2010, Forestier 2013]. Several other applications include detection of deviations from a standard procedure flow [Bouarfa 2012, Huaulmé 2017b], accurate estimation of remaining time and resource management [Maktabi 2017].

Due to the lack of automatic recognition, most applications use manually annotated surgical activities, which is a terribly tedious and time-consuming process. However, the automatic recognition of surgical activities is an extremely challenging task. Unlike phases and steps, activities are of shorter duration (minutes *vs.* seconds) and higher diversity in

terms of number (dozen *vs.* hundreds of distinct items), execution order (simple sequencing *vs.* great multitude of possible paths) and surgeon/practice-specific characteristics.

To facilitate the recognition process, the approaches proposed in the literature break down the activity into its meaningful elements (e.g., verb, instrument, and structure) then proceed with one-by-one detection. The activity is then deduced from one or a combination of elements. The elements to be detected are chosen depending solely on available signals, without any analysis of their relevance. The instrument is often considered a good indicator of the on-going task [Kranzfelder 2011, Bouarfa 2012, Maktabi 2017], even though it has been shown to have multiple functions that vary depending on the situation and surgeon [Mehta 2002]. The verb, which provides pertinent information about the activity context, is difficult to recognize due to a high variability of action execution [Meißner 2014], and often requires additional sensors. The anatomical structure, on the other hand, can be recognized from usually available image-based signals [Lalys 2013], without extra sensors needing to be brought to the operating room. However, no study as yet exists justifying the choice of elements to detect.

In this work, we propose to approach the problem from the opposite direction, using the data from Chapter 3. Assuming information on all three elements is available, we assess their impact on the performance of low-level activity recognition with the aim of defining a minimum set of required sensors and signals. This is the first large-scale multi-site study of elements' importance to activity recognition conducted on complex clinical data. This work's unique contribution consists in its original approach to information analysis for the optimization of operating room sensors, as well as its study results.

4.2 Problem statement

We made the following hypothesis at the beginning of this study: surgical activity recognition does not require sensors for all three activity elements. In order to prove the hypothesis and assess the impact of each semantic element on the overall recognition process we have used the following scheme. If we want to know how well a whole activity can be recognized solely knowing the used instruments, we can apply a *"010010"* mask to the activity tuple, which gives us *(unknown, forceps, unknown, unknown, scalpel, unknown)* for the example above. The element is considered known for both left and right hands, as in practice the same type of sensor is needed to recognize both. We also take into consideration a temporal context, meaning N activities having taken place before. The same mask is also applied to the N previous activities, ensuring that only available information is involved in the analysis. The problem then consists in mapping a sequence of partially hidden tuples to a full tuple (i.e., masked with *"111111"*) of a current activity. This problem was resolved using a deep neural network, described in the next section. This was performed for all other *configurations*, meaning the solely known elements, as well as their combinations.

Finally, the neural network model of each configuration was tested on the activity recognition task, and their performances were compared, as described in Section 4.4.

The relevance of the temporal context has previously been discussed in [Forestier 2015, Maktabi 2017]. We found out that working with deep learning and a fairly small dataset requires a careful choice of parameters, N in our case, in order to enable an effective learning process. We will talk about our choice of this parameter, defined as a function of factors like dataset size, number of unique activities, average number of activities per intervention, and complexity level, in Section 4.5.

4.3 Deep neural network for analysis

Today, deep learning methods are successfully applied to many different problems, starting from image labelling to natural language modelling and text generation. In the majority of cases, they outperform classical machine learning methods in terms of performance. Long Short-Time Memory (LSTM) recurrent neural networks enable analysis of long sequences with complex temporal dependences. We used LSTM in this study, hypothesizing that any hidden elements of an activity depend on currently known elements, as well as on the temporal context.

Long Short Term Memory networks were first introduced by Hochreiter and Schmidhuber in 1997 [Hochreiter 1997] and have undergone multiple modifications ever since. A complete recurrent neural network (RNN), as shown in Figure 4.1, starts with an input layer of neurons that take input values, which is then connected to a series of recurrent layers. A simple network may contain only one recurrent layer. The last recurrent layer, in turn, is connected to a dense or otherwise called fully connected layer, which outputs the final values.

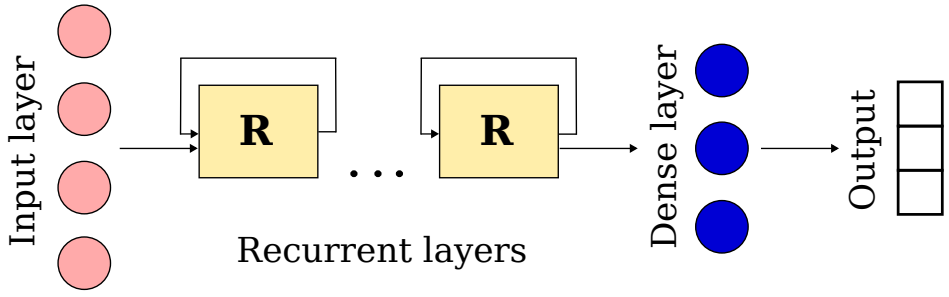

Figure 4.1: Recurrent neural network

Each recurrent layer has the form of a chain of repeating modules, one module per time step (i.e., item in a sequence). The unrolled recurrent layer is depicted in Figure 4.2. The RNN is actually called deep because of its depth in time.

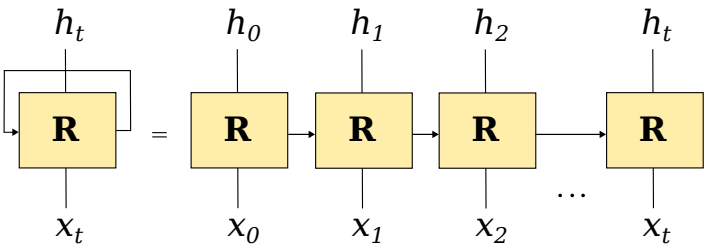

Figure 4.2: Unrolled recurrent layer. The letters *x* and *h* correspond to the input and output of a time step

The recurrent module of the LSTM network has the following structure and functioning (see Figure 4.3). A cell state C_t, represented by a horizontal line running through the top, is the core part of the module, and contains information about the sequence in form of a vector of real values. The module uses mechanisms called gates to control the content of the cell state and the output. They are composed of a sigmoid function and a point-wise multiplicator. The sigmoid function outputting numbers between zero and one defines the amount of information to pass through the gate (0 - none, 1 - all).

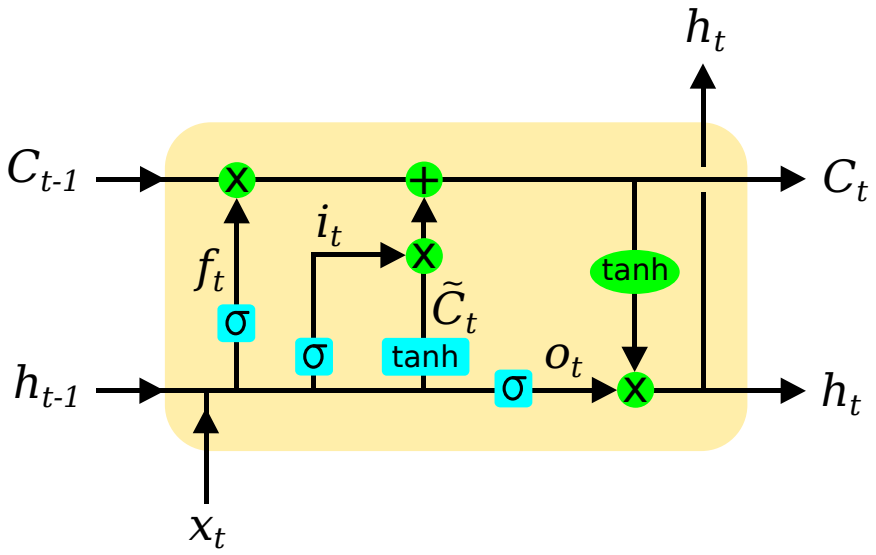

Figure 4.3: Recurrent module of an LSTM model

To process the data at the current time step t, its input x_t is first concatenated with the output of the $t-1$ module h_{t-1}. Then, the first *forget gate* decides what information has to be erased (i.e., forgotten) from the cell state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4.1}$$

The second sigmoid *input gate* defines the information to add and store in the cell state, and the tanh function prepares the actual input values by putting them in the range [-1;1].

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4.3}$$

A pointwise product of the outputs of these two functions is than added to the cell state.

$$C_t = f_t \circ \tilde{C}_{t-1} + i_t \circ \tilde{C}_t \tag{4.4}$$

The third *output gate* decides what part of the information contained in the cell state has to be outputted at this time step. As with the previous gate, the actual values from the cell state pass through the tanh function, and they are then multiplied by the output of the last sigmoid function.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.5}$$

$$h_t = o_t \circ \tanh(C_t) \tag{4.6}$$

The obtained output vector h_t is also passed to the module $t+1$ along with the modified cell state C_t. As in classical neural networks, W and b correspond to weight matrices and biases learnt by training which are called *internal learnable parameters*. Back propagation trough time is used for that [Werbos 1990].

The LSTM network has several important *hyperparameters* that have to be set: number of recurrent layers, number of hidden neurons, loss function, number of training epochs, batch size, learning rate, optimizer and dropout rate. The number of recurrent layers is usually selected according to the learning task and data representation. Every subsequent recurrent layer summarises shorter temporal dependences extracted by the previous layer to analyse longer dependences. For example, in our data, shorter dependences would be the connections between the elements within one activity, and longer the connections between the activities in the sequence. The number of hidden neurons represents the size of the LSTM module's cell and defines the quantity of stocked information. The loss function calculates the difference between the actual network output and its expected output. Its gradient is computed during back propagation to gradually correct the network's error. A training epoch is one forward and one backward pass through all training samples during which the internal parameters of the network are updated. The batch size defines the number of training samples that are going to be propagated through the network in one forward/backward pass. The learning rate defines the portion of the adjustment applied to the old neurons' parameters. The optimizer is the algorithm that defines how the internal parameters are updated (e.g., using first order or second order derivatives to minimize the loss function). The dropout is a regularization technique for reducing overfitting. Its rate determines the portion of neurons that are "disactivated" during back propagation.

While many LSTM models exist (e.g., GRU [Cho 2014]), all produce similar state-of-the-art results [Jozefowicz 2015, Greff 2017]. In this study, a variant of classic LSTM [Graves 2012] including three gates (i.e., input, forget, output), an output activation function, no peephole connections, dropouts and a full gradient training was used. We also tested different sets of aforementioned LSTM hyperparameters. All of the tested networks generated similar results with less than 5% difference. In order to recreate the same analysis conditions for all activity elements, the same LSTM model (that which provided the best results on preliminary tests) was used for all configurations and experiments, described as follows. The model had two stacked recurrent layers with dropouts of 0.2, each containing 256 hidden neurons. It was trained during 50 epochs with a learning rate of 0.001 by 128-size batches. Categorical cross entropy was used as the loss propagation function, together with Adam optimizer. The input data was transformed to a two dimensional matrix of size NxM, where N (see Section 4.2) subsequent activities within the temporal context corresponded to rows, and a one-hot vector of size M uniquely identifying each activity (i.e., an array where all values are set to "0" except one that set to "1", which defines the activity by its index in the vocabulary of all possible activities) to columns. The number M correponds to the number of unique activity tuples in a dataset. The neurons of the last dense layer output a probability distribution vector of size M.

4.4 Study design

In order to confirm our hypothesis and determine the essential sensors and signals that are necessary for activity recognition, we conducted a series of several experiments assessing the impact of each element and their combinations on recognition performance, as described below. The analysis was done on the seven datasets from Chapter 3 which were examined separately but using the same protocol. Three configuration types were considered: one-element configurations (V - only the verb information was available, I - only the instrument, and S - only the structure), two-element configurations (VI - both the verb and the instrument were known, VS - the verb and the structure, and IS - the instrument and the structure), and three-element configuration that played a role of a base line (VIS - all the elements were known).

All the experiments assumed the presence of element information used as input at each moment of surgery. This information was assumed to originate from underlying distinct processing and recognition algorithms, each taking care of its own element. The performances were compared based on an accuracy score. An activity was considered well recognized when all tuple items were correctly discovered. Given the relatively small amount of data we had available for deep learning, we performed a full-cross validation for each dataset in a leave-one-intervention-out manner. Moreover, since LSTM uses non-deterministic algorithms for training, we performed three runs for each fold, calculating an average recognition score for three models. The following values of N were tested in

each experiment: 5, 10, 20, 30, 40, 50, 75, 100. In several experiments, statistical tests were performed to estimate the significance of the results. All the tests were performed comparing average recognition accuracies of each intervention in a dataset provided by two different configurations. The order of interventions for both compared configurations was the same.

4.4.1 Experiment1: One-element configuration

The first experiment was designed to compare the activity recognition performances achieved with using each individual element as the only input. We also examined one-to-one relationships between the elements to assess how well one element can be recognized when another another is known. This experiment focused on the sequential aspect only, omitting timestamps and duration of activities. The model thus had only to predict the correct activity tuples in the correct order, without indicating moments of transition. Each recognized activity is supposed to start when all three elements appear in the operating scene (in theory, they have to be recognized at the same time), and end when they disappear, which then implicitly provides the activity duration.

4.4.2 Experiment 2: Two-element configuration

The second experiment compared combinations of elements, meaning that a pair of known elements was used to infer a complete activity. The same "no-time" condition was used, requiring a sequence of activities as output only.

4.4.3 Experiment 3: Left hand vs. right hand

As it was said earlier in Section 3.2.2, recognition of activities executed by both surgeon's hands is highly relevant to the situation understanding. In this study, any configuration was assumed to contain information about both hands at once. The present experiment, however, was designed to assess the individual contribution of each hand to the recognition performance. Thus, during the tests, the information of the left hand and the right hand was alternately masked for all three configuration types. A complete both-hand activity had to be discovered as output. Only the order of activities was considered.

4.4.4 Experiment 4: Activity duration

Contrary to three previous experiments where the workflow was considered as a sequence of activities only, in this experiment we added the duration of the activity (in seconds) at the end of its input tuple. We then analysed how knowledge of activity duration impacts inference process. However, the model was still required to predict a 6-tuple only, and no timing was taken into account when computing final accuracy. The constraint of

providing duration for input restricts the recognition process, as you have to wait for the on-going activity to finish. This could negatively reflect on on-line applications, yet it is still well adapted to cases where no immediate reaction is needed during the activity, or when only the order of activities is relevant.

4.4.5 Experiment 5: Noise in input data

The previous experiments were conducted with the assumption that all the input information was correct. In reality, raw signals coming from sensors may have a certain amount of noise, or some elements may be mislabelled by corresponding recognition algorithms. In this experiment, some element instances in activity tuples were randomly corrupted in order to simulate noise and create more realistic conditions for the analysis. For example, in one input tuple, the value of the right verb "cut" could be replaced by another existing verb "coagulate", the left instrument "needle-holders" by "classic forceps" or the right structure "ligament" by "fascia". In the simulation, the elements, as well as their left and right counterparts, were independently corrupted, meaning that noise occurred at different moments in time for all six items of the tuple. Four types noise were simulated:

— *Uniform distribution noise.*Using this kind of noise, a corrupted element instance is replaced by another instance of the same element group chosen randomly from a uniform distribution.

— *Frequency distribution noise.* Often, recognition algorithms tend to assign the labels of the most prevalent classes to incorrectly recognized samples. In our case, if an underlying recognition algorithm was trained over the entire procedure, the most commonly represented element instances (in terms of number of samples) would be those that appear in the operational scene for longer than the others. To simulate the behaviour of this type of noise, each instance has a chance to be randomly selected proportional to the frequency of its appearance in the dataset computed by duration.

— *Pairwise noise.* Another potential occurrence is when the samples of two major classes are mutually mislabelled (i.e., their labels are switched), this is known as pairwise noise.

— *No signal.* Sometimes recognition algorithms fail to identify a performed action or an object present in the scene, providing no label at all. In this experiment, a temporal absence of the sensor signal or the algorithm's disability to recognize an element is simulated by simply replacing the corrupted instance with the word "none".

Configurations of all three types were compared in this experiment. For the first two configurations (one and two known elements), the LSTM models from experiments #1 and #2 trained on noise-free data were tested on corrupted data. For the third configuration where all three elements were available with no need of an LSTM model, the activity was simply defined as their composition.

For all configurations, the noise was simulated at different rates: 5, 10, 15, 20, 25, 50 and 75% of the corrupted data in the procedure. For instance, for an algorithm recognizing instruments that provides a correct label 95% of time, 5% of all instrument instances in the intervention will have a wrong label. The same is applicable to other elements. Given that a correct activity is one where all tuple items are correctly recognized, with 5% noise for each element, a total amount of corrupted activity tuples may vary from 5 to 10% for one-element configurations, to 20% for two-element configurations and up to 30% for the base line. In the best case scenario, all items in all corrupted tuples have wrong labels, and in the worst case, no more than one item is wrongly labelled in each corrupted tuple. Giving this great variation, noise at each rate was simulated five times, and an average was calculated. No time aspect was involved in the analysis.

4.4.6 Experiment 6: Temporal delay

For the previous experiments, we mostly worked with only the sequencing aspect (with the exception of experiment #4). No time was taken into account when computing accuracy, which was strictly based on the order and correctness of activities, not on their durations. The elements were supposed to be recognized at the same time as they appeared in the scene. However, the underlying recognition algorithm may experience a certain delay before providing a label. In this experiment, we simulated such a temporal delay.

As in the fourth experiment, we first added the duration of each activity to its tuple, and then simulated delay for all available elements. The delay caused a change in activity duration, and, in some occasions, a shift in activities order by creating new tuples and deleting or altering existing ones. The last one changes the workflow of the intervention in terms of sequencing and number of activities. The goal of the LSTM model was to discover a sequence of activity tuples without giving their correct durations. However, they were accounted for when calculating final accuracy, which was computed as the sum of durations of all correctly discovered tuples divided by the total duration of all activities in the intervention. Delays of 1, 5, 10, 15, 20, 25 and 30 seconds were simulated. As in the previous experiment, we did not retrain the LSTM models on delayed data, and VIS base line was defined as combination of ground truth activity elements with no LSTM model.

4.4.7 Experiment 7: Phase recognition

Until now, the whole study was devoted to the recognition of low-level semantic activities only. One can also wonder how the choice of sensors may influence the phase recognition. In this last experiment, the surgical phase was recognized from the activity elements. All three configurations were tested as input. The same LSTM models as for experiments #1 and #2 were used, but the output labels contained the names of phases only. The accuracy was computed as the number of correctly recognized phase labels to the total number of labels.

4.5 Results

4.5.1 Experiment 1: One-element configuration

The accuracy of the activity recognition using each individual element is indicated in the upper part of Table 4.1. The experiment demonstrated that one element is not enough to confidently recognize activity. The instrument provided the best results for four out of seven datasets, yet no element is exclusively preferable for all procedures. The instrument and verb are tightly connected (Figure 4.4) and have a statistically significant (p-value $\leqslant$ 0.05) correlation, according to Spearman's Rho two-tailed test. While both elements provide a lot of information about each other, they contribute little regarding the structure, and vice versa. Cataract surgery, however, happens to be an exception, it is a short, highly standardized procedure with minimum of deviations and a small number of unique activities. Here, one element almost explicitly defined others, which explains exceptionally high scores.

During the experiment, different values of number N, defining the size of the temporal context, were tested. We observed that as the temporal context increased, the recognition scores tended to grow quickly until reaching a plateau. With a further augmentation of the number N, the recognition performances began to decrease. This behaviour is due to the working mechanism of LSTM. In order to obtain a clear picture of the relationship between the elements and activities, the network needs to consider the larger portion of the context. However, in order to clarify these connections, a larger set of training examples is necessary. Its size should increase in correlation with problem's complexity. Without sufficient amount of examples, the learning process becomes much less effective. That is why, calculating the optimal size of the temporal context depends on many aspects and differs for each presented dataset. The best results presented here correspond to $N = 50$ for ACDF procedures, $N = 20$ for LDH and PA, and $N = 5$ for CS.

Table 4.1: Average activity recognition accuracy (in %) achieved in experiments 1 and 2. Values in bold indicate the element(s) giving the best score for each dataset

	ACDF.L	ACDF.R	LDH.L	LDH.R	PA.L	PA.R	CS
V	49.72	66.29	52.64	62.32	48.63	68.23	**92.79**
I	**59.08**	**79.06**	59.69	**74.89**	58.17	**80.25**	90.40
S	54.50	63.27	**64.91**	68.29	**60.52**	60.08	90.18
VI	64.56	81.73	63.47	75.62	60.23	82.79	96.96
VS	84.99	83.33	91.12	90.11	85.16	87.73	97.06
IS	**94.18**	**96.54**	**97.30**	**97.40**	**97.10**	**96.81**	**99.82**

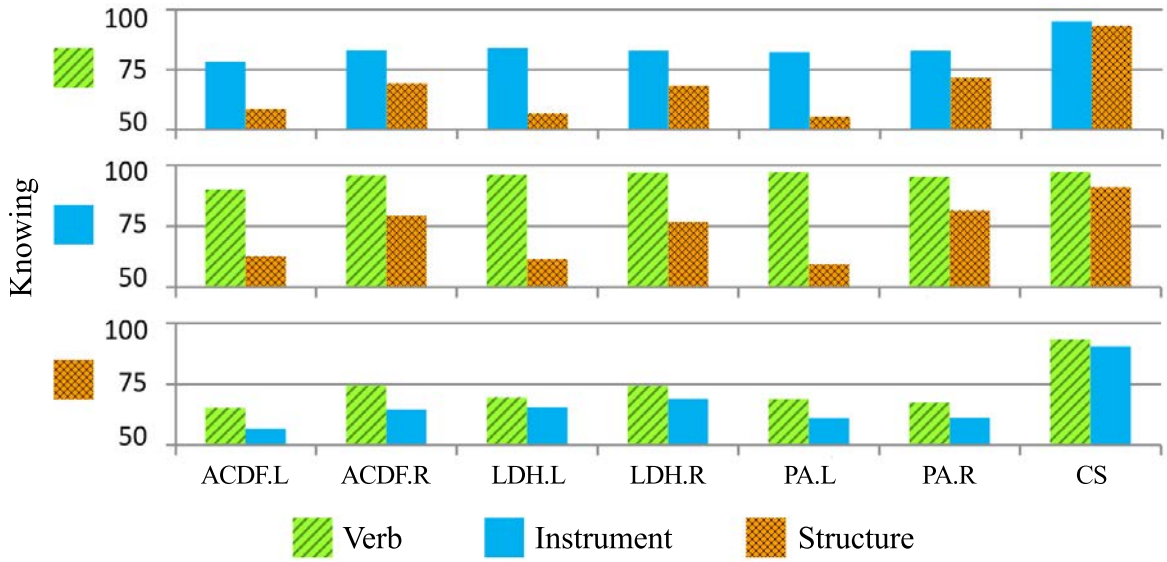

Figure 4.4: Average recognition accuracy (in %) for one element knowing another

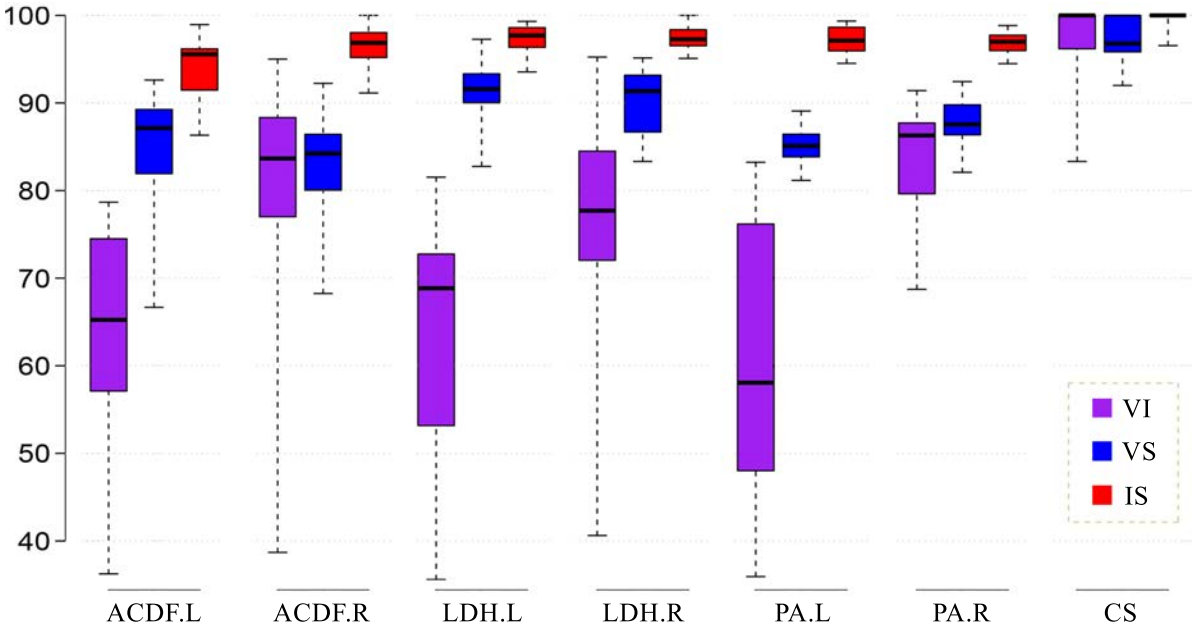

Figure 4.5: Activity recognition accuracy scores (in %) for element combinations. Center lines of the box plot show the medians, limits indicate the 25th and 75th percentiles, whiskers extend to minimum and maximum values. For each dataset, VI is on the left, VS in the middle, and IS on the right

4.5.2 Experiment 2: Two-element configuration

The recognition accuracy provided by combinations of activity elements can be found in the lower part of Table 4.1. As expected, the VI combination, providing redundant information with less clues about the structure, generated low performance results, proven insufficient for correct stable recognition. A VS combination produced relatively good results of approximately 85%, that would probably be acceptable for some purposes. For all procedures and sites, the IS combination was satisfactory to confidently recognize activities, producing a score of approximately 95% and higher. As in the previous experiment, these results were obtained with $N = 50$ for ACDF procedures, $N = 20$ for LDH and PA, and $N = 5$ for CS. We also performed a Wilcoxon signed-rank statistical test. For all datasets, the scores given by these three combinations significantly differed, with a medium to large effect size (p-value $\leqslant 0.01$ for a two-tailed test, except p-value $\leqslant 0.05$ for VI versus VS from PA.R and CS, and VI versus IS from CS; no significant difference between VI and VS from ACDF.R). A noticeable difference in scores can be also observed in Figure 4.5. This demonstrates that only two types of sensors are necessary for low-level activity recognition. In addition, the anatomical structure, present in two leading combinations, is an essential piece of information.

4.5.3 Experiment 3: Left hand vs. right hand

The results of this experiment can be found in Table 4.2. The R-VIS configuration, provided the highest accuracy of approximately 66%, was still insufficient for an accurate recognition. The R-IS configuration generated results comparable to R-VIS. The experiment revealed a 15-20% gap in performance between the left and right-hand elements. These results from Table 4.2 were generated using with $N = 20$ for ACDF and LDH procedures, $N = 10$ for PA, and $N = 5$ for CS. It was noticed as well that the L-VIS configuration helped to discover the right anatomical structure the best, and the R-VIS the left verb. The experiment showed that neither the left nor the right hand solely can accurately indicate the overall activity, even if all three elements of the same hand are known. This demonstrates the joint importance of hands in the recognition process, and suggests that the sensors have to capture information about both surgeon's hands.

Table 4.2: Activity recognition accuracy (in %) achieved by one-hand (left L and right R) configurations and averaged for all datasets

	V	I	S	VI	VS	IS	VIS
L	32.6±22.2	34.7±20.9	31.2±17.5	37.8±14.3	42.5±13.1	43.6±12.3	48.4±9.0
R	41.5±12.4	48.6±8.9	46.5±9.1	55.7±7.4	60.2±6.1	64.7±5.7	65.96±5.3

4.5.4 Experiment 4: Activity duration

The experiment examining the importance of activity duration showed that using it as additional input information only slightly improved results of activity recognition, having bigger effect on configurations with one known element than those with two. In average for all datasets, V configuration had a gain in accuracy of 3.7%, I - 4.3%, S - 4.1%, VI - 2.8%, VS - 1.3% and IS - 1.5%. Nevertheless, it allowed the IS combination to achieve a recognition accuracy close to 98-99%, which corresponds to our hypothesis. These results were obtained using the same N values as in the first and second experiments.

4.5.5 Experiment 5: Noise in input data

The results of the activity recognition from noisy data can be viewed in Figure 4.6 on the example of frequency distribution noise. We chose to present the results for the example of frequency distribution noise as this is the most common type of noise related to data recognition and classification. As expected, all the configurations had reduced ability to predict on-going activity when subjected to noise. Generally, those which previously providing higher recognition scores (i.e., having more useful information in them) were the most significantly affected (see Figure 4.7). While a ranking of one-element configurations was slightly altered for some datasets, the two-element combinations kept their order: IS was still the most informative combination, followed by VS then finally VI. The highest activity recognition accuracy for IS combination ranged from 79% to 84.4% with 5% noise, and decreased to an average of 6.6% with 75% noise. However, using an LSTM model encoding procedure history enables the effect of noise to be attenuated as well as "correcting" input tuples, especially for smaller amounts of noise.

It is interesting to see that the base line VIS combination always concedes to IS, and that it yields to all two-element combinations event with relatively small amounts of noise (starting from 10-15% noise). It quickly decreases reaching almost zero accuracy at 75% noise, and exhibits the most significant accuracy loss (see Figure 4.7), assuming that a perfect VIS combination would attain 100% recognition. VIS represents a naive approach of simply putting three elements together with no temporal model, and is thus unable to correct itself. Unlike other configurations, in the presence of noise it is automatically incorrectly recognized. The rapid drop in its performance quality can also be explained by the fact that an additional element in an activity tuple leads to a higher risk of its corruption, especially with greater noise. Thus, having less information is better than having lots with noise.

Continuing in our analysis of different types of noise, we found that at lower levels (up to 20%), the results for all noise types were similar with just a minor difference in accuracy (see Figure 4.8). This difference grows noticeable at higher noise rates, resulting in steeper or flatter curves. However, no statistically significant correlation between configurations and noise types suitable for all datasets was found at higher noise levels. The curve of the

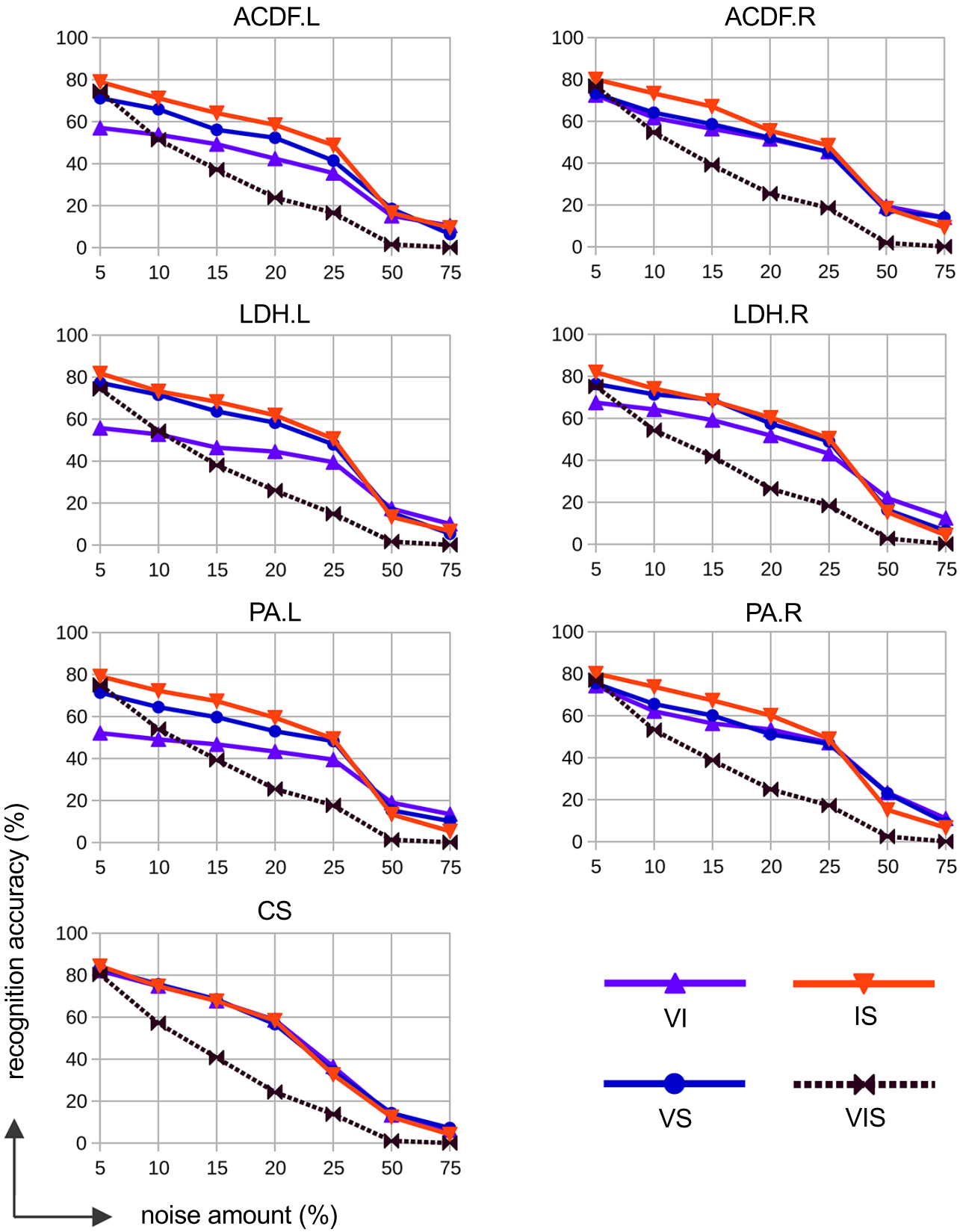

Figure 4.6: Impact of frequency distribution noise on activity recognition accuracy

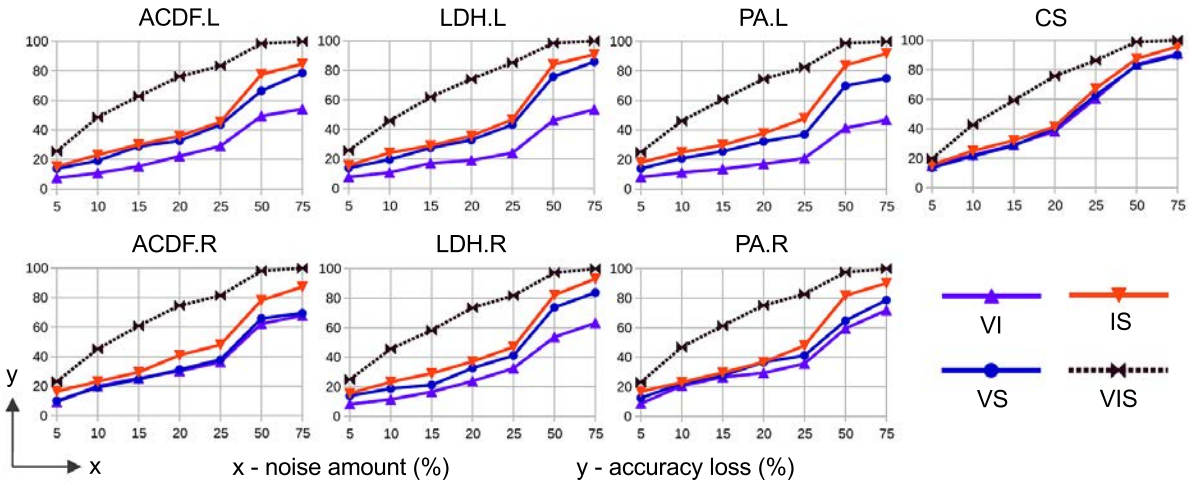

Figure 4.7: Relations between amount of frequency distribution noise and loss in accuracy

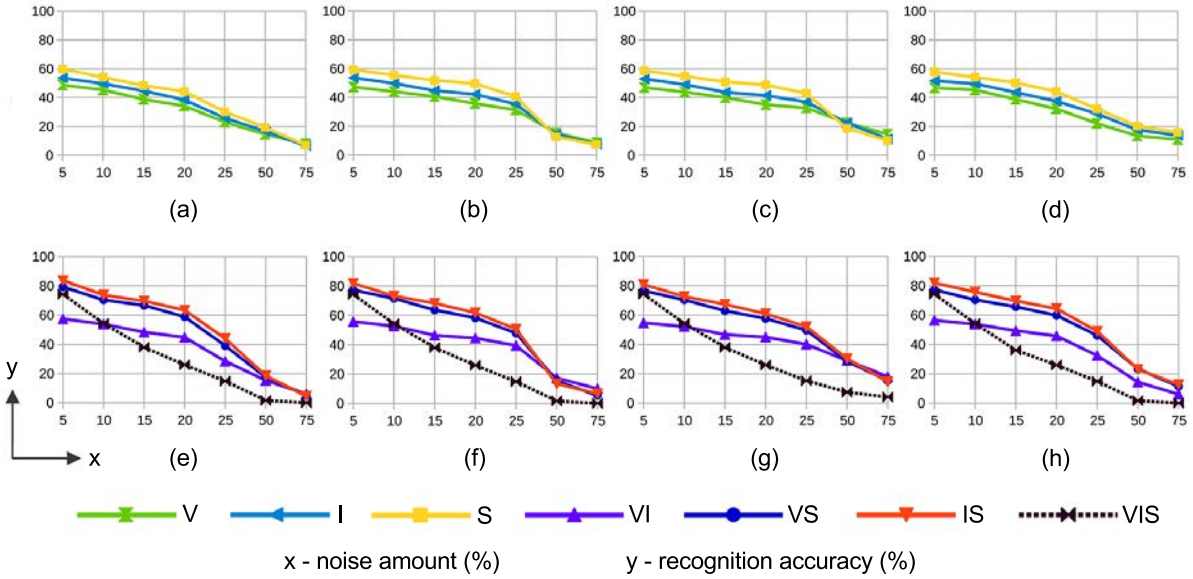

Figure 4.8: Influence of different types of noise on the activity recognition performance. Present diagrams correspond to LDH.L dataset. The uniform distribution noise is on the figures (a) and (e), frequency distribution on (b) and (f), pairwise noise on (c) and (g), no-signal noise on (d) and (h). The top shows the results for one-element configurations, while the bottom for two and three-element configurations

VIS base line is nearly the same for all noise types and datasets. In the case of VIS configuration, the quality of activity recognition on noisy data depends neither on the semantic content of the surgery nor on the type of noise, but rather only on the randomness of corruption. This is evident as an altered tuple item is wrong anyway, no matter its initial or received value.

Previous experiments have demonstrated that within certain limits, a wider temporal window is better for perfectly correct data. This experiment, however, showed that for noisy data, a small N value generates better results, as as chances to make a prediction based on false data are higher with a larger temporal context. In the exception of one-element configurations of ACDF.L and LDH.L datasets containing uniform noise, $N = 5$ was the best option for all other cases. Nevertheless, most of the time, the difference in accuracy scores given by N equaled 5, 10 and 20 was not statistically significant. The performance, however, significantly degraded starting from $N = 30$.

4.5.6 Experiment 6: Temporal delay

The results of the activity recognition from data with temporal delay can be viewed in Figure 4.9. As in the previous experiment, we found here that a delay caused all the configurations to progressively lose their recognition ability. Nevertheless, as before, the relationship between configurations remained the same with the IS combination achieving the best results. This combination keeps very high scores for a 1s delay, ranging from 91% to 97.9%, with an average of 94.1%. Even if some divergence is observable between IS and VIS curves for several datasets, as seen in Figure 4.9, the average IS and VIS curves are very similar with less than 1% difference at each delay point, with the exception of the 30s point where the IS configuration surpasses VIS by 1.8%.

As was the case with noise, the performance of VIS configuration is progressively impaired as the delay increases, as there is no temporal context of the procedure and no opportunity to correct tuple values. Two-element combinations, on the other hand, we found still able to discover on-going activity due to the history of the procedure represented by the LSTM. They nevertheless suffer from altered activity sequencing, making it difficult for LSTM to follow. We can observe that the biggest deficiencies for two-element combinations occurred in intervals from 1 to 10 seconds (a loss of approximately 15-20% each time). This can be explained by the fact that during these intervals, the most significant changes in workflows are made (i.e., creation and deletion of activities).

Similar to the experiment #5 with noise, smaller values of N were also preferable for recognition from data containing some temporal delay. The best performances exposed here were obtained with $N = 30$ for ACDF.L, $N = 20$ for ACDF.R and LDH procedures, $N = 10$ for PA and $N = 5$ for CS. Even so, their were mostly no statistically significant difference between N values from 5 to 30.

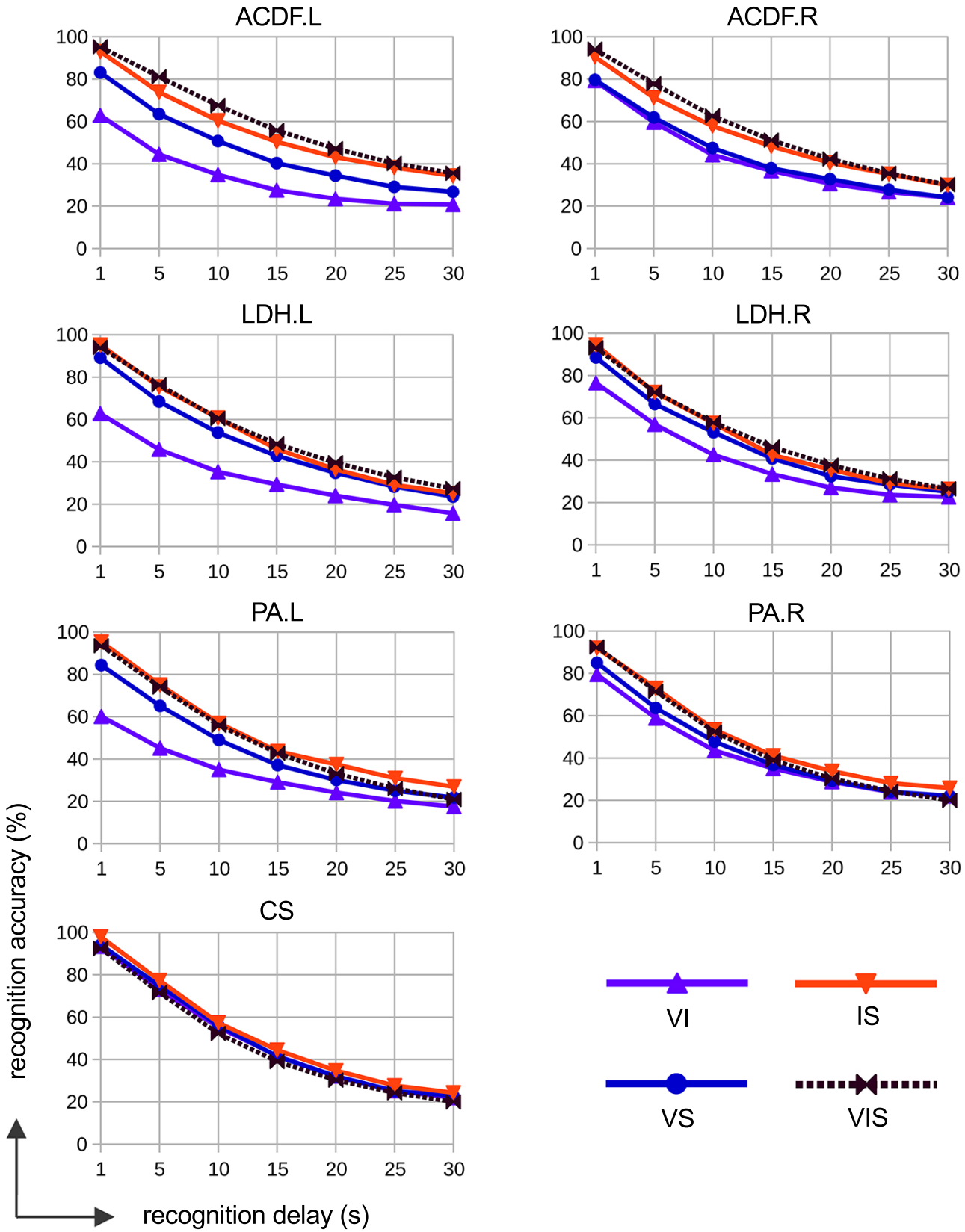

Figure 4.9: Impact of temporal delay on activity recognition accuracy

4.5.7 Experiment 7: Phase recognition

All the configurations provided similar results on average varying from 87.9% (the V configuration) to 92.3% (the VIS configuration) for all the datasets. According to the Wilcoxon signed-rank two-tailed test, there were no statistically important difference between the I, IS and VIS configurations at a p-value threshold of 0.01. This suggests that only one element (probably the instrument) should be used for phase recognition. Although, other signals (e.g., raw video) may be necessary to make a more accurate recognition.

In this experiment, a large temporal context was important for one-element configurations. Their best results were obtained using $N = 100$ for ACDF.L, $N = 70$ for ACDF.R, LDH.L and PA.L, $N = 50$ for LDH.R and PA.R, and $N = 10$ for CS. For neurosurgical datasets, there was no statistically significant difference between aforementioned values. They however provided significant improvement compared to values smaller than 50. For two-element configurations, the best results were obtained with $N = 50$ for ACDF, LDH and PA procedures, and $N = 5$ for CS.

4.6 Discussion

4.6.1 Experiments

This study proved our hypothesis that for accurate recognition of low-level surgical activity, not all of the activity elements need sensors to track. Though that sort of analysis should also be conducted for other surgical domains, the best choice for neurosurgery is the use of a combination of sensors recognizing the instrument and the anatomical structure. In the case of standardized simple procedures, such as cataract operations where one element is necessarily tightly bounded to two others, searching for the most informative elements is not worthwhile. However, two sensors are sufficient for activity recognition for these procedures as well. The experiments with noise and temporal delay also demonstrated the advantage of the instrument-structure combination over other configurations, including those uniting all three elements, suggesting that the VIS combination can be safely replaced by IS with no significant impairment.

In order to verify the aforementioned conclusions, further analysis in certain directions must still be undertaken. First of all, during our fifth experiment, for the sake of simplicity we assumed that all the elements had the same amount of noise in them, which is, of course, not necessarily the case in real-life procedures. The amount and type of noise in each element depends on the underlying algorithm for its recognition. The best way to get realistic estimations of scores is to use confusion matrices from these algorithms to simulate noise in data. We also proceed under the assumption that the perturbation in data was uniform in time, making each time-point equally available for corruption. This aspect should be explored more carefully, as it may not be valid in real surgeries. The same applies to the delay. It may also vary from one element to another in real-life situations,

as well as between different element instances. The combinations of different noises and delays must also be evaluated.

Secondly, in the last two experiments, the conditions under which the studied one-element and two-element configurations were compared to the VIS base line differed. An LSTM model was used for one and two-element configurations but not for VIS. Knowing that noise and delay were introduced in the test data only (i.e. the training data was perfectly correct), it would be impossible to use a deep neural network of this type for learning the mapping from correct complete activities to themselves in case of VIS configuration, because such learning would result in simple copying of the input activity without understanding temporal relationships between the elements. Suchwise, in these experiments, only the configurations with one and two known elements benefited from the temporal context of the procedure and correction of input tuples enabled by the use of an LSTM model.

4.6.2 Model

We demonstrated that in terms of recognition scores, the IS combination is capable of providing very accurate results. Nevertheless, a considerable drop in performance was observed in the presence of noise and delay (about 80% accuracy on average at 5% noise *vs.* 97% with no noise). This work sought to generate neither a high recognition performance nor a suggestion of an original efficient LSTM architecture. Nevertheless, in order to truly prove that two types of sensor are enough for surgical activity recognition, the overall performance should be enhanced. First, our main focus was on discovering relationships between activity elements using simple LTSM models. There are always subtle connections between the elements that influence the recognition process, however, regardless of which method is used. Thus, the conclusions drawn from the analysis would not considerably change using any other method or LSTM model. However, it should still be possible to find other more suitable deep models that could provide greater accuracy and maintain a strong performance even in the presence of noise. In our experiments, the chosen LSTM model was trained on no-noise data only. One can therefore imagine that retraining the network on simulated noisy data or using some preprocessing methods, as well as noise reduction techniques, could be beneficial. Secondly, the problem with delay can also be avoided. The procedural workflows used in our study were annotated manually in real time. Most of the very short activities are due to an annotator's late reaction or a surgeon's complex hand coordination. For most applications, such an extremely detailed annotation is unnecessary. Eliminating these very brief activities, causing a major change in activity sequencing in the experiment with delay, will make the recognition scores increase again. A larger amount of available data would also provide better results and make the network more robust. In addition, it should be noticed that not all clinical applications require absolute recognition accuracy. Certain errors or delay cause no harm and can be tolerated, consequently reducing the gap between developing activity recognition techniques and

their actual realization and use in operating theaters. Application-dependent metrics similar to [Dergachyova 2016] may be used to re-estimate this gap. Finally, one thing remains clear: placing more sensors in the operating theater is not a solution. The way forward is enhancing underlying algorithms recognizing verbs, instruments and anatomical structures.

4.6.3 Surgical practice

In addition to confirming our hypothesis about sensors, this study led to some interesting observations about surgical practices. During the experiments, we noticed that the elements of the right hand of the surgeon were obviously contributing more to the correct identification of the activity. However, despite the correlation between the surgeon's hand movements, neither the information about the right nor that of the left hand alone was enough to attain acceptable recognition results. This demonstrates how important both hands are in activity execution. The first two experiments also revealed a difference between practices in the Leipzig and Rennes hospitals. Using one-element configurations to discover the activity, the results for the procedures performed in Rennes were always significantly better than for those performed in Leipzig. At the same time, the instrument was clearly a better choice over other individual elements for Rennes, yet the same was not true for Leipzig. Moreover, in all of the procedures conducted in Rennes, we found there was a stronger bound between the instrument and structure, as well as between the verb and structure. Our resulting hypothesis is that, unlike in Rennes, the surgical instruments in Leipzig are more often used for new functions rather than their initially-intended application. This indicates that the procedures performed in Rennes are more standardized and have less variability in surgical workflow. Such observations are important for the analysis and understanding of surgical processes.

4.7 Conclusion

In this work we analysed the relationships between the essential elements of low-level surgical activity and their impact on recognition process. By performing a semantic analysis using deep learning, we demonstrated that two out of three elements are enough to confidently recognize an activity. The operated anatomical structure is a crucial element. The combined structure-instrument pair enables very confident activity recognition, followed by a structure-verb combination that provides slightly worse yet still acceptable results. This knowledge should facilitate the choice of right sensors to install in the operating room of the future for situation awareness. We also made some interesting observations about surgical practices that improve understanding of the surgical process.

Main findings

— Having sensors for two activity elements instead of three is enough

— A sensor for recognizing the anatomical structure is essential

— The instrument and the verb contain similar information
$\Rightarrow$ Only one needs to be recognized

— The combination of the structure and instrument is the most informative

— Recognizing the actions of both surgeon's hands is necessary

— For phase recognition, knowing one activity element is as good as knowing two or three $\Rightarrow$ Only one needs to be recognized (preferably the instrument)

— The anatomical structure relates more to the instrument and verb in Rennes than in Leipzig

— *Hypothesis 1*: The surgeries in Rennes are more standardized than in Leipzig

— *Hypothesis 2*: In Leipzig the instruments are more frequently used for new not initially-intended functions

CHAPTER

5 Knowledge transfer for prediction of surgical activities

Preamble

In this chapter, we describe our work on knowledge transfer that helps to improve performance of deep architectures and compensate for data deficit. We first explain the importance of knowledge transfer and related problems. Then, we introduce the first transfer method - word embedding, and next the second one - transfer learning. We also present advanced studies conducted to find what works the best for these two methods and to demonstrate their efficiency. Finally, we discuss both methods and conclude with interesting observations about the transfer and surgical practices in general.

Contents

5.1 Introduction

Comprehensive and correct training data always played an important role in machine learning. Now, in a new era of deep learning, the quantity of data becomes a major factor along with its quality. The sizes of massive datasets used as training basis for deep learning approaches, such as famous ImageNet, Microsoft COCO, recently released Google's Open Images and YouTube-8M, speak for themselves (Table 5.1). They contain millions of samples representing thousands of categories. A data-greedy tendency can also be observed thorough text-based corpora broadly used for natural language processing, such as Amazon reviews, Stanford's SNAP Social Circles: Twitter Database [Leskovec 2014] and DBpedia dataset [Auer 2007] raised from Wikipedia. They propose millions and billions of words to learn from. These datasets constantly continue to grow and new ones keep appearing regularly. The success of deep models is recognized to be a product of their ability to learn multi-level hierarchical representations, which simulates the processing pipeline of the human brain. But their real power comes from extensive amounts of data provided for training.

Table 5.1: Massive datasets for deep learning

	Dataset	Reference	Content
Images	ImageNet	[Deng 2009]	14M images, 22K classes
	Microsoft COCO	[Lin 2014]	300K images, 2M x 80 objects
	Google's Open Images	[Krasin 2017]	9M images, >6K classes
	YouTube-8M	[Abu-El-Haija 2016]	7M videos, 4.7K classes
Text	Amazon reviews	[McAuley 2013]	35M reviews, 3B words
	SNAP Social Circles	[Leskovec 2014]	>1.7M tweets
	DBpedia dataset	[Auer 2007]	3.4M concepts

Unfortunately, sometimes, it happens that a learning task has to be performed on a domain of interest represented by a small set of data. In these cases, *knowledge transfer* may come into play. In a large sense, knowledge transfer involves methods that use resources from other domains of interest, where the data may have different distribution and be in a different feature space, to improve learning of a targeted task. As Pan et Yang say in their survey [Pan 2010], "The study of transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions". With deep models, the learnt knowledge of one network can easily be transferred to another. Transfer learning, a technique of knowledge transfer, is now widely used together with Convolutional Neural Networks (CNN) for tasks related to visual content (e.g., image/video recognition, captioning, segmentation and detection) benefiting from freely available massive datasets [Oquab 2014, Karpathy 2014]. It is also broadly applied to tasks of sequence analysis as speech and language processing [Huang 2013], document classification [Dai 2007] and sentiment analysis [Glorot 2011].

Surgical domain also starts opening the doors to deep learning methods yet the amount of available data is still a hot discussion topic. Multiple constrains stand in the way of a proper data acquisition: ethical approvals, patient's and medical staff's consents, limited amount of cases, expensive installation of data acquisition equipment and time-consuming manual annotations requiring medical experience. As with other domains of interest, knowledge transfer is an option. For example, Shin et al. [Shin 2016] used transfer learning within CNN for computer-aided detection of thoraco-abdominal lymph nodes and classification of lung diseases. Twinanda et al. [Twinanda 2017] used transferring approach to classify images of laparoscopic surgery into surgical phases.

The problem of data deficiency for surgical activity recognition was emphasized in Chapter 2 by pointing out the small amount of interventions per study. We propose to use knowledge transfer to solve this problem. A transfer of surgical process knowledge, however, has not been done before. It has several difficulties compared to classic visual or text-based transfer. First of all, the transfer of knowledge from other non-related do-

mains is complicated by a particular data representation. Meanwhile, for tasks related to visual content, any image from another domain can easily be brought to a needed format (size, color channels, etc.) with very small information loss. The text for any natural language related task can be represented as a sequence of individual words. Secondly, a domain with a similar underlying logic has be found to make the transfer work. Despite the differences in contexts, scenes, objects and viewpoints, CNN capture information of any image in a hierarchical manner from basic visual features to shapes, objects, their collections and scenes. The hidden layers farther away from the output encode fundamental visual information and low-level characteristics shared with other visual domains [LeCun 1999, Yosinski 2014]. Text-related domains share semantic, syntactic and grammatical features. Surgical processes can also share some knowledge with other complex processes. Yet, the absence of comprehensive processual datasets limits transfer options.

In the previous chapter, we used deep learning to find relevant sensors facilitating surgical activity recognition. In this chapter, complementing the previous one, we use deep learning and knowledge transfer to bypass the lack of activity workflow data and improve the recognition performance. For this purpose, the task of next surgical activity prediction for neurosurgical procedures from Chapter 3 was chosen. This chapter introduces two methods of knowledge transfer. The first method of word embedding serves to extract semantic knowledge about surgical terms from medical texts. The second represents transfer learning that enables capturing important information about the surgical process and transferring it from one surgery to another. Thus, the presented work consists of two parts separately described in Sections 5.3 and 5.4. These two parts are, however, connected since the second method is based on the first one and integrates word embedding results. For both methods we conducted studies on different sources and types of transfer described in corresponding sections. The results of these studies also include interesting observations concerning surgical practices. Finally, we conclude the chapter with a discussion and findings that can also serve as a guidance for dataset enlargement for multiple recognition and analysis tasks involving surgical workflow.

5.2 Predicting next surgical activity

The learning task of predicting next surgical activity can be formalized as follows. Let $\mathbb{P} = \{\alpha, \mathbb{S}\}$ be the domain describing a surgical process consisting of two components. The first component is a set of m possible surgical activities $\alpha = \{A_1, A_2, ..., A_m\}$, where, according to the definition given in Section 3.2.2, each activity A_i is represented by a 6-tuple $< L(V, I, S), R(V, I, S) >$ containing a verb, instrument and structure for both left and right surgeon's hands. The second component of the surgical process is a set $\mathbb{S}$ containing recorded surgical interventions represented as an ordered sequence of activity tuples $Seq = (a_1, a_2, ..., a_l) \in \mathbb{S}$, where $a_i \in \alpha$, and the sequence length l is different for each intervention. Let $Seq_t^* = (a_{t-n+1}, a_{t-n}, ..., a_t)$ be a partial sequence defining the workflow of

an on-going intervention within a temporal context of size $n < l$ during the current activity $t < l$. The learning task $\tau = \{\alpha, f(\cdot)\}$ consists in learning from training data an objective predictive function $f(Seq_t^*) = P(a_{t+1} = A_j | Seq_t^*)$ which predicts the next surgical activity for a sequence of already known performed activities.

5.3 Word embedding

The goal of the first part of the work was to extract and encode the knowledge about the surgical domain and the meaning of surgical terms from medical literature, as well as to create a common basis for data representations enabling further transfer learning. According to our hypothesis, different surgeries share certain procedural knowledge that can be transferred from one to another in order to enlarge training database. Nevertheless, every surgery also expresses general surgical knowledge characterized by specific concepts and relations. The surgical process is described by a vocabulary. The words of this vocabulary serve to define activities but they in fact contain semantic knowledge about the surgical domain. In a classic deep model (i.e., method used in the last chapter), the activity elements get transformed into unique keys representing their position in the vocabulary. This enables unique identification of activities but makes elements lose their semantics. Moreover, every surgery has its own activity vocabulary with a particular size. The vocabulary size in turn determines the number and dimensions of the network's internal parameters. This size constraint hinders further transfer of knowledge from one surgery to another. In this chapter, we propose to treat each activity description as a sequence of words for which we find appropriate unified semantic representations. These representations are made using word embedding technique explained in this section. The word embeddings help to reintegrate the semantic knowledge about the surgical domain and to bring activity descriptions of any surgery to a common size.

5.3.1 Main concept

Word embeddings are a family of methods originating from natural language processing (NLP) domain that seek to map semantic meaning of words into a geometric space. This is done by associating a vector of real numbers to every word in the dictionary so that the distance between the vectors forming a so-called *embedding space* captures semantic relationships between the corresponding words. The vector values are found based on the words' co-occurrence information, meaning the frequency of their mutual appearance in a large text corpus. For instance, a famous example of *"king"* - *"man"* + *"woman"* = *"queen"* explains that simple arithmetic operations on the embedding vectors of "king", "man" and "woman" can (approximately) give a word embedding for "queen", if correctly defined. Word embeddings are also capable of incorporating syntactic information (e.g., "cat" to "cats" relates as "dog" to "dogs", or "clear" to "clearer" as "strong" to "stronger"). The

size of the vector is an important parameter that defines how the relationships between the words are presented and how much of information is stocked in the vector. Generally, the size of the embedding vector grows with the size of the text corpus and the number of distinct words.

The term word embedding was originally introduced by [Bengio 2003] as a part of an NLP model. The power of pre-trained word embeddings was then demonstrated by Collobert and Weston in [Collobert 2008]. In 2013 Mikolov et al. proposed *word2vec* toolkit for training word embeddings via neural networks [Mikolov 2013]. Their model learnt embedding vectors in order to improve the ability of predicting a target word from context words. They also proposed the way to reduce computational complexity for learning high dimensional vector representations on a large amount of data. A year later, a GloVe model was proposed by Pennington et al. [Pennington 2014], which learnt embeddings by performing dimensionality reduction on the words co-occurrence matrix. These two models, described in Sections 5.3.3.1 and 5.3.3.2, replaced classical Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) methods, and are now widely used for many NLP and other problems. Not requiring any annotation effort, they are also tightly connected to unsupervised learning.

5.3.2 Word corpora

Word corpus is a collection of texts on a subject brought to the form of a plain sequence of words separated by single spaces. It is widely used in NLP for production of an embedding space encoding semantic relationships between the words describing the subject. Various large corpora (e.g., Amazon reviews, SNAP, DBpedia, etc.) are available on the internet, as well as their pre-trained embeddings. Despite the abundance of corpora on general subjects, in the current moment there are very few corpora from medical or surgical fields. Besides using already available corpora for training, a new corpus on a specific subject can be created from scratch using two following methods. The first way is to automatically download web pages on the subject using crawler tools. The second method consists in obtaining materials from special search engines having access to large databases trough dedicated interfaces. The second one is often used in biomedicine-related research [Huang 2016, Muneeb 2015]. In this work, we collected three corpora, jointly containing 708K unique words and 175M in total, using some available datasets and by requesting the data trough scientific search engines. The process of their collection and transformation is presented below.

5.3.2.1 Medical Transcriptions

Our first corpus contained Medical Transcriptions (MT) - voice-recorded reports dictated by healthcare professionals converted into text format. The web site www.medicaltranscriptionsamples.com proposes samples of transcribed medical reports

from many specialities that can be used by learners or working medical transcriptionists for reference purposes. The site proposes 103 samples from neurosurgery that represent post-operative reports recorded by the surgeon to describe the performed procedure and the patient's state. The neurosurgical samples include all three procedures from the clinical data used in this study (i.e., ACDF, LDH and PA). We downloaded these samples from the iDASH repository [1] which offers the same transcriptions in a form of a collection of text files. First, we concatenated the neurosurgical samples in a single file. Then, we formatted the file to eliminate all digits, punctuation marks, special symbols, and carriage returns. The obtained **MT corpus** consisted of 58975 words in total and contained 4469 unique words. Nevertheless, the common vocabulary of all clinical procedures studied in this work contains 79 unique words and only 66 of them can be found in the MT corpus.

5.3.2.2 PubMed abstracts

We composed the second corpus by collecting all abstracts referenced by PubMed relevant to the query "neurosurgery OR anterior cervical discectomy OR lumbar disk herniation OR pituitary adenoma" (searched in all fields). PubMed is a free search engine that provides references and abstracts of publications in life science and biomedical field kept in MEDLINE database. American National Library of Medicine maintains the database as part of the Entrez Global Query Cross-Database Search System for information retrieval purposes. The Entrez system can be accessed trough different interfaces available in many programming languages. We downloaded the abstracts in bulk through Entrez using Biopython library. The 62489 downloaded abstracts formed the **PubMed corpus** contained >57M words in total, 188K of which are unique. It also contains all 79 words from the clinical data vocabulary.

5.3.2.3 PubMed Central full-text articles

We created the last corpus by collecting full-text articles from PubMed Central (PMC) corresponding to the query "neurosurgery OR anterior cervical discectomy OR lumbar disk herniation OR pituitary adenoma AND free+full+text[filter]". PMC is a free digital repository developed by the National Center for Biotechnology Information (NCBI) that contains publicly accessible full-text scientific articles published within biomedical and life sciences journals. This time, we also used Entrez to retrieve the data from the repository. PMC referenced 97611 articles corresponding to the query. However, only 32271 of them were actually present in the PMC repository itself and directly available for downloading in a form of xml files. The **PMC corpus** created from these articles contained 638K unique words, including 118M words in total and all the words from the clinical data vocabulary.

1. iDASH repository https://idash-data.ucsd.edu) is supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460.

5.3.3 Embedding methods

In this chapter, we propose to use two methods to create word embeddings: word2vec and GloVe that are described below. We also integrate word embeddings into a deep model to enable on-the-fly transformation of activity descriptions into meaningful representations. Their integration is explained in Section 5.3.4.

5.3.3.1 Word2vec

Word2vec is an unsupervised embedding method published by Google's researchers in 2013 [Mikolov 2013]. The method used neural networks with three layers (an input layer, a single linear projection layer and an output layer) to learn vector representations of words in a corpus. Two architectures with their efficient implementations were proposed: **continuous bag-of-words (CBOW)** learning a word from a context, and **skip-gram** learning a context from a word (Figure 5.1). For example, for the phrase "the method proposes efficient architectures for learning vector representations", the CBOW model outputs the word "architectures" for the set of words "the", "method", "proposes", "efficient", "for", "learning", "vector" and "representations" provided as input. The skip-gram does the opposite - outputs a set of surrounding words from one current word. Before training the model, a vocabulary containing all unique words from the corpus has to be created. Let V be the size of this vocabulary. Each word in the corpus is then represented as a V-dimensional one-hot vector - an array where all values are set to zero except one cell, which defines the index of the word in the vocabulary, set to 1.

In the *continuous bag-of-words model* model, the input layer is formed by the C context words, and the output layer by a V-dimensional probability vector. The model has two weight matrices to learn that are shared for all words: W_1 between the input and projection layers, and W_2 between the projection and output layers. The matrices are of size VxS, where S is the size of embedding vectors. The activation function of the projection layer is linear: the value passed from the projection to output layer is a sum of the W_1 rows, corresponding to the input words, divided by C. The second weight matrix W_2 serves to compute a probability score for every word in the vocabulary. In output layer, a softmax is used to obtain a posterior distribution of words. The training objective of the model is to maximize the conditional probability of observing the actual output word given the input context words.

For the *skip-gram model*, learning to "predict words in a certain rank before and after current word" [Mikolov 2013], the current word makes a single input vector, and the output layer is now formed of C V-dimensional distributions instead of one. It has also two VxS weight matrices. However, the projected vector is now simply the W_1 row corresponding to the input word. The training objective is to minimize the summed prediction error across all context words in the output layer.

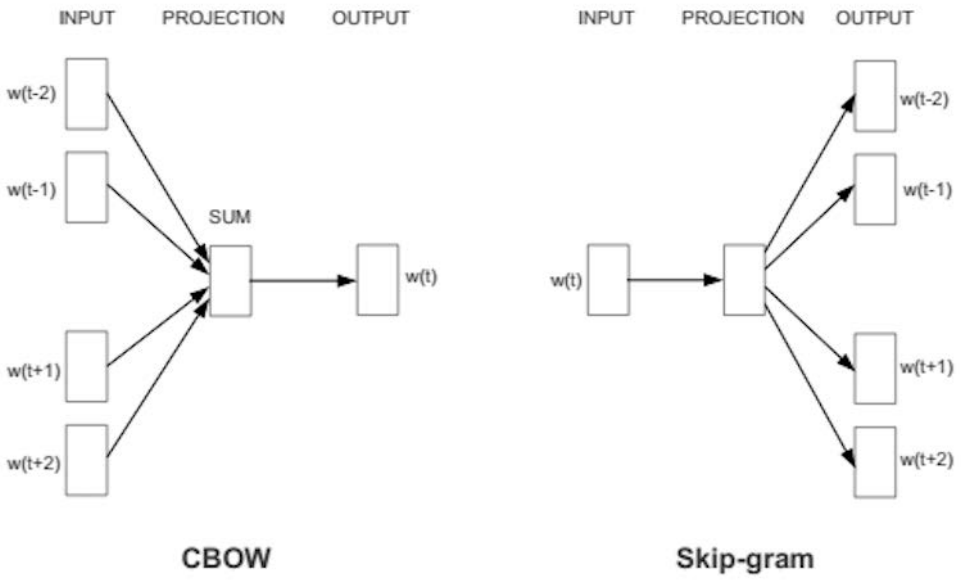

Figure 5.1: Continuous bag-of-words and skip-gram architectures of word2vec embedding method. *Source: [Mikolov 2013]*

In both models, all the words are projected and outputted into the same position meaning that their order in the phrase does not influence the projection. The training thus consists in giving pairs of words to the model that learns the statistics from the number of times each pair appears in training samples. Two words in the corpus laying "nearby" may form a pair. The notion of closeness is actually defined by the size of a context window, usually 5 neighbouring words at both sides. The intuition behind the method is that if two words are situated nearby, they are semantically or syntactically related. The network will learn similar embeddings for the words appearing nearby, and dissimilar embedding vectors for the words appearing far-away from each other in the text.

The word2vec models have been substantially inspired from [Bengio 2003], but their main advantage comes from the tweaks proposed by the authors to optimize the training process. A successful learning requires a large corpus which leads to a great number of words in the vocabulary. As a result, a vanilla model has a very high number of internal parameters to train $O(V \text{x} S)$. Running gradient descent on such a structure is very slow. The authors proposed two efficient optimizations. First is to subsample frequent words to decrease the amount of meaningless training samples. For each word the probability of its keeping in the vocabulary is computed based on the frequency of its appearance in the corpus:

$$P(w_i) = (\sqrt{\frac{z(w_i)}{s}} + 1) \cdot \frac{s}{z(w_i)}), \tag{5.1}$$

where $z(w_i)$ is the fraction of the word w_i in the corpus, and s is a subsampling rate usually set to 0.001. In this way, too frequent words (e.g., "the", "a", etc.) that are mostly irrelevant for capturing semantic information are eliminated from the vocabulary. The second optimization is so-called negative sampling consisting in using each training sample to update only a small percentage of weights, which is also shown to improve the final representations. Normally, during training, given a pair of words, ("blue", "ink") for example, the weights has to be adjusted so that the output neuron corresponding to "ink" outputs 1, and all others output 0. In negative sampling, instead of adjusting the weights of all neurons only a small number of randomly selected neurons is updated. According to the authors, selecting 5-20 words works well for smaller corpora and 2-5 for large ones. In that way, only a small part of all internal parameters has to be computed at one pass which greatly reduces computational complexity.

5.3.3.2 GloVe

GloVe is an another unsupervised learning algorithm for obtaining vector representations of words proposed by Stanford's researchers [Pennington 2014]. The authors argued that the scanning approach used in word2vec was suboptimal since it did not fully exploit statistical information about the word co-occurrences. The GloVe model is based on the statistics extracted from the word co-occurrence matrix. "The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning". Glove is trained to learn embeddings so that the dot product of word vectors equals the logarithm of word's probability of co-occurrence. As a logarithm of a ration equals a difference of logarithms, the ratio of probabilities is associated to the difference between two vectors.

The GloVe method consists in the following. A large co-occurrence matrix of VxV size, where V is the number of words in the vocabulary, is constructed from the corpus. The matrix is filled within a single pass trough the corpus by counting for each word how many times it has been seen in the context of other words. For a large corpus, this may be computationally expensive, but it is a one-time cost. Then, the matrix is recursively factorized to a lower-dimensional one until it attains the desired dimensions of VxS, where S is the size of embedding vectors. The goal is to find a lower-dimensional representation conserving most of the variance of the high-dimensional data. The factorization is done by minimizing reconstruction loss. These training iterations are done much faster because the number of non-zero matrix entries is typically much smaller than the number of words in the corpus.

5.3.4 Integration into LSTM model

In this study, we performed the prediction of next surgical activity using LSTM. A recurrent model with the same set of hyperparameters as in the last chapter (except some changes explained in the following) was used. The changes include a different data representation and one additional embedding layer placed before the recurrent layers (Figure 5.2). In the initial model, the input data was transformed to a two dimensional matrix of size NxM, where the N subsequent activities within the temporal context corresponded to rows, and the one-hot vectors of size M uniquely identifying activities were columns. This time, each input activity tuple was first decomposed into a sequence of separate words (each instance of one element was broken down into composing it tokens). The sequence was then normalized (padded or truncated if needed) to have a common size of 15 words. Finally, the network input consisted of a concatenation of the word sequences of all the activities within the temporal context. The added embedding layer serves to transform the sequence of words into a matrix 15NxS, where S is the desired size of embedding vectors. The transformation is encoded in the layer's weight matrix, and can be obtained in three ways: 1) *train* - the weight matrix is obtained by training the layer from scratch along with all other layers, 2) *set* - the weights are set from pre-trained word2vec or GloVe embeddings, and the layer is frozen during training, and 3) *set & train* - the weights are initialized with pre-trained embeddings but still updated in the training process.

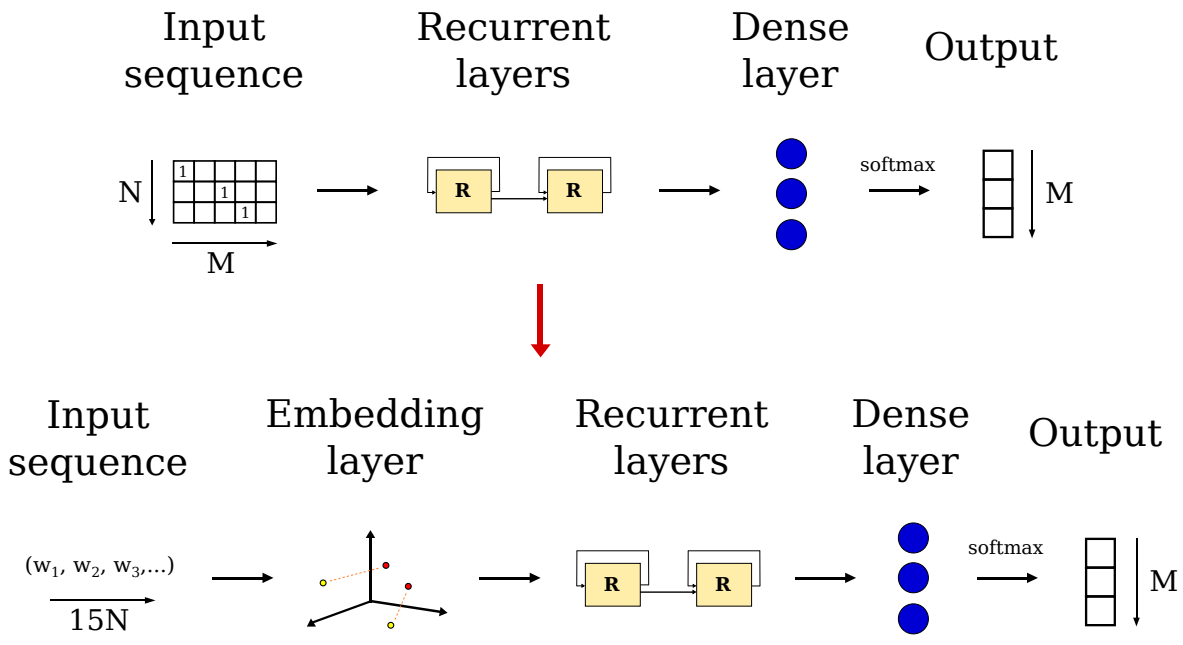

Figure 5.2: Initial LSTM model and the changes integrating word embeddings

5.3.5 Study design

We used the neurosurgical procedures from Chapter 3 to demonstrate that the semantic meaning encoded in word embeddings can improve prediction results. The main factor observed trough the entire study (both in the word embedding and transfer learning parts) was Δ - the amount of prediction improvement between different configurations (i.e. training without and with embeddings) that is the difference in prediction accuracy. Prediction accuracy score was computed as the total amount of correctly predicted activities to the number of test samples. An activity was assumed to be correctly predicted if all its elements of both hands were correctly predicted. A 10-fold cross validation within every dataset was performed each time, averaging the results of three training runs for one fold.

First of all, a base line was defined as the prediction accuracy of the initial model with one-hot vector like representations and no embedding layer. The only varying parameter was the size of the temporal context N indicating how many of previously performed activities were observed in order to predict the next one. The values of N equal to 5, 25, 50, 75 and 100 were tested.

Secondly, the model that trains word embeddings from scratch was assessed. The embedding layer of the model was trained at the same time as the others on the clinical datasets only, involving no collected corpus. That is why, in this case, the obtained embeddings represented a modified format of data yet did not contain any particular semantic relationships between the words. The model had two varying parameters: the size of the temporal context N and the size of embedding vectors S. The same values of N were tested as for the base line. The values of S equal to 25, 50, 100 and 200 were tested.

Finally, the impact of the pre-trained embeddings containing semantic information was estimated. Series of experiments were conducted varying the following parameters of the model: configuration (set or set + train), embedding method (word2vec CBOW or GloVe), word corpus (MT, PubMed, PMC or all combined), and embedding vector size (S equalled 100, 300 or 500). The N values provided the best results within the previous model were used. The weight matrix of the embedding layer was populated with the embeddings of the words found in the clinical datasets only. The embeddings of the words from the datasets not existing in the corpus (e.g., MT) were set by zero vectors. The results were compared to the base line and their statistical significance was measured using two-tailed Wilcoxon rank-sum test.

5.3.6 Results

The best results achieved by the initial model are given in Table 5.3. These scores established the base line to which the other configurations were compared. The average for all datasets accuracy reached 67.4% with average standard deviation of 12.5%. These results were obtained by setting N to 75 for ACDF.L and LDH.L, and to 50 for all other datasets.

The best accuracy scores reached by the model training the word embeddings from scratch are presented in Table 5.3. The change of representation and addition of an embedding layer provided only a minor not significant increase in accuracy of 0.8%. The new average accuracy attained 68.2% with 12.2% of standard deviation. The presented results were obtained with $N = 50$ for ACDF.L, ACDF.R and LDH.L, and $N = 25$ for others, as well as $S = 25$ (giving results significantly different from other S values). A modest embedding space seems to suit more small vocabularies.

Table 5.2: Prediction activity scores (in %) for the initial model and the model training embeddings from scratch. The values in bold indicate the datasets that had minimum and maximum scores. The difference in accuracy Δ between these two configurations is in red

	Configuration	ACDF.L	ACDF.R	LDH.L	LDH.R	PA.L	PA.R	Avg	Δ
Avg	No embedding	66.28	67.71	**62.83**	64.18	70.34	**73.15**	67.42	-
	Embedding (train)	66.39	68.21	**64.15**	65.07	71.08	**74.23**	68.19	0.77
Std	No embedding	12.54	12.15	14.53	13.70	11.24	11.02	12.53	-
	Embedding (train)	12.55	11.91	13.22	13.52	10.93	10.87	12.17	-0.36

Table 5.3: Prediction accuracy scores (in %) for different sets of parameters. The values in bold indicate maximum accuracies obtained by "set" and "set + train" configurations. The stars indicate statistical significance as regards to the base line: * p<0.05, ** p<0.01

		Word2Vec			GloVe		
Corpus	S	set	set + train	max Δ	set	set + train	max Δ
MT	100	65.8 ± 13.1	68.2 ± 11.8	0.8	63.4 ± 14.4	67.1 ± 12.4	-0.3
MT	300	63.5 ± 14.5	67.6 ± 11.3	0.2	62.1 ± 14.9	65.2 ± 13.3	-2.2
MT	500	61.0 ± 15.8	67.4 ± 11.3	-0.1	62.7 ± 14.7	65.2 ± 13.5	-2.3
PubMed	100	69.6 ± 10.5	$74.4 \pm 9.4^{*}$	7.0	68.0 ± 11.2	72.1 ± 10.2	4.6
PubMed	300	70.1 ± 10.5	$75.3 \pm 9.3^{*}$	7.9	69.5 ± 11.3	74.5 ± 10.0	7.1
PubMed	500	70.9 ± 10.7	$75.0 \pm 8.5^{*}$	7.6	70.7 ± 10.1	$74.7 \pm 9.5^{*}$	7.3
PMC	100	72.4 ± 10.2	$77.0 \pm 7.2^{*}$	9.6	72.2 ± 11.3	$77.5 \pm 8.9^{*}$	10.1
PMC	300	73.7 ± 9.8	$77.4 \pm 8.2^{*}$	10.0	73.8 ± 10.1	$77.9 \pm 8.5^{*}$	10.6
PMC	500	73.8 ± 9.5	$77.9 \pm 7.5^{*}$	10.5	$74.7 \pm 9.6^{*}$	$78.5 \pm 7.8^{**}$	11.1
ALL	100	73.5 ± 9.7	$77.3 \pm 8.3^{*}$	9.9	$74.8 \pm 9.3^{*}$	$78.3 \pm 7.9^{**}$	10.9
ALL	300	$75.1 \pm 8.6^{*}$	$78.8 \pm 7.5^{**}$	11.4	$75.2 \pm 8.7^{*}$	$78.9 \pm 7.4^{**}$	11.5
ALL	500	$\mathbf{75.7} \pm 9.1^{*}$	$78.8 \pm 7.4^{**}$	11.4	$75.4 \pm 9.1^{*}$	$\mathbf{79.1} \pm 7.5^{**}$	11.7

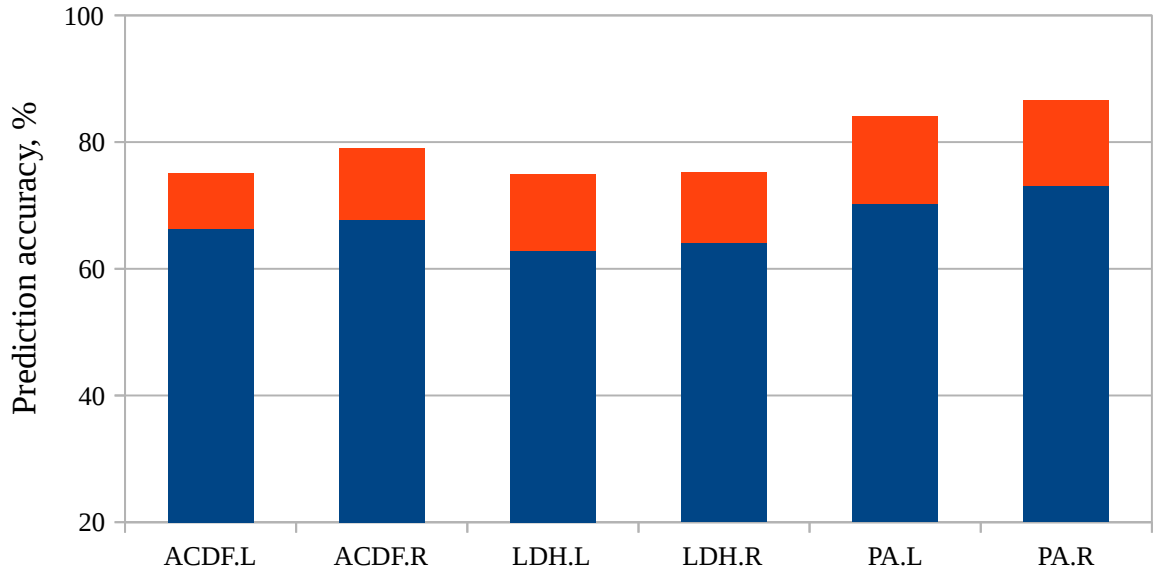

Figure 5.3: Prediction accuracy scores (in %) for different datasets with the best embedding configuration. The blue color indicates the scores of the base line, and the red color indicates the improvement

The Table 5.3 displays the average for all datasets prediction accuracy obtained with pre-trained embeddings and different sets of parameters, as well as the Δ (maximum between "set" and "set + train" configurations) comparing each set to the base line. The highest results (average accuracy = 79.1%, average standard deviation = 7.5% and average delta = 11.7%) were achieved using GloVe method on a combined corpus (ALL) with 500-dimensional embedding vector additionally trained for the learning objective. For this set of parameters, the ACDF.L dataset had the smallest Δ of 8.7% and PA.L the highest of 13.8% (Figure 5.3). It was observed that the accuracy went up with the growth of the corpus size. The same happened with the increase of the embedding vector size S, except for the MT corpus (it seems to be too small for higher values of S). The word2vec method performed better than GloVe with smaller corpora, but both provided similar results for larger corpora. The additional training of the embedding layer fairly improved the results. The pre-trained embeddings were made for a learning objective quite different form activity prediction. The additional training for the current objective adjusted the embeddings to reflect more the semantics of the surgical process. By these experiments, word embedding was shown to be a great tool for unification of data representations and introduction of higher semantic meaning that improves prediction results.

5.4 Transfer learning

5.4.1 Why and How

This whole section describes the second part of the work on knowledge transfer between surgeries. The work was motivated by the following hypothesis. Sequences of activities representing a surgical process encode some form of abstract knowledge about a given procedure, surgical practice and the process in general. This knowledge can be extracted and exploited to improve all sorts of operations on surgical process data, including analysis, recognition and prediction. It was particularly assumed that the knowledge obtained from one procedure might improve prediction of surgical activities for another procedures. The knowledge in question may include dependences between activities in a sequence, relationships between the elements inside an activity, and connections between individual elements of different activities spaced in time. In view of the implicit nature of the knowledge that can hardly be formalized, deep recurrent neural networks were chosen as a method able to extract and transfer it.

The deep neural networks have an interesting property that enables the network to store extracted information in a distributed hierarchical way. It means that more fundamental information common for many domains is stored separately from the features describing particularities of a specific domain of interest. It implies that this information can be shared with another learning objectives meaning an another training task or domain. The information encoding in recurrent neural networks is less studied compared to convolutional networks. However, the layers closer to the input are considered to learn shorter dependences between the items situated nearby in the sequence, while the layers closer to the output tend to learn long-term dependences between the items (or groups of items) situated far away from each other. In deep models, the knowledge learned from the data is encoded in the weight matrices representing the layers' internal parameters. In order to establish the values for the internal parameters, the model is first trained on a domain containing a large amount of training samples. Then, depending on the amount and quality of the data in the actually targeted domain, three transfer options exist. First, the same trained model can be directly used on new data if it is close enough to the one used for training, and if the task has not changed. The second option is to use the weights (all or only a part) from the trained model as initialization for the new model. This is suitable for cases where a reasonable amount of new data is available for training. The third option is called fine-tuning and often used when the new domain contains only a small number of samples. It consists in importing the trained weight matrices to the new model, but "freezing" some layers during training, usually those that contain more fundamental features. Setting the weights pre-trained on another data is often more optimized that a random initialization. The network can benefit from the already learned knowledge so that it concentrates its "attention" on specific features of the new data. The next section will describe how the transfer of surgical process knowledge was done in this work.

5.4.2 LSTM models

In this work, the transfer learning was based on the word embeddings presented earlier. They served not only to encode semantics of surgical terms but also to unify the data representations which enables learning from another procedures. Thus, the LSTM models used for transfer learning were the modified versions of the model from Section 5.3.4 that integrated word embeddings. Let us introduce some notions and designations helping to explain the structure and the functioning of the new LSTM models. We call the dataset containing the knowledge extracted for transfer *source dataset* $\mathbb{D}_{\mathtt{src}}$, and the model extracting the knowledge by training of the network on source dataset *source model* $\mathbb{M}_{\mathtt{src}}$. Analogically, the dataset that benefits from the transferred knowledge is called *destination dataset* $\mathbb{D}_{\mathtt{dist}}$, and the model that gets initialized with the weights of the source model is called *destination model* $\mathbb{M}_{\mathtt{dist}}$. Let M be the number of unique activities in a dataset, and V the number of unique words in the dataset vocabulary. Let subscript $\mathtt{src} \cup \mathtt{dist}$ define the activities and words belonging to both source and destination datasets. Let N be the size of the temporal context, and S the size of the embedding vector.

In this study, we used two LSTM models for transfer learning (Figure 5.4 and Figure 5.5). The *first model* $\mathbb{M}^1$ has the following structure. It takes a sequence of words of length 15N as input (as described in Section 5.3.4) that defines the content of the temporal context. The input layer is then connected to an embedding layer with a weight matrix of size $V_{\mathtt{src}\cup\mathtt{dist}}$xS. The embedding layer is followed by two stocked recurrent LSTM layers, where the last one passes a single vector to the final dense layer. The neurons of the last dense layer output a probability distribution vector of size M of the targeted dataset, it means that the number of neurons in the dense layer may be different for the source and destination models. The next predicted activity is the activity corresponding to the cell of the output vector with the highest probability value. The transfer between the source and destination models of this type is made as follows. For the source model $\mathbb{M}_{\mathtt{src}}$, the embedding layer is set from 500-dimensional embeddings pre-trained on ALL corpus using GloVe algorithm (the configuration that showed the best results in the experiments in Section 5.3).

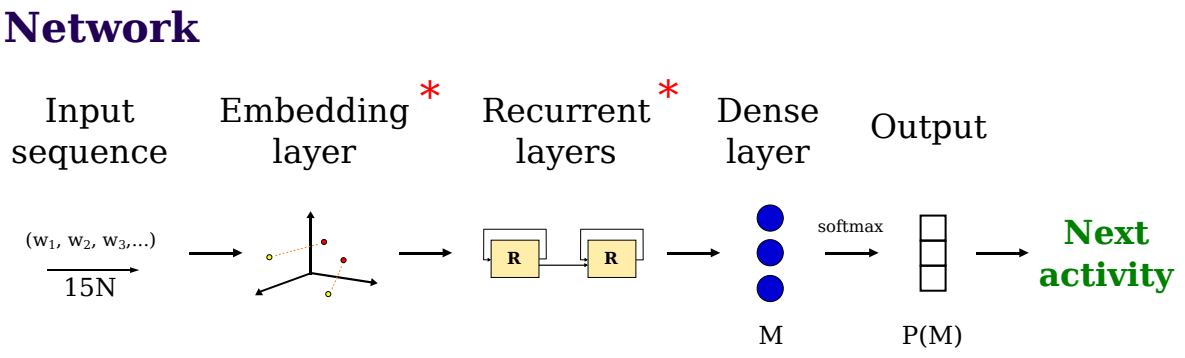

Figure 5.4: First type of model for transfer. The weights of the layers marked with a star can be transferred.

Only the $V_{src \cup dist}$ words are imported. The recurrent and dense layers are initialized randomly. Then, the model is trained on the source dataset $\mathbb{D}_{src}$. The weight matrices from the embedding and both recurrent layers are exported and saved for further transfer. For the destination model $\mathbb{M}_{dist}$, the embedding and both recurrent layers are first set with the weights from corresponding source layers. The dense layer is randomly initialized. The $\mathbb{M}_{dist}$ model is then trained on the destination dataset $\mathbb{D}_{dist}$ updating the weights of all layers.

The *second model* $\mathbb{M}^2$ has the following structure. It also takes a 15N-dimensional sequence of words as input, which is then passed to an embedding layer of the same size $V_{src \cup dist}$xS. The embedding layer is connected to two subsequent recurrent layers. The last recurrent layer passes a time-distributed sequence of vectors encoding temporal dependences to a dense layer. Thereby, the dense layer outputs 15 $V_{src \cup dist}$-dimensional probability vectors that have to be additionally processed outside the network to actually predict the next activity. The output vectors are transformed to a sequence of 15 corresponding words according to maximum probabilities. The obtained sequence is compared to all M (M_{src} or M_{dist}) activities in the targeted dataset, which were transformed into word sequences beforehand, to find the closest one. The closest found activity is taken as a prediction. The distance between two word sequences is the sum of the distances between the words respecting the order of their appearance in the sequence. The distance

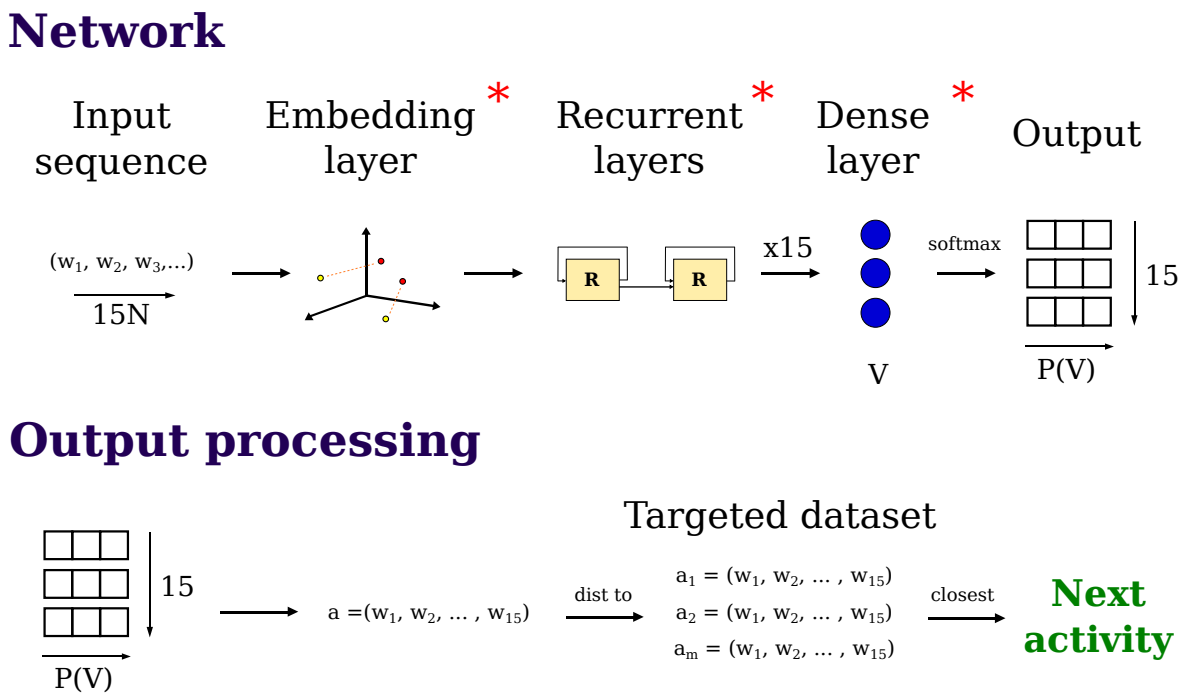

Figure 5.5: Second type of model for transfer. The weights of the layers marked with a star can be transferred.

between two words, in turn, is computed as a cosine distance between the corresponding embedding vectors. The embedding vectors used for search of the closest activity are the initial GloVe embeddings that have not been retrained neither on $\mathbb{D}_{\text{src}}$ nor on $\mathbb{D}_{\text{dist}}$. The transfer process is analogical to $\mathbb{M}^1$, except that the weights of the source dense layer are also exported and used to initialize the destination dense layer.

Our use of two different models may be explained by their properties that suit different purposes. The weights from the source model of the first type can only be partially transferred to the destination model since the dense layer is tied up with the number and content of activities in the source dataset. As shown in Chapter 3, despite a large fraction of common element instances, different procedures may still have a small number of common activities. That is why, making a dense layer consisting of $M_{\text{src}\cup\text{dist}}$ activities would be suboptimal. Withal, the main processual knowledge is supposed to be encoded in the recurrent layers, while the dense layer only serves to connect observations to the activities. The second model type, on the other hand, enables application of $\mathbb{M}_{\text{src}}$ to $\mathbb{D}_{\text{dist}}$ as it is. This property can be used to demonstrate the potential of transfer learning and to give an idea of how much of knowledge two procedures may have in common. However, it is less effective for the prediction, as the network can actually output a non existent activity.

5.4.3 Study design

The current study was designed to estimate how much knowledge can be transferred within a domain of neurosurgery (i.e., between neurosurgical datasets from the third chapter) and to find the best source-destination pairs of datasets in order to improve the prediction capacity of the model. Effective extraction and transfer of knowledge, however, require an analysis of surgical processes. That is why we conducted several experiments exploring the similarities and differences between the datasets in terms of surgical workflow and practices. Every experiment, detailed in the corresponding paragraph of Section 5.4.3.1, had its own purpose and examined a particular aspect of the surgical process. We also tested three types of transfer described in Section 5.4.3.2. All statistical comparisons of the results were performed using two-tailed Wilcoxon rank-sum test, where the results were considered significant for p-values $\leqslant 0.05$.

5.4.3.1 Experiments

Base line. This experiment let us establish a base line for the following experiments in order to be able to measure the improvements in prediction obtained by transfer learning. The $\mathbb{M}^1$ and $\mathbb{M}^2$ models were trained and tested separately on all six initial datasets (one at a time). Both models had the same parameters for all the datasets, except the number of neurons in the dense layer. In the $\mathbb{M}^1$ this number was equal to the number of unique activities M in the dataset, and in $\mathbb{M}^2$ to the number of unique words V. The number of activities N in the temporal context was set to 25 for all the datasets in this and all the

following experiments. The model provided the highest results (i.e., prediction accuracy average for all datasets) was chosen as the standard to which the results of the subsequent experiments were compared.

Mix. This experiment was performed to find out how putting different data together changes the prediction accuracy. It is different from transfer learning, as the model training is made on all the data at once. On one hand, the advantage of joining the data from different datasets is having more samples to learn from. On the other hand, each dataset has its own feature domain often quite different from all the others. This greatly rises the complexity of the learning problem. In this experiment, the data was mixed in four different ways. First, all interventions from both sites within one procedure were put together (i.e., L + R in ACDF, LDH and PA). This helps to understand, how similar every procedure in two sites is. Secondly, all interventions from one site including all three procedures were mixed together (i.e., ACDF + LDH + PA in L and R). This shows how similar the general practice in two sites is. Next, different pairs of procedures including interventions from both sites were tested as input (i.e., ACDF (L+R) + LDH (L+R), ACDF (L+R) + PA (L+R) and LDH (L+R) + PA (L+R)). This demonstrates general similarity/difference between two procedures. Finally, all the data was joined to form one input dataset (i.e., ACDF.L + ACDF.R + LDH.L + LDH.R + PA.L + PA.R) in order to see how close all the neurosurgical procedures are. The models $\mathbb{M}^1$ and $\mathbb{M}^2$ were trained and tested separately on each type of data union. The vocabularies of unique activities and words at each time included the activities and the words found in the mixed datasets only. For each type of mix, the model provided the highest results was compared to the base line standard by computing the difference Δ between their average values. The third mix type was also compared to the first one to measure the actual increase in accuracy.

Split. The initial datasets contained the operations performed by both senior and junior surgeons. This experiment, splitting the data by surgeon's expertise level, was performed to see if the prediction can be improved by separating seniors from juniors. Splitting the data may seem counter-intuitive, as it reduces the number of training samples in each set. However, separating interventions from surgeons having different expertise can actually also reduce the variability in workflow, thus, the complexity of the learning problem. In this experiment, three different splits were made. The first split divided the data on interventions performed by juniors and seniors within each procedure (i.e., ACDF (J), ACDF (S), LDH (J), LDH (S), PA (J) and PA (S)). The second split was made within each surgical site (i.e. L (J), L (S), R (J) and R (S). The third split divided all the initial data (i.e., D (J) and D (S)). These splits let us perceive the difference between the expertise levels in different contexts: procedure-related, hospital-related and in general. The models $\mathbb{M}^1$ and $\mathbb{M}^2$ were trained and tested separately on each type of data split. The vocabularies of unique activities and words were kept identical for both levels, as their difference was minor. For each

type of split, the model provided the highest results was compared to the base line standard. The first, second and third splits were also compared to the first, second and fourth mixes respectively.

Raw use. This experiment served to estimate the amount of knowledge shared between the datasets that could be used without any additional training. For this, the entire model was first randomly initialized and tested on the destination dataset $\mathbb{D}_{\mathtt{dist}}$. This was needed to measure how many predictions could be correctly made without any knowledge at all. Then, the model was trained on the source dataset $\mathbb{D}_{\mathtt{src}}$ and tested on the destination dataset $\mathbb{D}_{\mathtt{dist}}$ again without being retrained on it. This shows how much knowledge learnt on one dataset is useful to predict the activities from another dataset. The pairs of tested source and destination datasets were the same as in the next experiment on actual transfer learning, and are described in details in Section 5.4.3.2. Only the $\mathbb{M}^2$ model was used in this experiment, since only its architecture enabled direct model use on an another dataset with a different feature space. However, the word vocabulary used in the embedding and dense layers included the words from both source and destination datasets.

Transfer. We conducted this experiment to measure the improvement in prediction accuracy provided by the actual transfer learning. For this, the source model $\mathbb{M}_{\mathtt{src}}$ was first trained on the source dataset $\mathbb{D}_{\mathtt{src}}$. Its weights were extracted and saved apart. They were then used as initialization for the destination model $\mathbb{M}_{\mathtt{dist}}$, which was then trained and tested on the destination dataset $\mathbb{D}_{\mathtt{dist}}$. This shows how the knowledge from one dataset can improve the training process and prediction results for another dataset. We performed three types of transfer using different combinations of source and destination datasets: inter-site, inter-procedure and inter-expertise transfers. Section 5.4.3.2 defines each type of transfer. Only the $\mathbb{M}^1$ model was used, as it proved to provide better results within one feature space. That is why, only the weights from the embedding and recurrent layers were transferred, and each activity vocabulary was restricted to the activities from one input dataset only. For each type of transfer, the results were compared to the base line standard and to the corresponding mixes and splits as well.

5.4.3.2 Types of transfer

In this study, we examined three different types of transfer: inter-site, inter-procedure and inter-practice transfers (Table 5.4). The *inter-site transfer*, which can also be considered as intra-procedure transfer, was performed to estimate the efficiency of transfer between the interventions performed in different hospitals belonging to the same procedure. This type of transfer involved the following dataset pairs: L $\rightarrow$ R and R $\rightarrow$ L for ACDF, LDH and PA. In these pairs, the source dataset is on the left of the arrow sign, and the destination dataset is on its right. The *inter-procedure transfer* was performed to estimate the efficiency of transfer between different procedures belonging to the same surgical domain. This type

Table 5.4: Source and destination datasets for different types of transfer. Plus sign indicates a mix of datasets

Inter-site transfer		**Inter-procedure transfer**	
$\mathbb{D}_{src}$	$\mathbb{D}_{dist}$	$\mathbb{D}_{src}$	$\mathbb{D}_{dist}$
ACDF.L	ACDF.R	ACDF.L + ACDF.R	LDH.L + LDH.R
ACDF.R	ACDF.L	LDH.L + LDH.R	ACDF.L + ACDF.R
LDH.L	LDH.R	ACDF.L + ACDF.R	PA.L + PA.R
LDH.R	LDH.L	PA.L + PA.R	ACDF.L + ACDF.R
PA.L	PA.R	LDH.L + LDH.R	PA.L + PA.R
PA.R	PA.L	PA.L + PA.R	LDH.L + LDH.R

Inter-expertise transfer		
Within	$\mathbb{D}_{src}$	$\mathbb{D}_{dist}$
procedure	ACDF.L(J) + ACDF.R(J)	ACDF.L(S) + ACDF.R(S)
	ACDF.L(S) + ACDF.R(S)	ACDF.L(J) + ACDF.R(J)
	LDH.L(J) + LDH.R(J)	LDH.L(S) + LDH.R(S)
	LDH.L(S) + LDH.R(S)	LDH.L(J) + LDH.R(J)
	PA.L(J) + PA.R(J)	PA.L(S) + PA.R(S)
	PA.L(S) + PA.R(S)	PA.L(J) + PA.R(J)
site	ACDF.L(J) + LDH.L(J) + PA.L(J)	ACDF.L(S) + LDH.L(S) + PA.L(S)
	ACDF.L(S) + LDH.L(S) + PA.L(S)	ACDF.L(J) + LDH.L(J) + PA.L(J)
	ACDF.R(J) + LDH.R(J) + PA.R(J)	ACDF.R(S) + LDH.R(S) + PA.R(S)
	ACDF.R(S) + LDH.R(S) + PA.R(S)	ACDF.R(J) + LDH.R(J) + PA.R(J)
D [1]	D(J)	D(S)
	D(S)	D(J)

of transfer involved the pairs as ACDF $\leftrightarrow$ LDH, ACDF $\leftrightarrow$ PA, LDH $\leftrightarrow$ PA. The double-side arrow means that the transfer was made from the left (source) dataset to the right (destination), and the other way around. The *inter-expertise transfer* was performed to estimate the efficiency of transfer between the interventions performed by the surgeons having different levels of expertise (i.e., seniors and juniors). For this transfer type, the tests were performed separately for the procedures (i.e. pairs J $\leftrightarrow$ S for ACDF, LDH and PA), sites (i.e. pairs J $\leftrightarrow$ S for L and R) and composition of all initial datasets (i.e. pair J $\leftrightarrow$ S for D).

1. D represents composition of all initial datasets that is ACDF.L + ACDF.R + LDH.L + LDH.R + PA.L + PA.R

5.4.4 Results

5.4.4.1 Base line

The experiment, putting all the initial datasets in the same conditions, showed that at best the next activity could be predicted with accuracy of 78.91±7.53% on average for all datasets (see Table 5.5). This result was provided by the $\mathbb{M}^1$ model and was taken as the standard to which all the following advances were compared. As expected, the model $\mathbb{M}^2$ was less accurate (76.78±8.41%). The PA activities were predicted the best and those belonging to LDH the worst. Better predictions were also made for the procedures performed in Rennes than Leipzig. Lower results may indicate a higher complexity of the learning problem (including factors as the number of unique words, elements and activities, procedure length, number of possible workflows, etc.), but also a higher variability in the data. This variability is partially inter-personal, meaning that it comes from the number of different surgeons performed the interventions.

Table 5.5: Base line prediction accuracy (in %) for six initial datasets

		ACDF.L	ACDF.R	LDH.L	LDH.R	PA.L	PA.R	Avg
Avg	$\mathbb{M}^1$	74.35	78.73	74.46	75.23	84.12	86.57	78.91
	$\mathbb{M}^2$	71.17	77.95	73.13	73.00	82.04	83.36	76.78
Std	$\mathbb{M}^1$	9.77	8.32	9.10	8.12	5.37	4.76	7.53
	$\mathbb{M}^2$	10.14	8.27	9.25	9.87	6.39	6.55	8.41

5.4.4.2 Mix

Mixing the interventions belonging to the same procedure but performed at different surgical sites provided at best 81.42% of accuracy, generating a 2.51% boost compared to the base line standard (Figure 5.6). This score was obtain using $\mathbb{M}^1$ model. The $\mathbb{M}^2$ model provided a comparable but a slightly lower score of 81.27% (Δ = 2.36%). The increase of the learning problem complexity due to a mix of different future spaces was bigger for $\mathbb{M}^1$ and smaller $\mathbb{M}^2$ because of their differences in architecture and data handling. That is why, in this case, $\mathbb{M}^2$ was able to catch up the $\mathbb{M}^1$. LDH had the biggest increase in accuracy, while ACDF the lowest. The higher increase in accuracy may indicate the bigger similarity between two sites within a particular procedure. The Δ of all three procedures were statistically significantly different from each other. All new resulting accuracies, except for ACDF using the $\mathbb{M}^2$ model, were also statistically significantly different from the base line.

Mixing all procedures within one site only decreased the prediction accuracy (Figure 5.7). The $\mathbb{M}^1$ model showed an accuracy of 75.89% loosing 3.02% compared to the base line standard. The $\mathbb{M}^2$ generated 76.19% loosing 2.72% compared to the base line. This time,

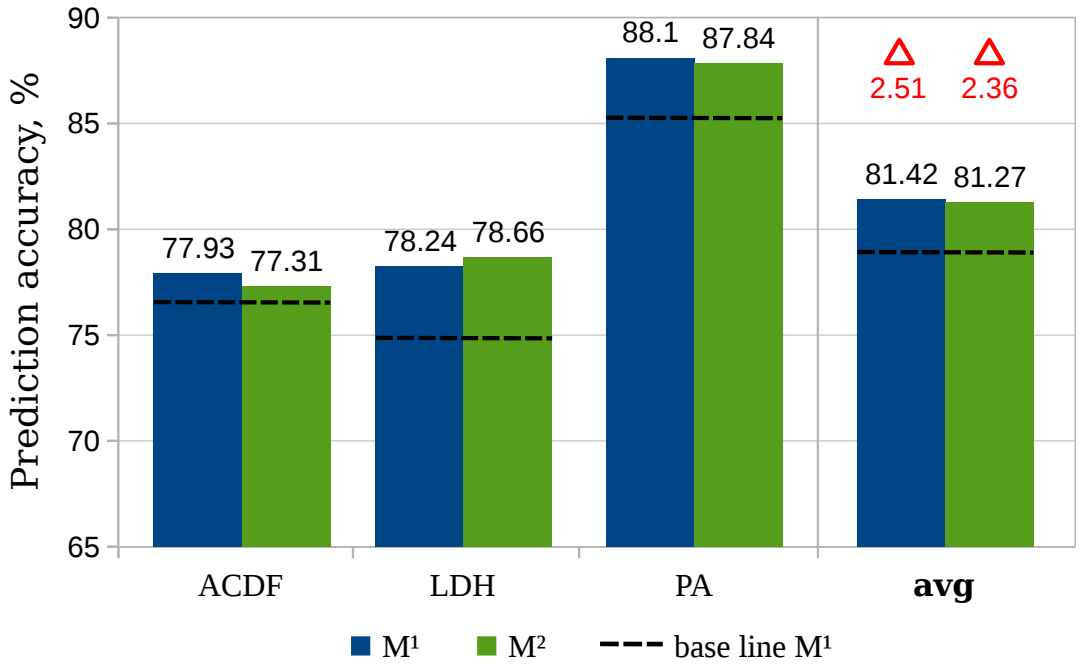

Figure 5.6: Mix of sites within one procedure

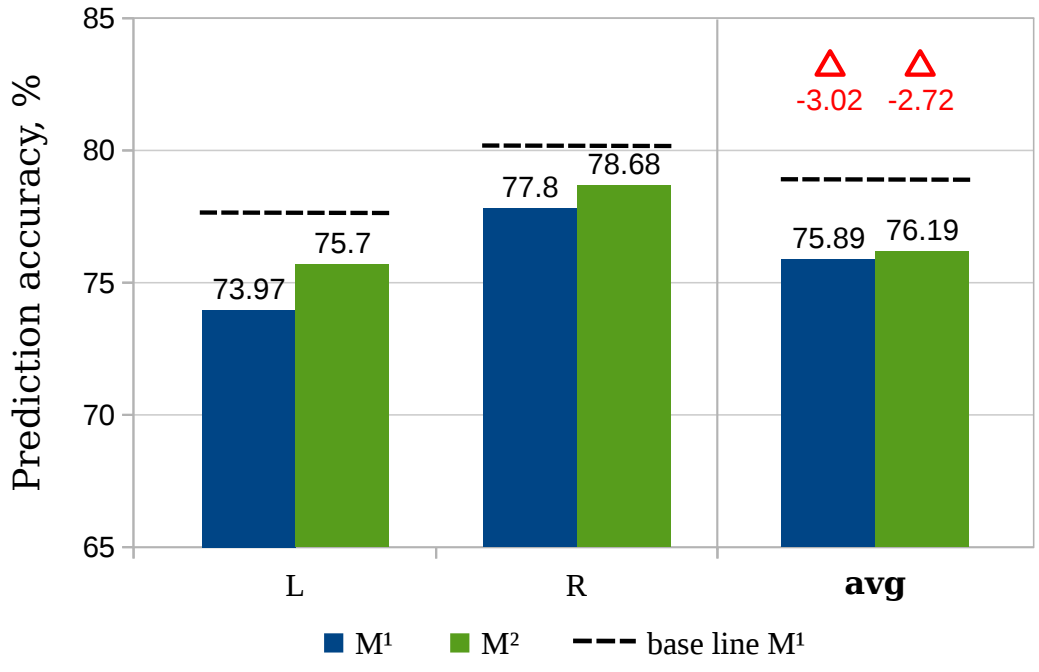

Figure 5.7: Mix of surgeries within one site

$\mathbb{M}^2$ provided better results than $\mathbb{M}^1$ due to even bigger dissimilarity between the mixed datasets in terms of unique activities but less in terms of words. Comparing to the base line scores, the datasets from Leipzig lost in accuracy more than those from Rennes (3.67% vs. 2.37% for $\mathbb{M}^1$ and 1.97% vs. 1.49% for $\mathbb{M}^2$). The statistical significance of this difference let us presume that these three procedures are more alike in Rennes than in Leipzig. This time, all new resulting accuracies for both mixes and both models were also statistically significantly different from the base line.

Mixing different procedures from both sites provided at best 79.96% of average accuracy with Δ = 1.05% compared to the base line standard (Figure 5.8). This result was provided by the model $\mathbb{M}^2$ yet it had no statistically significant difference with the $\mathbb{M}^1$ that gave 79.32% with 0.41% Δ. The combination ACDF + LDH had 4.98% gain over the base line (statistically significant), while the ACDF + PA lost 2.16% (significant), and LDH + PA gained 0.30% only (not significant). As the sites were mixed together with the procedures, it would be fair to compare the results to those from Figure 5.6, by taking an average score of two procedures each time, in order to determine the actual Δ. Such wise, the mix of procedures actually lost in accuracy with average Δ = -1.46%, and only the ACDF + LDH combination improved prediction performance of 3.62%. These results allow us to believe

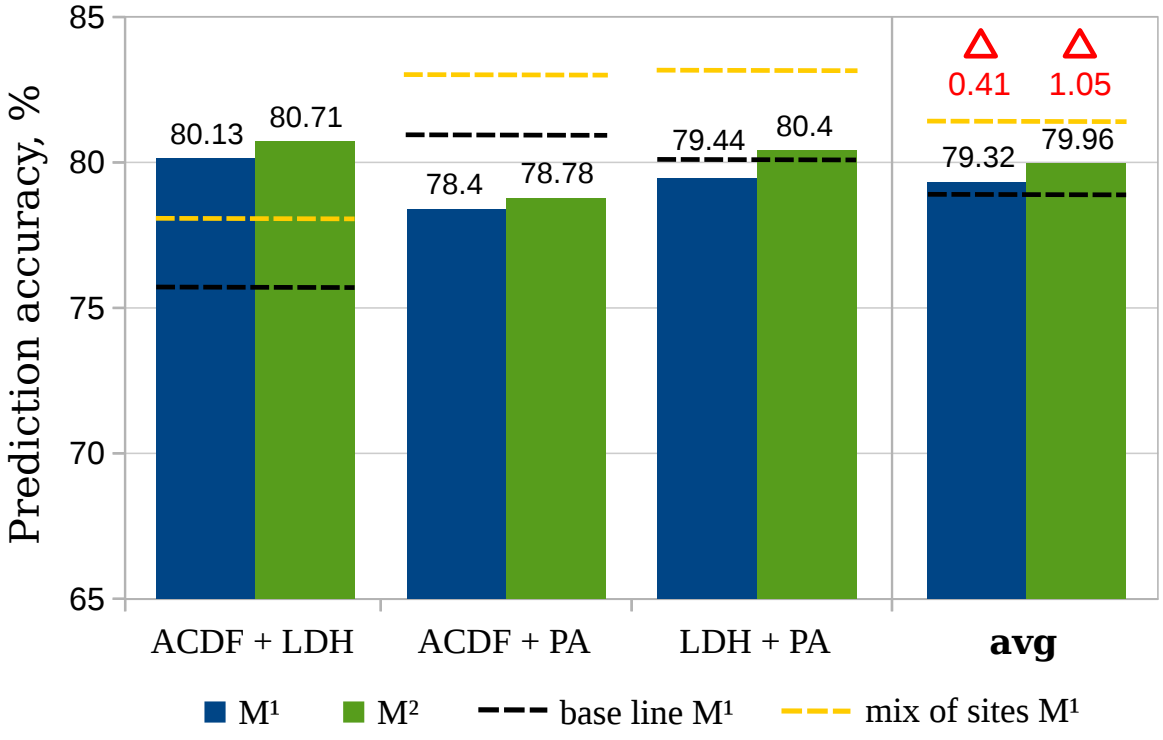

Figure 5.8: Mix of different procedures from both sites

that the ACDF and LDH procedures have much in common, but ACDF and PA, as well as LDH and PA are too different to be mixed together.

Mixing all initial datasets together also only deteriorated the prediction ability of both models, $\mathbb{M}^1$ more than $\mathbb{M}^2$. The model $\mathbb{M}^1$ showed 75.73% accuracy with Δ = -3.18%, and $\mathbb{M}^2$ 77.54% accuracy with Δ = -1.37%. This mix had the highest data variability in terms of unique activities which was too hard to handle for $\mathbb{M}^1$, that is why $\mathbb{M}^2$ provided statistically significant higher accuracy and lower decrease compared to the base line.

5.4.4.3 Split

Splitting interventions within one procedure by expertise level provided the average accuracy of 80.16% and a 1.15% improvement compared to the base line (Figure 5.11). This results were provided by the $\mathbb{M}^1$ model. The model $\mathbb{M}^2$ had slightly lower but comparable results of 79.94% (Δ = 1.03). For all procedures, the results for senior surgeons were higher than for juniors, but only for PA this difference was statistically significant. However, no junior surgeon performed PA in Rennes. That is why, in this case, a low score was probably caused by a drastic decrease of the amount of training samples for the junior group. It also make sense to compare the results with those from Figure 5.6 since the interventions from different sites were mixed within a procedure. Hereby, splitting by expertise levels actually deteriorated the results compared to the mix of sites (Δ = -1.26% on average). This can be explained by the fact that for the same complexity of the learning problem (identical vocabularies of unique activities and words), reduction of training samples only causes degradation of the prediction performance of the model. Yet, thanks to the split, each group had less variability in it which prevented the accuracy from falling to much.

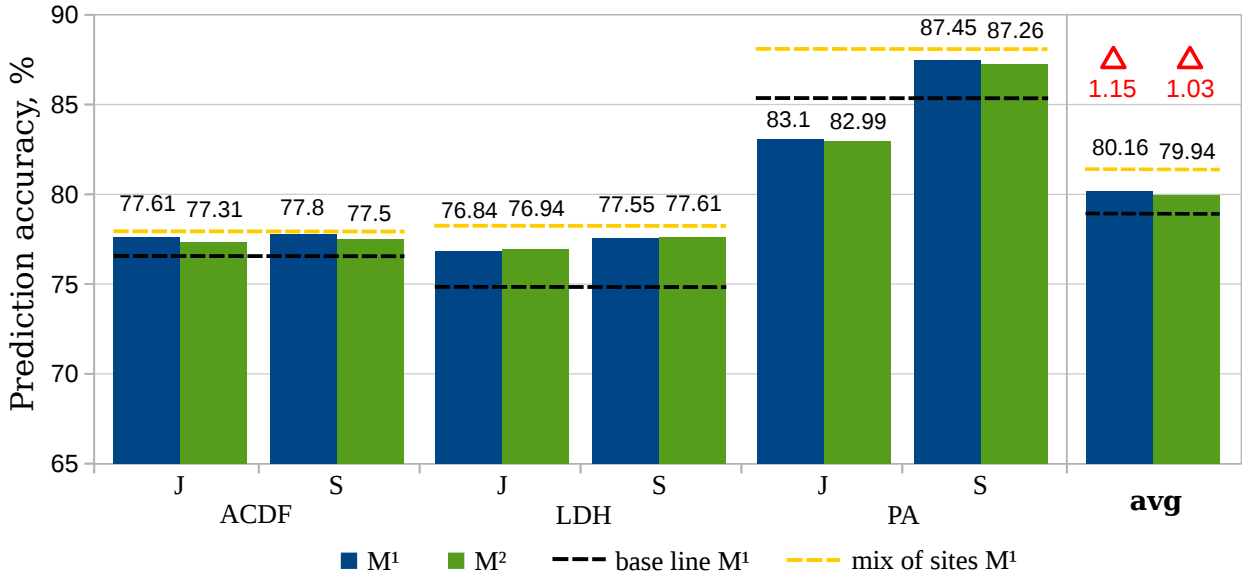

Figure 5.9: Split by expertise level within a procedure

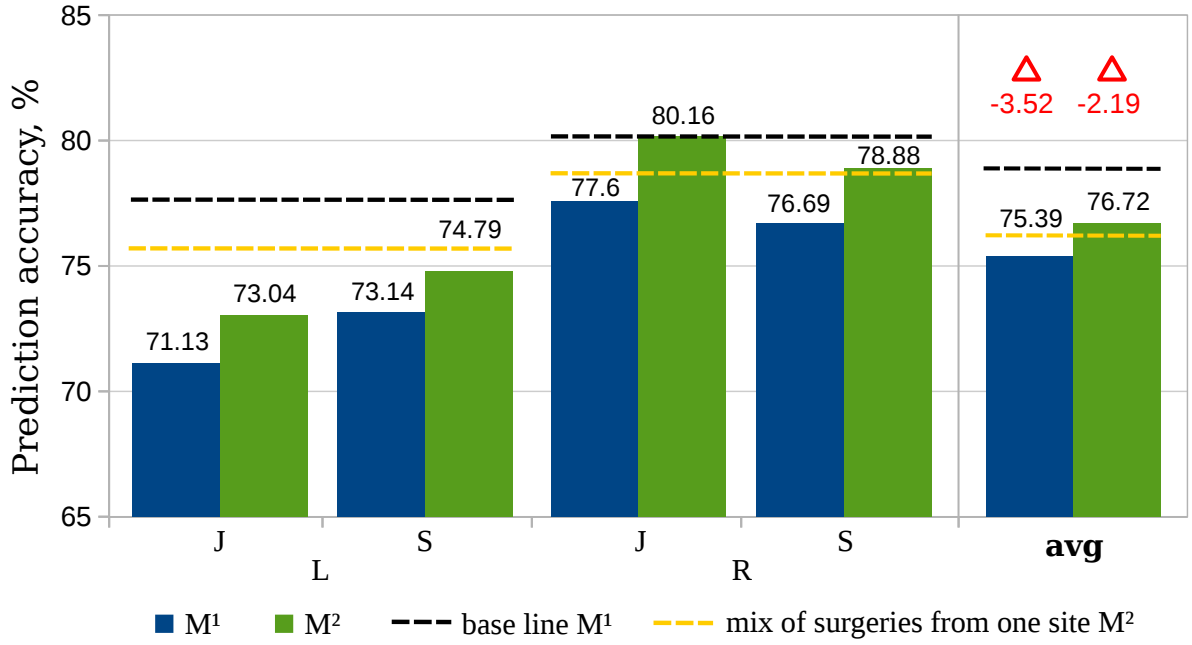

Figure 5.10: Split by expertise level within a site

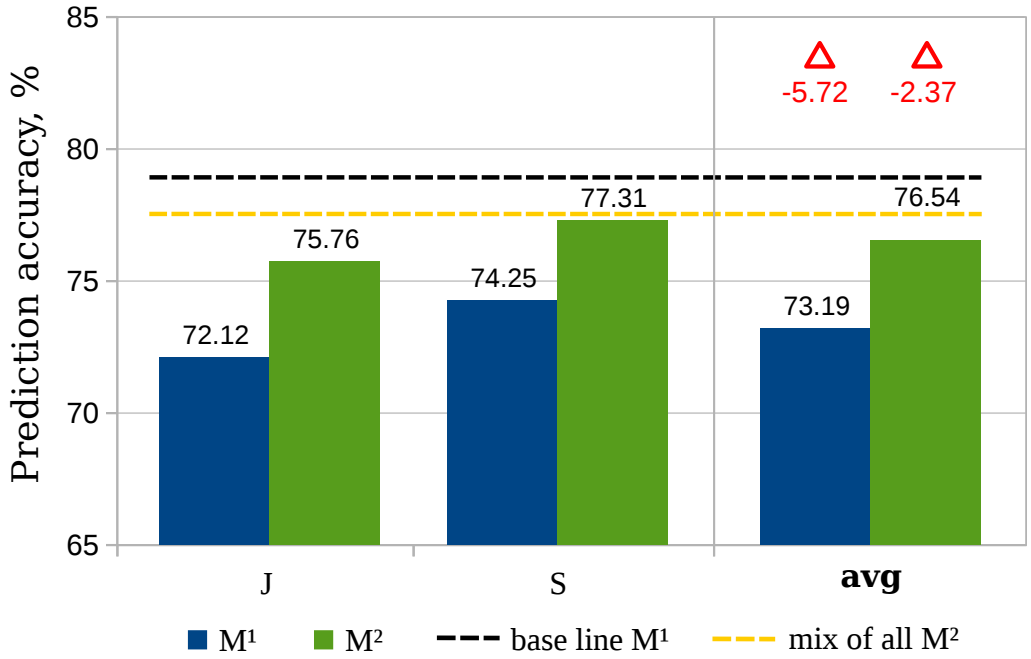

Figure 5.11: Split by expertise level of all data

Splitting by expertise level within a surgical site decreased the accuracy of both models compared to the base line (Figure 5.11). The model $\mathbb{M}^2$ allowed the least impairment giving 76.72% accuracy and Δ = -2.19%, while $\mathbb{M}^1$ 75.39% with Δ = -3.52%. Comparing to results to the mix of all procedures from one site (Figure 5.6), the model $\mathbb{M}^2$ actually had a gain of 0.53% (not significant). However, this gain was caused by a spike in $\mathbb{M}^2$ results for R (J) group. In this group, the PA procedure was not represented at all (no junior surgeon operated), that is why the complexity of the problem was much lower than in other cases.

Splitting all the data by expertise level only worsened the situation (Figure 5.11). The model $\mathbb{M}^2$ had the smallest decrease in accuracy compared to the base line (76.54% in average with Δ = -2.37%), while $\mathbb{M}^1$ had 73.19% accuracy and Δ = -5.72%. The prediction performance also dropped compared to the mix of all data (Δ = 1.00%). In this experiment, the difference between the expertise levels became statistically significant due to a very high complexity of the learning problem (three procedures from both sites put together). In the view of this result, it is also possible to assume that not only juniors are different from seniors, but the dissimilarity inside a group is higher for juniors than seniors.

5.4.4.4 Raw use

In the raw inter-site use, the model trained on one site was used (without any additional training) to predict the activities of the same procedure from the second site. The results can be found in Table 5.6. On average for all procedures and source-destination pairs, raw use provided predictions with 36.12% of accuracy, while a random initialization of the model only 5.77%. It shows that roughly 30% of knowledge is shared between the sites, although not all of it may be useful for improving prediction accuracy of the model. As in the experiment with the mix of sites, a raw model use within the LDH procedure provided the best results, and within ACDF the worst. This confirms our observation that the surgical practises in Leipzig and Rennes are closer for LDH procedure, then, for PA and, finally, for ACDF.

Table 5.6: Prediction accuracy (in %) for raw inter-site use

	ACDF	LDH	PA	Avg
L -> R	30.14	38.82	36.23	36.12
R -> L	33.18	43.97	34.49	

In the raw inter-procedure use, the model trained on one procedure including both sites was used (without any additional training) to predict the activities of an another procedure. The results of the experiment are exposed in Table 5.7. On average for all source-destination pairs, the raw inter-procedure use demonstrated 26.08% accuracy, while a random initialization of the weights provided only 5.02%. However, the learnt knowledge was

Table 5.7: Prediction accuracy (in %) for raw inter-procedure use

		Destination			
		ACDF	LDH	PA	Avg
Source	ACDF	-	42.30	17.25	26.08
	LDH	40.12	-	20.46	
	PA	16.77	19.55	-	

Table 5.8: Prediction accuracy (in %) for raw inter-expertise use

	ACDF	LDH	PA	Avg	L	R	Avg	D	Avg
J -> S	60.72	58.14	65.61	62.81	56.75	40.33	51.66	54.19	55.01
S -> J	63.12	60.49	68.78		56.40	53.14		55.83	

much more useful in the case of the ACDF $\leftrightarrow$ LDH combinations than ACDF $\leftrightarrow$ PA or LDH $\leftrightarrow$ PA. This demonstrates again, that ACDF and LDH are the most similar procedures, and ACDF and PA are the most dissimilar.

The results of the raw inter-expertise use, when the model was trained on one expertise level to be applied on another, are in Table 5.8. The experiment showed, that learnt knowledge was more useful when applied within a procedure (62.81% on average), than within a site (51.66%) or with all the data (55.01%). A random initialization provided an accuracy of 4.51% on average. The results for J $\rightarrow$ S use inside Rennes were much lower than the others because that source model did not have any PA samples (no junior surgeon operated) to learn from, thus, was unable to correctly predict its activities from the destination dataset. Giving generally high accuracy values (higher than in other types of raw use), the following assumption can be made. The knowledge encoding the difference between the expertise levels is less important for a correct activity prediction, or, in another words, this type of knowledge is less captured by LSTM models.

5.4.4.5 Transfer

The results of the inter-site transfer with additional training on the destination dataset are shown in Figure 5.12. On average for all procedures and source-destination combinations, the prediction performance of the model achieved 85.97% having a statistically significant increase of 7.06% compared to the base line, and 4.55% compared to a simple mix of sites (significant as well). As expected, LDH benefited from the knowledge transfer the most. However, ACDF had a more important increase that PA. It shows that even if two sets of data are less similar, one can still have some form of knowledge (common or not) that enhances the learning process of the second one.

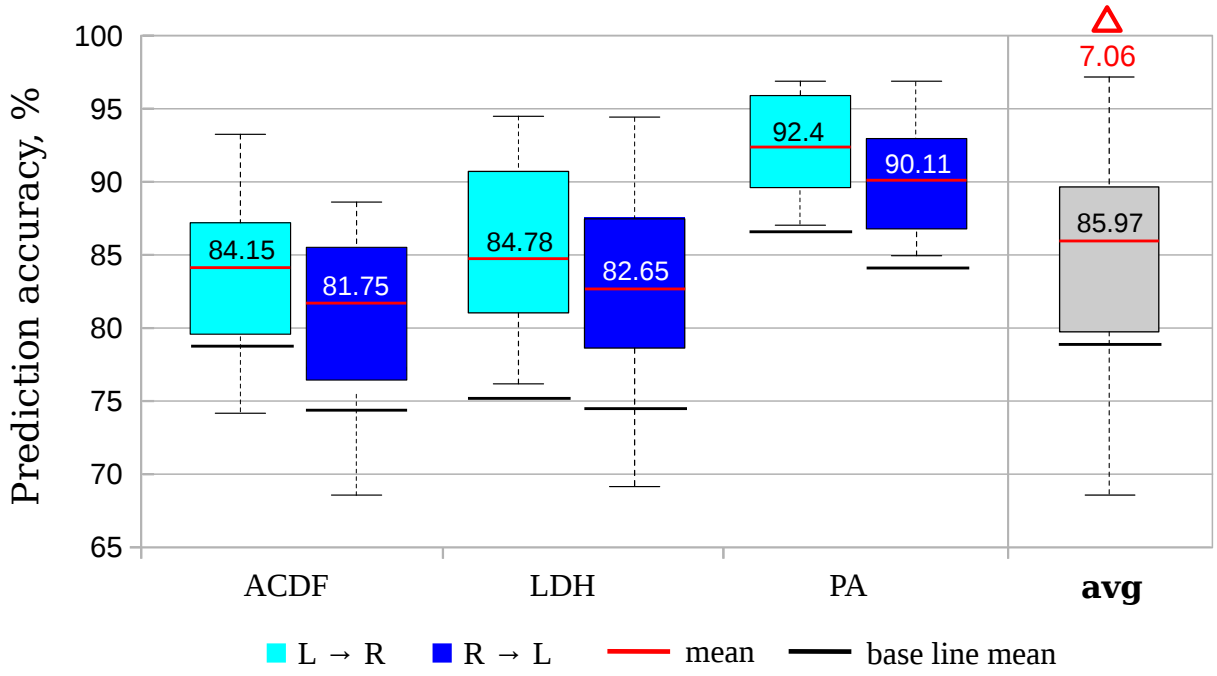

Figure 5.12: Inter-site transfer within a procedure

The results of the inter-procedure transfer are exposed in Figure 5.13. On average for all tested combinations of source and destination datasets, the prediction accuracy reached 86.49% with a statistically significant boost of 7.58% compared to the base line, and 6.53% compared to a simple mix of procedures (significant). However, 2.51% of this boost came from the mix of both sites in the source and destination datasets. Thus, the actual increase in accuracy equalled 5.07%. On the other hand, if we choose only the most appropriate procedures for transfer (e.g., ACDF for LDH, LDH for ACDF and ACDF for PA), the average accuracy can be recomputed to 89.09% with the Δ of 10.18% for the base line and 7.67% for the sites mix. The highest improvement was made for LDH when transferring knowledge from ACDF, and the lowest for ACDF when transferring from PA. This experiment also demonstrates, that even if the source and destination procedures are quite different from each other, both still encode some fundamental knowledge about the surgical process in general that can help in training.

The results of the inter-expertise transfer can be found in Figure 5.14, 5.15 and 5.16. The experiment showed that the transfer within a procedure provided 80.34% accuracy on average and made an improvement of 1.43% (not statistically significant) compared to the base line but only 0.28% (not significant) in comparison with a simple split. The transfer within a site actually impaired the results providing 76.09% which is 2.82% smaller than the base line (significant) and 0.63% smaller than the split (not significant). The transfer

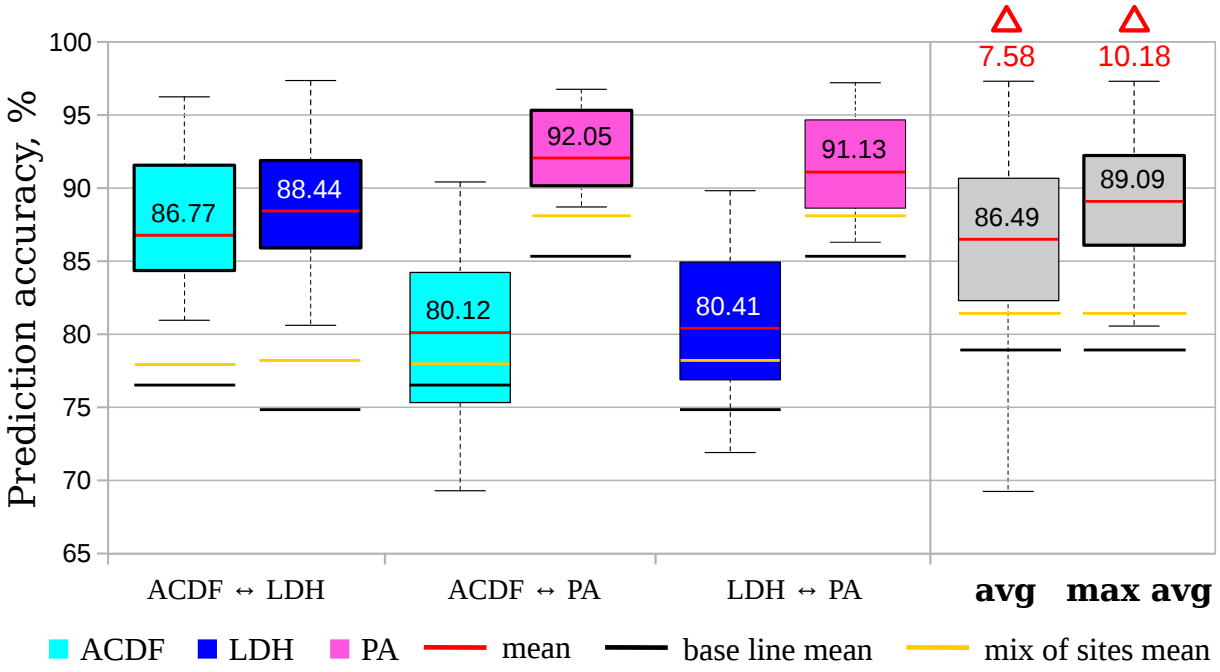

Figure 5.13: Inter-procedure transfer

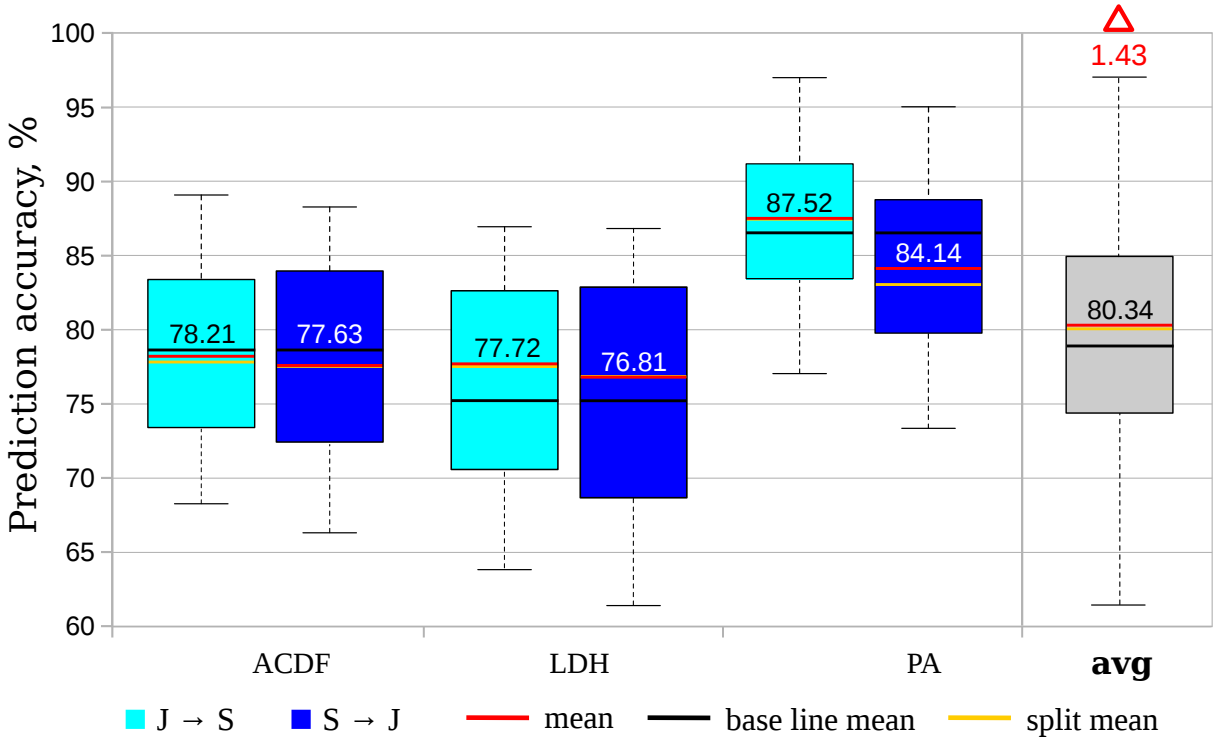

Figure 5.14: Inter-expertise transfer within a procedure

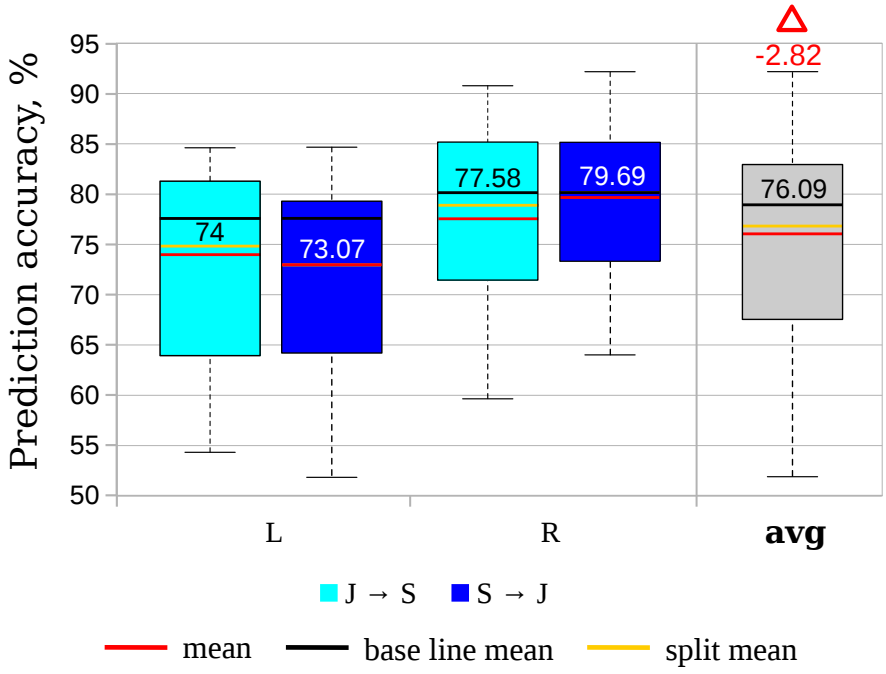

Figure 5.15: Inter-expertise transfer within a site

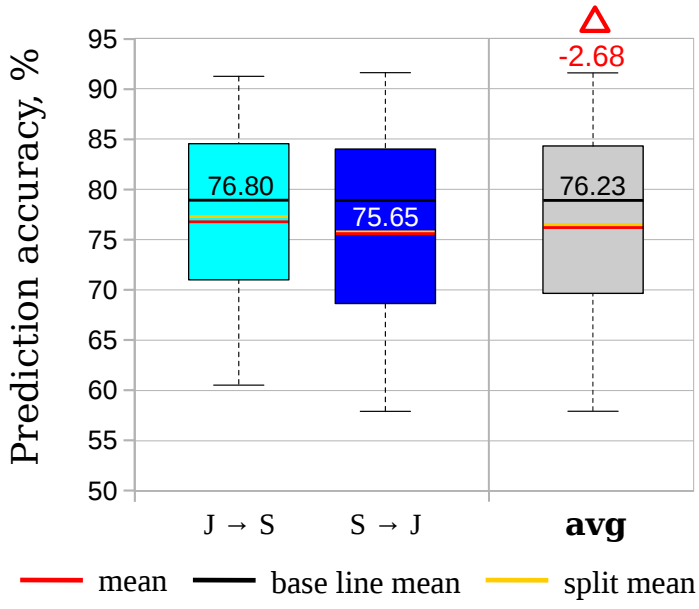

Figure 5.16: Inter-expertise transfer for all data

from one expertise level to another using all data also led to an impairment of prediction: 76.23% accuracy with Δ = -2.68% for the base line (significant) and Δ = -0.31% for the split (not significant). The transfer causing a decrease in accuracy is a good example of negative transfer, a well-known phenomenon [Pan 2010]. This demonstrates that some knowledge (related to the expertise level in this case) should not be transferred as it may hurt the learning process. In all cases, the difference between J $\rightarrow$ S and S $\rightarrow$ J transfers was not statistically significant.

5.5 Discussion

In this chapter, we proposed two methods of knowledge transfer to overcome the problem of data deficit. Thanks to both of these methods and extensive experiments conducted on clinical data, the performance of a basic LSTM model was increased by almost 22%. On average for all used datasets, the accuracy of prediction reached 89%. Forestier et al. in their work on activity prediction for the LDH and ACDF procedures [Forestier 2017] were able to attain 95% accuracy. The authors, however, were interested in prediction of right hand activities only, which represents a simpler problem. Yet, knowing the activities of both hands is important for a complete understanding of the situation. For instance, too frequent use of suction may indicate an excessive blood loss. On the other hand, the improvement of prediction by means of knowledge transfer but not a high performance was the main objective of our work. Similarly to the work on sensors and signals presented in the last chapter, only a very basic LSTM architecture was used. A more elaborated architecture would probably provide better results. Moreover, the used approaches of knowledge transfer may easily be applied to any surgical workflow related problem (e.g., recognition, analysis, etc.). Another contribution of this work consists of multiple observations made during the experiments that allowed a better understanding of surgical practices. These observations are discussed in Section 5.5.3.

5.5.1 Word embeddings

We demonstrated that the word embeddings trained on corpora constructed from medical and scientific texts indeed contained certain semantic meaning that enabled a boost of the learning process and played an important role in knowledge transfer. Unfortunately, the embeddings encode information in a human-unreadable way. Another drawback is their inability to solely encode only chosen parts of the information. It means that some useless knowledge gets probably encoded along the way as well. Since the initial embeddings were trained for a different learning objective (word context discovery), their representation still needed to be retrained within the LSTM model for the activity prediction task. Despite that, the initially encoded information contained in them guided the training process in the right direction.

Generally, the word corpora used for embedding training in NLP are much bigger than those that were used in this study (billions vs. millions of words). The relatively small corpora created in this study were an attempt to test their impact on the prediction of activities. The experiments showed that the success of training grew along with the corpus size. That allows us to assume that the results can be further improved with bigger corpora. The created corpora were based on post-operative transcriptions and scientific articles on a related subject. It would also be interesting to create a word corpus specifically dedicated to describe the surgical process. For instance, on-line commenting of an actual process would take less effort than annotation of the same amount of interventions.

The best results were obtained with bigger sizes of embedding vectors. However, in this study, the size did not exceed 500. In the literature, the bigger size is normally considered when the corpus contains several billions of words [Mikolov 2013]. The importance of vector's size regarding the number of learning words was also demonstrated: a small MT corpus provided much lower results with vectors of size 300 and 500. Moreover, the vectors are encoded differently (have quite different values) by word2vec and glove methods. It shows that different information gets encoded, but in practice, the overall relationships between the words are preserved.

5.5.2 Transfer learning

The method of transfer learning was used in this work to increase the prediction capacity of a deep neural network by transferring processual knowledge. The studies exploring different ways to operate the data (i.e., mix, split, raw use and transfer) revealed both positive and negative manipulations discussed below.

Positive experience. The manipulation "mix of sites + inter-procedure transfer", consisting of two steps enabled the best enhancement. First, the interventions from different surgical sites but belonging to the same procedure were mixed together in one input dataset used for pre-training. In the second step, the internal parameters of the model trained on the mixed data (i.e., weight matrices containing the learnt knowledge) were transferred to the model which was trained on another procedure. The most effective transfer happened, of course, between the procedures that resemble the most - ACDF and LDH. However, the transfer between less related procedures, as ACDF and PA or LDH and PA, was still effective. This showed that even if two procedures have very few activities in common, they still can contain certain fundamental knowledge or common information about the surgical process which can turn out to be very helpful. Another manipulation that gave slightly lower but still good results was the inter-site transfer. There were four possible reasons why it conceded to the inter-procedure transfer. First the the inter-procedure transfer prior benefited from the mix of the surgical sites that also boosted the performance. Secondly, thanks to the mix of data, the models involved in the inter-procedure transfer had more training data than the models in the experiments with the inter-site transfer. In third place, other

procedures may contain some useful knowledge not present in the same procedure from another site. Finally, the 10% boost of the inter-procedure transfer was computed using only the best source-destination combinations. Whereas, in inter-site transfer only one source option existed for each destination dataset.

Negative experience. The negative manipulations included the split of data by expertise level and inter-expertise transfer. Splitting the data in current conditions negatively reflected on the prediction performance, probably because it brutally reduced the number of training samples each time. Yet, given a sufficient amount of data, it potentially could provide positive results by reducing the variation in the way the surgery is performed, making two clusters of data more homogeneous. Nevertheless, one should be careful with such a split. The interventions performed by junior surgeons seem to be less consistent that those from the senior group. It is possible that when used together in one dataset, the senior group helps to balance out the variation inside the junior group. It would be interesting to see what happens when the data is split by expertise level within the most homogeneous data - only one procedure and one site (i.e., within initial datasets). This has not been done because no junior surgeon operated pituitary tumour in Rennes. The split within a procedure made only a not significant improvement of the results. Whereas, the expertise split of the data mixing different procedures together made the performance degrade. Such composition of data turned out to be too confusing for the network. It is, however, interesting to see that a raw inter-expertise use gave good results but the the transfer did not work. It means that both for juniors and seniors the models learns very similar information, and the expertise-related knowledge from one group is not helpful to another. This particular knowledge from another group may even hurt the training process and cause the phenomenon of negative learning. Yet, this problem should be researched more to have a clearer understanding of its influence.

Learning process. Deep neural network is an intricate mechanism, where a balance between all the parameters has to be found. The quantity of training data is well known to be a great factor influencing the results of training. However, the requirement of an important amount of data is not the only condition for a successful training. Its success also depends on the complexity of the learning problem including factors as number of output classes (number of unique activities in our case), complexity of temporal dependences to discover, as well as inter and intra class variability. That is why the complexity of the problem grows when putting too much different data together. For instance, mixing dissimilar procedures increases the amount of possible activities, which increases the risk of a bad choice. Different procedures have different sequencing of activities leading to higher variability in workflow. It means that the model basically has to learn two different types of temporal dependences at once. That is why, learning separately on a more uniform set of data describing only one type of dependences is better. In such conditions, the network learns more efficiently and extracts more of useful information. Using that information,

another network pays more attention to the particularities of its own dataset resulting in a stronger performance. This explains better results for the inter-procedure transfer than the mix of procedures.

Knowledge. We demonstrated that LSTM network is capable of learning and transferring surgical process knowledge. This knowledge probably contains both procedure dependent and independent features. The procedure dependent features characterize the terms and concepts specific to a given surgery. The procedure independent features, conversely, represent the information about the temporal sequentiality inherent to any surgical process. Unfortunately, for now, a neural network, especially a recurrent, is a black box. It is difficult to see what happens inside and to know how, why and what exactly it learns. Inability to explain network's functioning has been a major critique of deep learning methods. Understanding of the learning process is a hot research topic nowadays. However, current research is mainly interested in visualization of the process for image-based data, and much less in sequential or other types of data. In consequence, the conclusions made about the transfer strategies and surgical practices can only be made by observing the correlation between the input and the output. The encoding of knowledge in a form not readable by human is another drawback of the method. The encoded knowledge is not formalized, and for now can only be used within deep net architectures. Hopefully, one day, it will be possible to combine it with formal representation of knowledge (e.g., ontology).

Perspectives. The results obtained in this work are encouraging but leave a room for further exploration. In the future, a multi-level transfer has to be tested. It consists in training on one dataset, using the weights as initialization for training on a second dataset, and transferring the weights again for training on a third dataset. Such wise, the inter-site transfer can be done first, and then followed by the inter-procedure transfer. Another option is subsequently training on all three procedures, or a mix of procedures as a source. In any case, a good order and combination has to be found first. In addition, it is also possible to try importing the weights of different recurrent layers separately. The study results also allow us to think that the surgeries from other specialities can also be used for transfer. Furthermore, there is also a hope that the information about the process in general, as a sequence of actions towards an objective, gets encoded as well. In that case, pre-training on annotations of different sorts of processes (not necessarily related) would also help. This opens the doors to many opportunities, as the acquisition of such process models is less constrained than in the surgical field, and may be used to create massive datasets to learn from.

5.5.3 Surgical practice

Along with the analysis of different types of transfer, we also made interesting observations about surgical practices. The results obtained during the conducted experiments suggested that the workflows of interventions performed by the surgeons having different levels of expertise are closer to each other than two interventions of the same type performed in different hospitals. Withal, the interventions from different hospitals are more alike than the interventions from different procedures. It means that the difference between the expertise levels is lower than the difference between the surgical sites, which in turn, is lower than the difference between the surgical procedures. Indeed, Forestier et al. arrived to analogical conclusions concerning the expertise-hospital difference in their multi-site study of surgical practices made on the ACDF procedure [Forestier 2013]. During hierarchical clustering performed in that study, the outer clusters were represented by the surgical sites and the inner clusters by the expertise levels. However, in our study (Section 5.4.4.2) it was also shown that the difference between two sites depends on the procedure. For example, the LDH procedure is performed more similarly in Rennes and Leipzig than PA or ACDF.

Another discovery demonstrates that the procedures in Rennes resemble more to each other than those in Leipzig. The base line scores from Section 5.4.4.1 were better for procedures performed in Rennes. However, this might be caused by a greater complexity of Leipzig's surgical process in general. This complexity can be observed in the description of the data through a longer duration, greater length (number of activities per intervention) and bigger number of unique activities. Yet, strong evidences of the difference in question were found during the experiments on mixing and transfer (Sections 5.4.4.2 and 5.4.4.5). The intra-site differences between Rennes and Leipzig may be explained by several things. The procedures in Rennes could be more standardized than in Leipzig. This assumption would also suit the hypothesis from the last chapter about the more frequent use of surgical instruments not according to their initial functions in Leipzig. However, another explanation would correlate the complexity of the surgical process to the patients' conditions. With a relatively small dataset like this, it is statistically possible that surgeons in Leipzig more often operated patients with difficult cases (e.g., multi-level discectomies, later stage of the disease, abundant bleeding, etc.).

Unlike with sites and procedures, the experiments on expertise level were less conclusive. Forestier et al. performing classification of surgical workflows revealed that "seniors can have different operating techniques and preferably sequences of activities that differ from one senior to another" [Forestier 2012]. The results of our study suggested that juniors were different from seniors, but each group, especially juniors, also had a high variability inside the group. That could explain lower prediction results of junior datasets. This is coherent with [Riffaud 2010] stating that with practice, the number of gestures is reduced and coordination of hands is optimized which leads to a more standardized process in the senior group. However, lack of evidence may also be caused by a blurry frontier between

juniors and seniors. In the data used in this study, the seniority (i.e., level of expertise) was defined by the number of performed operations but not as the result of an objective evaluation of surgical skills. Knowing that each surgeon has its own operating style, it would be interesting to split the procedures by individuals. Yet, such an experiment would require considerably larger amounts of data.

The described conclusions on surgical practice were made observing the relationships between the input and output (i.e., prediction results) of the conducted experiments. Nevertheless, the nature of deep learning discussed in Section 5.5.2 does not allow to see the true dependences inside the network. Thus, these conclusions remain hypothesises to be proven.

5.6 Conclusion

In this chapter, we focused on the problem of data deficiency and proposed to use knowledge transfer methods in order to compensate a small amounts of training data. To demonstrate the power of knowledge transfer the task of next surgical activity prediction was chosen as an example. Two transfer methods were proposed in this work augmenting the prediction accuracy by almost 22% in total. The first method was the word embedding technique vastly applied in natural language processing. The word embedding was used to extract semantic knowledge describing the relationships between surgical terms from specially created medical word corpora. Its was shown that using even a relatively small corpus can significantly improve the results of the prediction. The word embeddings also helped to uniform the data from different datasets to enable transfer learning within deep neural networks. The second method transferred the knowledge from one dataset of annotated interventions to another. It used the internal parameters of the neural network to encode the information about the surgical process. Such transfer learning was also very effective in improving the prediction. Positive results were obtained when transferring information from one procedure to another, and between the surgical sites. Several observations about surgical practices were made along with the experiments conducted to evaluate the transfer methods. We observed that the difference between the expertise levels is smaller than the difference between the surgical sites which is smaller than the difference between the surgical procedures. The procedures conducted in Rennes seem also to be more alike than those performed in Leipzig. This work is the first study in the literature applying knowledge transfer on surgical processes.

Main findings

- Bigger word corpus and bigger embedding vector size lead to higher performance
- About 30% of knowledge is shared between the sites, although not all of it is useful for transfer
- The best results in transfer learning were achieved by transfer between procedures
- The transfer between the surgical sites was also effective
- The transfer between the expertise levels was not effective
- The knowledge about the expertise level is less important for activity prediction
- Two dissimilar sets of surgical workflows can still have some knowledge that enhances the learning process of each other
- Deep recurrent network learns and transfers not only a procedure-specific knowledge but the knowledge about the surgical process in general as well
- The workflows of junior surgeons are different from the seniors, and the heterogeneity inside the junior group is bigger than in the senior group
- The difference in workflow between the expertise levels is smaller than between the surgical sites which in turn is smaller than between the procedures
- The surgical workflow in Leipzig is more complex than in Rennes
- The neurosurgical procedures are more alike in Rennes than in Leipzig
- ACDF and LDH have more in common than LDH and PA which in turn have more in common than ACDF and PA
- The workflows from Leipzig and Rennes resemble the most for LDH, less for PA and much less for ACDF

CHAPTER

6 Application-dependent validation metrics

Preamble

In this chapter, we address validation-related issues of the surgical workflow domain. We propose new metrics and performance estimation approaches particularly adapted for assessing workflow recognition and prediction methods. We propose separate metrics for methods dealing with surgical phases and activities, and demonstrate their use on concrete examples. The suggested metrics help to measure important parameters of the performance and to value it as regards to the actual clinical application.

Contents

6.1 Introduction

The surgical workflow and its recognition have been in the center of attention of many research projects for at least fifteen years. A great amount of innovative approaches were presented during this period. Today, in an era of rapid technological progress, the need for such systems becomes vital. Many researchers working in this field participate in a never ending race for performance. Each year more and more methods are proposed overcoming previous state-of-the-art results. However, one important aspect remains often neglected - proper validation and advanced analysis of errors. In most publications on the topic, inner parameters of the system and their influence on the performance are well studied, though the actual validation is reduced to a measurement of a few standard performance scores. This attitude limits objective comparison of approaches since multiple aspects of their performance stay veiled from the reader. As shown in the literature review from Chapter 2, most of the time, only global scores as accuracy, precision and recall are exposed in the result section. First of all, this provides only a general and vague picture of the method's capacities and flaws. Secondly, these scores are highly disconnected from practical clinical applications. They are too strict and not informative enough to capture the readiness of the method for its use inside the OR and its utility for the surgical process. In order to be able to proceed with a system's integration inside the OR, its performance should satisfy the requirements of the concrete application for which it was designed. The standard performance scores, commonly used in the domain of machine-learning, are not enough to express these requirements. That is why, it becomes highly difficult to estimate how close the method actually is to its application. It explains why many published methods get stuck in the validation stage and do not proceed with real-case system evaluation. Among all the publications reviewed in Chapter 2 (except [Dergachyova 2016] featuring the contribution of this chapter), only [Forestier 2017, Lalys 2011, Malpani 2016, Primus 2016, Twinanda 2017] proposed original metrics destined to express their method's behaviour in a truly relevant way. None of these

publications, however, evaluated such important parameters as recognition delay (except [Primus 2016] which separately measured it for each surgical phase, but proposed no metric to estimate the overall delay) and consistency, or connected the method's results to the targeted application.

The challenges the surgical workflow recognition annually organized by community members enable comparison of the methods on a common data basis. Yet, the comparison criteria can not be fairly called objective, as only standard performance scores are usually used for the assessment. It is obviously hard to find a perfect criterion for all approaches, as they have been initially designed for different purposes. Despite that, the need for new more objective validation strategies and meaningful metrics is apparent. Our work is an attempt to give a deeper insight of the clinical needs and to emphasize the relevance of this topic. In this chapter, we propose a set of application-dependent metrics and approaches for performance estimation for two workflow granularity levels: phases and activities. We also demonstrate the use of these metrics on several existing recognition methods, discussing the obtained results and their utility.

6.2 For surgical phase recognition

In the first part of this chapter, we introduce four metrics destined for the assessment of methods performing surgical phase recognition. First, in Section 6.2.1 we pose the problem by formally describing concepts that are used for definition of the metrics proposed in Section 6.2.2. The example of metrics application and the obtained results are presented in Section 6.2.3. The recognition methods used here as example are the methods submitted to the MICCAI M2CAI 2016 challenge on workflow recognition from endoscopic videos [1] organized by the research group CAMMA from the University of Strasbourg (France) and Technical University of Munich (Germany). The dataset provided for the challenge [Twinanda 2017, Stauder 2016] consisted of 27 cholecystectomy videos available for training, containing 8 surgical phases, and 14 test videos. Test videos contained both complete and incomplete surgeries. The submitted methods required to perform an on-line recognition (based on the information from the past but not from the future). Dice coefficient was used to assess the methods and determine the final rating. The organizers of the challenge kindly provided us the access to the ground truth annotations and the automatically recognized sequences submitted by the participants, which were used here for computation of the proposed metrics.

1. http://camma.u-strasbg.fr/m2cai2016/index.php

6.2.1 Problem formalization

1. Let $P = \{p_1, p_2, ..., p_m\}$ be the set of m possible phases.
2. Let $Seq^{GT} = (a_1^{GT}, a_2^{GT}, ...a_n^{GT})$ and $Seq^R = (a_1^R, a_2^R, ...a_n^R)$ be the ground truth (i.e., observer's manual annotation of the data) and recognized sequences respectively, where each item in the sequence $a_i(t \in \mathbb{N}, p \in P)$ represents a moment in time, uniquely defined by a timestamp t and having a label of phase p, and n is the number of analysed time moments in a given intervention.
3. Let $\tau = \{tr_1, tr_2, ..., tr_k\}$ be the set of k possible transitions between phases, where tr_i is defined by a $p_{prev} \rightarrow p_{next}$ transition; $p_{prev} \in P$ is the current phase before the transition, and $p_{next} \in P$ after.
4. Let us call *transitional moment* (TM) a moment in time when a transition between two phases takes place.
5. Let $TM^{GT} = (tm_1^{GT}, tm_2^{GT}, ..., tm_s^{GT})$ and $TM^R = (tm_1^R, tm_2^R, ..., tm_{s'}^R)$ be the sequences of transitional moments for the ground truth and recognition respectively, where $tm_i(t \in \mathbb{N}, tr \in \tau)$ is composed of a timestamp t indicating the moment in time when the transition takes place, and a transition tr respecting the condition $tm_i^{GT}(tr) \neq tm_{i+1}^R(tr)$. The values of s representing the number of transitions in TM^{GT} and s' in TM^R may differ from each other.
6. Let $TM^* = \{\mu_1, \mu_2, ..., \mu_l\}$ be the set of l pairs $\mu_i = < tm_x^{GT}, tm_y^R >$ indicating the closest matching transition moments, where $tm_x^{GT}(tr) = tm_y^R(tr)$, and $abs(tm_x^{GT}(t) - tm_y^R(t))$ is less than for all other candidate pairs having the same tr. The values x and y may or may not be equal but always unique within TM^*.
7. Let us call *transitional interval* (TI) a period in time between two matching transitional moments.
8. Let $TI^* = \{\rho_1, \rho_2, ..., \rho_l\}$ be a set of transitional intervals constructed from TM^*, where a pair $\rho_i = < t_{start}, t_{end} >$ indicates the start and end time of the i-th transitional interval.

6.2.2 Definition of metrics

6.2.2.1 Average transitional delay

Automatic recognition of current surgical phase enables providing the right computer assistance at the right time. For a continuous process as surgery, it is also important to accurately detect transitional moments between two phases. This allows to switch between different types of assistance at the right time. Generally, sequence-based recognition methods, that keep track of the procedure, do a better job of detecting TMs than frame-wise approaches. However, whatever the approach, the detected TMs may differ from the real ones determined by the ground truth annotation, resulting in a transitional

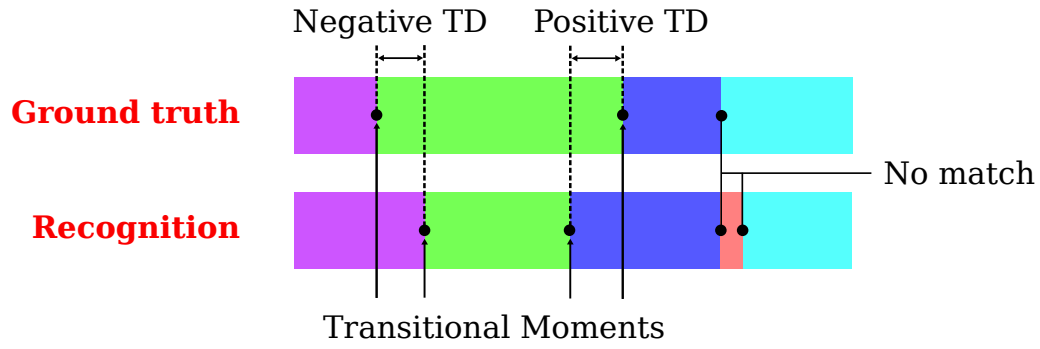

Figure 6.1: Examples of negative and positive transitional delays

interval formally defined above. We call a *transitional delay* (TD) the distance it time between the real and the detected TMs. This delay may be negative and positive (Figure 6.1). A negative delay indicates that the transition between the phases is detected with a lateness with regards to the ground truth (i.e., the detected TM has a bigger timestamp). A positive delay, conversely, means that the system decides to switch the phase too early, before the actual transition (i.e., the detected TM has a smaller timestamp). From the applicative point of view, knowing transitional delays is important to have an estimation that helps to decide whether the method is suitable for the targeted application.

We define the *average transitional delay* (ATD) metric to measure the delays produced during all transitions in all available interventions in order to have an average estimate of the delay. Negatives and positives delays are measured separately and used to define an interval of values for the average transitional delay. First, the sign of each delay (positive or negative) is determined with the function $sign(i)$ as in equation 6.1. Then, the negative ATD is computed as in equation 6.2 and positive as in 6.3. The final ATD interval is defined as in 6.4.

$$sign(i) = \begin{cases} 1 & \text{if } \rho_i(t_{start}) = \mu_i(tm^R(t)) \text{ and } \rho_i(t_{end}) = \mu_i(tm^{GT}(t)) \\ -1 & \text{otherwise} \end{cases} \tag{6.1}$$

$$ATD^- = \frac{\sum_{i=1}^{l} p_i(t_{start}) - p_i(t_{end}) \mid sign(i) = -1}{\sum_{i=1}^{l} sign(i) = -1} \tag{6.2}$$

$$ATD^+ = \frac{\sum_{i=1}^{l} p_i(t_{end}) - p_i(t_{start}) \mid sign(i) = 1}{\sum_{i=1}^{l} sign(i) = 1} \tag{6.3}$$

$$ATD = [ATD^-; ATD^+] \tag{6.4}$$

It may happen that some ground truth TMs are completely missed in the recognized sequence, or that a certain existing transition is detected several times in the recognition sequence. The missed or false positive TMs are mostly due to an actual failure of recognition and do not relate to delay. The ATD metric, on the other hand, measures the reaction time of the system only. The problem of recognition failure is addressed using the metric from Section 6.2.2.3. Meanwhile, the above formulas consider only transitional moments that can be uniquely matched. Following the same logic, the metric supposes that the correct detected TM is the closest to the ground truth TM. However, one can measure the delay between the first occurrences of the TM having the same transition tr.

6.2.2.2 Detected-to-real transition ratio

Stability and consistency of recognition is another important factor of a successful surgical computer assistance. Let us take an example of context-aware system automatically controlling devices and adjusting the brightness of the lights in the OR. A frequent incorrect detection of phase transition may highly disturb the surgeon and even cause complications. Now, imagine two recognitions as shown in Figure 6.2 made by two different systems giving the same accuracy. The first recognition has some transitional delays but a completely correct order of phases, being an example of a stable system. The second, contrariwise, has almost no delay, but a great amount of short lasting false positive transitions. This demonstrates the importance of measuring the number of transitions in the recognized sequence since the standard scores as accuracy do not represent this information. We propose to compute a detected-to-real transition ratio (TRR) between the number of detected transitional moments and the number of real ones (equation 6.5). It serves as indicator of system stability and reflects its robustness, as a system with a high TRR is probably less tolerant to intrinsic changes in input data. This ratio also gives a simple intuitive idea of how many of incorrect TM were detected as regards to their actual number.

$$\mathrm{TRR} = \frac{s'}{s} \tag{6.5}$$

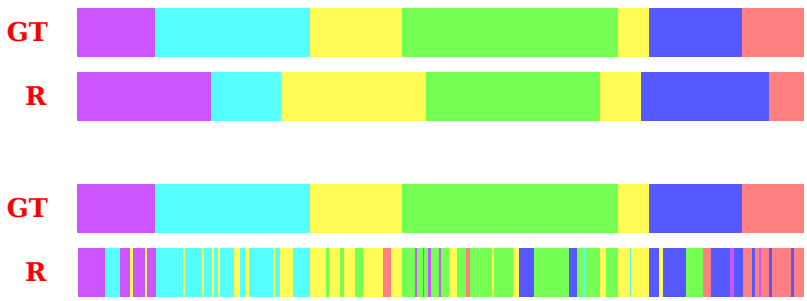

Figure 6.2: Examples of two systems providing different phase recognitions but having the same accuracy

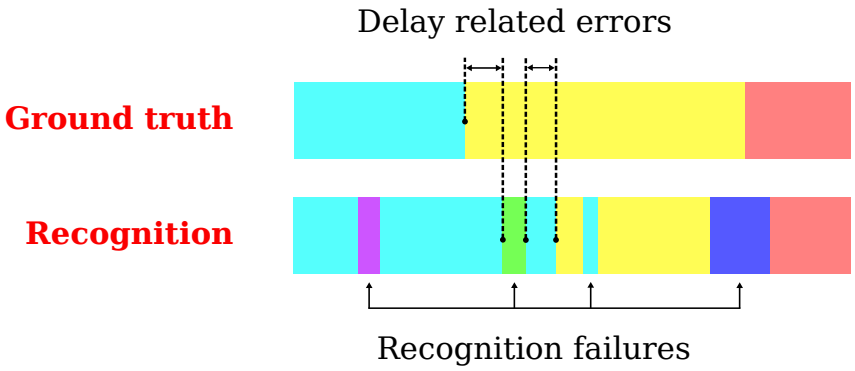

Figure 6.3: Examples of two types of error

6.2.2.3 Actual failure rate

The errors in the recognized sequence can be divided into two categories: delay related and actual errors cause by recognition failure (Figure 6.3). The time delay problem is, generally, very hard to solve but less harmful. A time delay means a belated but a correct recognition. Whereas, recognition failure, oppositely, indicates that the system fails to interpret and understand the on-going situation. That is why, it is important to know the actual amount of error (i.e., moments of recognition failure) that has to be worked through in the first place. We propose to measure this with a failure rate computed as a ratio of the number of mislabelled time moments not caused by any transitional delay to the total length of the sequence (see equation 6.6). The algorithm 6.1 explains the computation of the $error()$ function used in the equation. In this algorithm, the label of a time moment inside a transitional interval is still checked to ensure that it actually makes part of the delay.

$$AFR = \frac{\sum_{i=1}^{n} error(i)}{n} \tag{6.6}$$

6.2.2.4 Application-dependent scores

Standard performance scores are too strict and meaningless for system assessment within the applicative context. They do not allow to objectively estimate the readiness and appropriateness of the system to a specific use. Some applications do not require a frame-by-frame phase identification. They may tolerate a certain time delay in detection with no essential impact on the provided assistance. We propose to re-estimate standard scores for concrete clinical applications. For this, we introduce a notion of *transitional window* - an interval of time centred on a real transitional moment and, at both ends, authorizing a delay d acceptable for the chosen application (Figure 6.4). We also redefine the

notion of "true positive" TP replacing it by "application-dependent true positive" TP'. In this case, if the examined time moment is inside the transitional window and happens due to a delay, it is counted as true positive : $a_i \in TP'$ if $\exists tm : abs(tm^{GT}(t) - a_i(t)) \leqslant d$ and $error(i) = \text{false}$. Any standard score including accuracy, recall and precision can be re-estimated using the notion of application-dependent true positive. This approach helps to see the actual progress that has to be done before integrating the system in the OR.

Algorithm 6.1 Computation of $error(i)$ function determining if the i-th item in the sequence represents an actual recognition failure

```
function ERROR(i)
    e ← true
    if a_i^GT(p) = a_i^R(p) then e ← false
    else
        for 1 ⩽ j ⩽ l do
            if ρ_j(t_start) ⩽ a_i^R(t) ⩽ ρ_j(t_end) then
                if sign(j) = 1 and a_i^R(p) = μ_j(tm^GT(tr(p_next))) then e ← false
                else if sign(j) = −1 and a_i^R(p) = μ_j(tm^GT(tr(p_prev))) then e ← false
                end if
            end if
        end for
    end if
    return e
end function
```

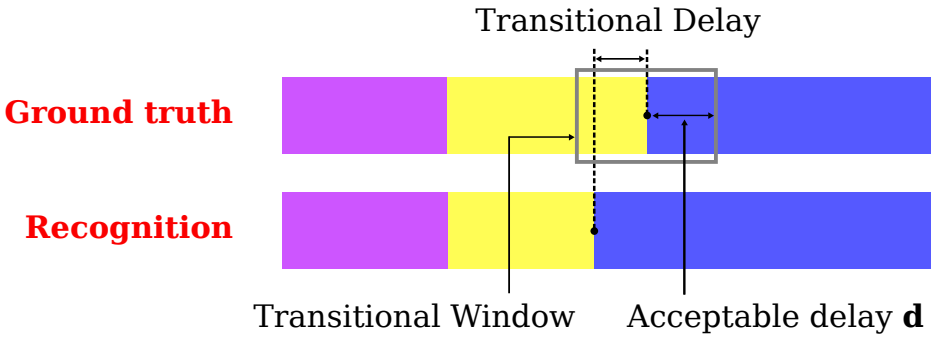

Figure 6.4: Transitional window

6.2.3 Metrics application

We applied the proposed metrics to three different phase recognition approaches submitted to the M2CAI 2016 challenge. These approaches are briefly described below; their detailed information can be found in the technical reports referenced for every method. Section 6.2.3.2 presents the results obtained with the metrics, discusses the differences between the methods and appropriate applications.

6.2.3.1 Recognition methods

1. Cadene et al. The approach proposed in [Cadene 2016] is a deep learning based approach. It used a convolutional neural network with time smoothing and a classical hidden Markov model to perform the phase recognition task. The network presented for the challenge was based on a Residual Network-200 pre-trained on ImageNet, last layer of which was replaced by a new fully connected output layer corresponding to 8 possible surgical phases. It was then fine-tuned on the M2CAI dataset using on-line data augmentation. A temporal smoothing was performed on the log-probability output vectors of the network. The smoothed vectors were then passed to the HMM to correct possible classification errors as regards to the previously recognized frames.

2. Twinanda et al. The authors of [Twinanda 2016] also proposed a method based on deep learning. They took the pre-trained AlexNet as a base, replacing the output layer, and fine-tuned it with the M2CAI training dataset. The fine-tuned network was called PhaseNet. The image features extracted by the CNN found in the second last layer of the PhaseNet were used as input for a one-vs-all linear SVM. A hierarchical HMM was also put on top of the SVM classifier to reinforce the temporal constraint.

3. Dergachyova et al. The method proposed in [Dergachyova 2016] represented a classical machine-leaning approach consisting of several steps. The first step consisted in describing input images by extracting visual features characterizing color, form and texture of the image. At the next step, an intermediate classification was done using several AdaBoost cascades. Finally, a hidden semi-Markov Model was used to provide a definitive phase label.

6.2.3.2 Results

The results of ATD, TRR, and AFR metrics are presented in Table 6.1. For ATD and AFR, their part in the total error was also computed. The application-dependent (AD) scores were re-estimated for two possible applications. The first was information display and device triggering that required a relatively fast reaction. For this application the acceptable threshold d for the delay was set to 15 seconds, meaning that all incorrect labellings caused

by a transitional delay of 15 seconds or less were counted as good recognitions. The second application was estimation of remaining surgical time. The d value was fixed to 1 minute with no essential impact on the estimation of time left. The results of standard and AD scores (accuracy, recall and precision) for both cases can be found in Table 6.2.

Table 6.1: ATD, TRR and AFR metrics for phase recognition

	ATD			TRR	AFR	
	interval	of total time	of total error		of total time	of total error
1. Cadene et al.	[-30s ; 1min 8s]	[1.6% ; 3.5%]	44.2%	1.8	7.3%	55.8%
2. Twinanda et al.	[-23s ; 54s]	[1.2% ; 2.8%]	34.1%	8.6	12.6%	65.9%
3. Dergachyova et al.	[-45s ; 1min 10s]	[2.3% ; 3.6%]	25.2%	2.7	22.7%	74.8%

Table 6.2: Standard and AD scores (%) for phase recognition

		Standard scores			AD scores (d = 15s)			AD scores (d = 1min)		
		Acc.	Rec.	Pres.	Acc.	Rec.	Pres.	Acc.	Rec.	Pres.
Avg.	1. Cadene et al.	84.2	69.2	74.2	88.6	75.8	80.9	92.4	80.9	86.5
	2. Twinanda et al.	79.2	67.6	74.0	81.5	72.6	78.8	85.9	79.9	84.5
	3. Dergachyova et al.	68.6	61.9	62.1	73.1	67.3	66.2	78.6	73.4	71.1
Std.	1. Cadene et al.	10.9	19.0	13.3	11.2	17.1	12.7	11.2	17.2	12.2
	2. Twinanda et al.	11.6	19.4	11.6	11.5	18.2	10.6	12.8	16.6	10.8
	3. Dergachyova et al.	5.8	16.8	20.8	6.5	16.2	19.0	7.3	15.1	17.1

From the results presented in Table 6.1, it is clear that the method #2 has the shortest transitional delays. That is why, its accuracy was improved less than others using AD metrics. This method is more suitable for applications requiring a fast system reaction. However, it makes too many incorrect transitions between phases (8.6 times more than should be). The method #1, on the other hand, provides recognitions with larger delays, but much less spikes of incorrect phase transitions (TRR = 1.8). However, the periods of incorrect recognitions last longer. Its recognitions are much more consistent compared to the method #2. Thus, the method #1 suits more the applications requiring high stability. The method #3 has also a low TRR. However, most errors come from the actual failures of recognition (74.8%), which last for longer periods of time. Short spikes of recognition failure, like high-frequency noise, are simpler to filter in real-time applications compared to longer periods of recognition failure. However, longer periods with less incorrect transitions have more chances to be successfully corrected in off-line applications, taking benefit from a complete sequence. This shows that the methods #2 is better for on-line use and

#1 and #3 for off-line. The results of Table 6.2 show how the performance scores can grow using appropriate AD transitional windows. This helps more clearly estimate how close the methods actually are to the clinical use within a concrete application.

It is also important to mention, that these three works can not be discriminated or, in this case, objectively compared with the proposed metrics. The methods were designed and trained to maximized accuracy but not the proposed scores. It is possible, that other sets of methods' parameters would result in better ATD, TRR, AFR and AD scores. Yet, using such metrics for the challenge assessment could change the ranking.

6.3 For surgical activity recognition and prediction

The second part of this chapter is devoted to application-dependent approaches of performance re-estimation for methods providing recognition and prediction of surgical activities. Section 6.3.1 provides a formal explanation of a usual performance assessment. Section 6.3.2 introduces new approaches of error estimation, and Section 6.3.3 demonstrates their use. We tested the proposed error estimation approaches on the method of next activity prediction from the previous chapter. The best transfer configuration "mix of sites + inter-procedure transfer" (transfers ACDF $\rightarrow$ LDH, LDH $\rightarrow$ ACDF, ACDF $\rightarrow$ PA) was used, and the scores averaged for all three procedures were exposed. The task of prediction was taken because a corresponding method was already available within the current work. However, the formalisation and metrics can be applied to the activity recognition task as well.

6.3.1 Problem formalization

Let $Seq^{GT} = (a_1^{GT}, a_2^{GT}, ...a_n^{GT})$ and $Seq^{P} = (a_1^{P}, a_2^{P}, ...a_n^{P})$ be the ground truth and predicted sequences respectively, where each activity $a_i < L(V,I,S), R(V,I,S) >$ is composed of an action verb V, surgical instrument I and anatomical structure S for both left L and right R hands. The activity is usually considered correctly predicted only if all 6 elements are correctly predicted. For instance, for one sequence prediction accuracy is computed as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n-1} f(i)}{n-1}, \tag{6.7}$$

where the function $f(i)$ evaluates correctness of the prediction for one activity as :

$$f(i) = \begin{cases} 1 & \text{if } (V_{L_i}^P = V_{L_i}^{GT}) \text{ and } (I_{L_i}^P = I_{L_i}^{GT}) \text{ and } (S_{L_i}^P = S_{L_i}^{GT}) \text{ and} \\ & (V_{R_i}^P = V_{R_i}^{GT}) \text{ and } (I_{R_i}^P = I_{R_i}^{GT}) \text{ and } (S_{R_i}^P = S_{R_i}^{GT}) \\ 0 & \text{otherwise} \end{cases} \quad (6.8)$$

6.3.2 Definition of metrics

6.3.2.1 Rational accuracy

The usual definition of the $f()$ function constrains the output to a binary value without counting the number of correctly predicted elements. Imagine two predicted sequences as in Figure 6.5. In the first sequence, all incorrect predictions have errors in five or six activity elements at once, while in all incorrectly predicted activities of the second sequence only one or two elements are wrongly labelled. Accordingly to the accuracy computed with $f()$ from equation 6.7, both systems that produce these predicted sequences have the same score but, in fact, quite different performances. For a more precise estimation of the system performance, it is important to know the number of correctly predicted elements. The present metric redefines the $f()$ function so that it returns a rational value for each predicted activity, based on the number of correctly predicted elements as in equation 6.9. However, one should keep in mind, that a system with a high rational accuracy of 83.3% (when in all activities only 5 from 6 elements are predicted) may still provide no truly correct prediction.

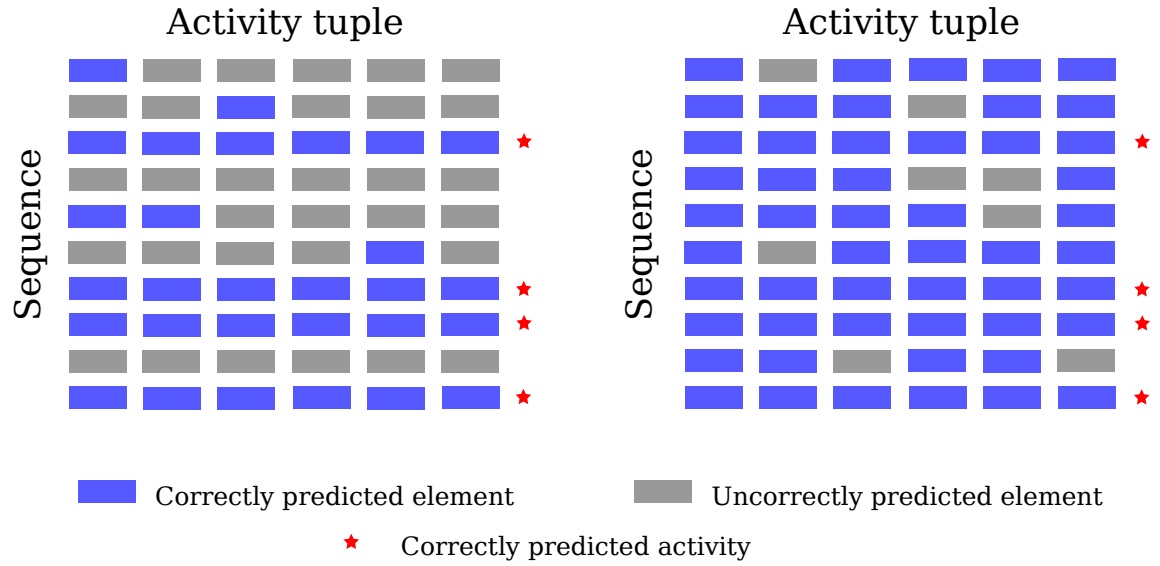

Figure 6.5: Examples of two systems providing different predictions of activity elements but having the same accuracy

$$\begin{aligned} f(i) := \tfrac{1}{6}(V^P_{L_i} = V^{GT}_{L_i}) + \tfrac{1}{6}(I^P_{L_i} = I^{GT}_{L_i}) + \tfrac{1}{6}(S^P_{L_i} = S^{GT}_{L_i}) + \\ \tfrac{1}{6}(V^P_{R_i} = V^{GT}_{R_i}) + \tfrac{1}{6}(I^P_{R_i} = I^{GT}_{R_i}) + \tfrac{1}{6}(S^P_{R_i} = S^{GT}_{R_i}), \end{aligned} \tag{6.9}$$

where the sign := means assignment, and the sign = is a boolean operator testing if the instance on the left is the same as on the right, returning 1 if yes, and 0 if not.

6.3.2.2 Inverted hands

It may happen that the labels given for the elements of a two-hand activity are actually inverted (i.e., right elements get predicted as left and vice-versa). This may occur if the surgeon changes hands on purpose, which creates a rare activity in the dataset, or if a computer-vision recognizer is confused by an unusual position of hands. Regarding the $f()$ function from equation 6.7, this condition will be treated as incorrect prediction. Yet, for some context-aware applications, such inversion would not hurt the overall understanding of the situation. For these cases, we redefine the function $f()$ as $f'()$ allowing the inversion.

$$f'(i) = \begin{cases} 1 \quad \text{if} \quad f(i) \text{ or } ((V^P_{L_i} = V^{GT}_{R_i}) \text{ and } (I^P_{L_i} = I^{GT}_{R_i}) \text{ and } (S^P_{L_i} = S^{GT}_{R_i}) \text{ and} \\ \qquad\qquad (V^P_{R_i} = V^{GT}_{L_i}) \text{ and } (I^P_{R_i} = I^{GT}_{L_i}) \text{ and } (S^P_{R_i} = S^{GT}_{L_i})) \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{6.10}$$

6.3.2.3 Unordered elements

There are applications, such as context video or database retrieval, where the search can be executed by a set of unordered key words. For these applications, the order of elements in the activity has poor importance while all correct element instances can be found in the prediction. Thus, the $f()$ function can be redefined as in equation 6.11.

$$f(i) = \begin{cases} 1 & \text{if} \quad |\{a^P_i\} \cap \{a^{GT}_i\}| = 6 \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

6.3.2.4 Significant element

For some applications, certain activity elements may have higher significance than others. As example, the automatic prediction of next surgical activities can be used to assists novice scrub nurses in anticipation process. Displaying next activity on the screen would help them to stay tuned and to prepare necessary instruments in advance. In this case, prediction of a whole activity is desirable yet only an accurate prediction of the instruments is vital. Equation 6.12 shows the definition of $f()$ function adapted to the prediction of instruments. It can be analogically defined for two other elements as well.

$$f(i) = \begin{cases} 1 & \text{if } (I^P_{L_i} = I^{GT}_{L_i} \text{ and } I^P_{R_i} = I^{GT}_{R_i}) \\ 0 & \text{otherwise} \end{cases} \tag{6.12}$$

6.3.3 Metrics application

The results of the proposed metrics applied to the task of next activity prediction can be found in Table 6.3. The rational accuracy was higher than the standard accuracy by almost 5%. On average, 3 to 4 elements of incorrectly predicted activities were still correctly discovered. Only 2.3% of activities got inverted in the prediction. It mostly happened when only one hand was involved in the activity. When the actions of the left hand are performed with the right hand (e.g. blood suction), it may indicate that a particular event (e.g. bleeding) is taking place. The authors of [Huaulmé 2017b] argue that deviations from the standard procedure are manifested by certain precursor patterns and are possible to predict. Yet, our system would apparently fail to correctly predict such precursor activities. On the other hand, the system predicted the elements in an appropriate order - its output was constrained to that. However, in some rare cases (0.8%) the prediction of one element of the left hand got inverted with the same element of the right hand. The verb was predicted better than other elements. It probably comes from the fact that the procedures involved in transfer shared more verbs than instruments or structures. That is why this information was learnt and transferred easier. With the new error estimation metrics, the system got 6% closer to the application of situation anticipation for scrub nurses.

Table 6.3: Results of activity prediction with application-dependent accuracies

Standard accuracy	89.1% ± 5.1%	Significant element	
Rational accuracy	94.7% ± 3.8%	Verb	96.3% ± 3.0%
Inverted hands	91.4% ± 4.9%	Instrument	95.3% ± 4.1%
Unordered elements	92.2% ± 4.6%	Structure	93.2% ± 6.3%

6.4 Discussion

6.4.1 Other metrics and strategies of error analysis

This chapter demonstrated the importance of meaningful validation. The proposed metrics and error-estimation techniques are the examples of how a recognition method can be bound to a targeted clinical application in order to approach its integration in the OR. Other methods of error analysis can nevertheless be applied. As an example, systems providing assistance during particular phases require high recognition recall for these

phases only. In addition, the acceptable threshold for the transitional window may vary for different phases. Other relevant information would be the order of phases in the recognized sequence. Levenshtein distance used in [Malpani 2016] and frequently applied in sequence alignment algorithms as DTW measures how much two sequences resemble. Another important requirement for medical systems that has be addressed in the validation process is robustness. "The robustness of a system refers to its performance in the presence of disruptive factors such as intrinsic data variability, data artefacts, pathology, or inter-individual anatomic or physiologic variability"[Jannin08]. Simulating noise in input data and measuring its impact on the recognition, as done in Chapter 4, is one of the options to estimate the robustness. Unfortunately, among all the reviewed methods only [Bouarfa 2011, Forestier 2015, Franke 2015a] actually did that. Finally, an intuitive way of understanding the errors is visualization of recognized sequences against the ground truth. This enables qualitative analysis. Such visualizations were displayed in [Dergachyova 2016, Forestier 2015, Forestier 2017, Franke 2015a, Lalys 2011, Lalys 2012, Tran 2017, Twinanda 2017, Zia 2017].

Similar strategies are conceivable for surgical activities. Particular attention should also be paid to the sequencing of activities. Although being highly appropriate, sequence distances as Levenshtein's leave a room for consideration. Such distances would treat all activities equally distant from each other. Nevertheless, the activity tuples are formalized representations of semantic knowledge about the surgical process. Some sort of semantic distance between the activities would be highly pertinent in this situation. An effort in [Huaulmé 2017b] was done to adjust the Levenshtein distance so that it takes the elements of the activity into account in its computation as well. However, the proposed distance still made no difference between the activity elements. Word embeddings, encoding semantic distance between activity words in form of vectors, would be a good option. Though another questions arise. How one should interpret the distance? Is a threshold separating the activities on similar or dissimilar has to be fixed, and how? How to validate a threshold if there is one? Making an actual mistake or interpreting the situation in different terms is not the same thing. Semantic distance has the potential to make this distinction yet the arose questions have to be answered first. In addition, the distance can also be computed as a factor of activity significance, which also leaves us with the following question. How to certainly define if an activity is important regarding the entire process? All this provides a lot of matter for discussion.

Finally, since the effectiveness of learning and performance of recognition methods also depend on the quality of input data annotation, its correctness should be examined and the amount of uncertainties and errors in annotation has to be taken into account when estimating method's performance as well.

6.4.2 Validation standards

Introducing the problematic of this chapter, we implicitly divided existing validation metrics into standard machine-learning scores and performance estimation approaches relevant to surgical workflow. In our understanding, relevant metrics should better describe the behaviour of the system (e.g., nature and sources of errors, recognition progression in time) and express its performance as regards to the requirements of targeted clinical applications. The concept of relevance, however, has to be explicitly defined. A standard referencing all crucial parameters to measure for surgical workflow recognition system has to be established. Another important question is "When can we say that the method is ready to be used?". Our application-dependent metrics, loosening up the recognition constraints, assume that the method is ready when its reaches perfect accuracy (i.e., the score equals to 100%) for a chosen acceptable delay or output permutation. Equivalent performance targets have to be defined for all relevant metrics.

During validation process, the recognized and predicted sequences are compared against manual workflow annotations, called ground truth. One should remember that this ground truth has several bias. The first bias originates from the definition of the surgical phase or activity itself. For instance, in reality, the transition between phases rather represents a period of their overlap than a sharp frontier between the two. Even the experts do not completely agree about the moment on the time line when the transition should be marked. The second bias comes from the person annotating the data. First, the transitional point is marked according to observer's perception. Secondly, particularly for short activities, a lag may occur between the time of the real transition and the time marked in the annotation. That is why, there is less sense in trying perfectly match ground truth transitional moments than in providing a sustainable recognition in between. Besides adequate metrics assessing the performance, a great attention should also be paid to a deeper analysis of errors and discovery of their sources. The main questions would be: When and why do the errors happen? Of course, the methodology of error exploration greatly depends on the recognition algorithm itself. However, general heuristics have to be proposed to facilitate the process and motivate the community.

Despite the progress in the field of surgical workflow, the appropriate concept of validation has not been fully developed. A careful and rigorous process of validation is often neglected. In the domain of medical image processing, a considerable effort in standardization of the validation process was done by [Yoo 2000, Jannin 2006b, Jannin 2008]. The assessment of the actual advantage in clinical use occupied one of the central places in these works. Similar ideas can be adapted for surgical workflow since, for now, no substantial common standard exist. There is also a need for standardization of concepts and terminology. Greater attention should be paid to ontologies structuring data. Their widespread use would establish common basis needed for objective comparison and validation.

6.5 Conclusion

The performance metrics usually used for validation of approaches recognizing or predicting surgical workflow are often too strict, not informative enough and disconnected from clinical applications. In this chapter, we proposed new adapted metrics that help to extract more of relevant information from the method's results and to re-estimate the standard performance scores as regards to a particular targeted application. Metrics such as average transitional delay, detected-to-real transitions ratio, actual failure rate and application-dependent scores were proposed for phase recognition. They were tested on the methods submitted to the MICCAI M2CAI 2016 challenge on workflow recognition for endoscopic surgeries. For activities, we proposed the following new error estimation approaches: rational accuracy, inverted hands, unordered and significant elements. These approaches were tested on the method of next activity prediction from the last chapter. The results revealed some interesting properties and demonstrated how the methods can be bound with the clinical application.

Acknowledgements

We would like to thank the organizers of the MICCAI M2CAI 2016 challenge on workflow recognition for kindly providing the data and resulting annotations. The dataset for the challenge was collected thanks to [Stauder 2016, Twinanda 2017]. We also thank Andru Putra Twinanda and Rémi Cadene, the first authors of the used recognition methods, for granting the access to their results.

Part III

Conclusion and perspectives

CHAPTER

7 Conclusion and perspectives

Preamble

This last chapter draws a line under the work exposed in this dissertation. Section 7.1 recalls important issues of the domain of surgical workflow discovered during the review of scientific literature and addressed in this thesis. Section 7.2 briefly reminds the main stages of the conducted research and summarises our contributions. Section 7.3 exposes the limits of this work and suggests several improvements for its enhancement. Section 7.4 presents perspectives and proposes ideas that could take the research on the subject to the next level.

Contents

7.1 Motivation and objectives

Through couple of decades, incredible advances have been made in making computers understand the situation inside the OR automatically recognize surgical workflow. Numerous applications of situation and workflow awareness have been proposed . Nevertheless, up to this moment, a considerable gap between automatic recognition systems and their application can be observed through the content of publications in the field. A certain apartness of computer scientists and engineers from clinicians can also be felt. Another tangible blank is a lack of exploration and analysis before to the recognition stage. In our literature review, three considerable obstacles to evolution of surgical workflow recognition were emphasized.

1. The publications proposing automatic recognition methods for semantic surgical activities are much more rare than those for surgical phases or gestures. The deficit is explained by an extreme difficulty of the task, especially for complex surgeries. Semantic information about activity elements as action verb, surgical instrument and operated anatomical structure plays an important role in the recognition. Yet, no study analysing their impact on the recognition process exists. We decided to perform such an analysis to find input signals and sensors that could facilitate the recognition and provide better results.
2. Lack of clinical data in the domain is a well-known problem. This creates obstacles for recognition and analysis of surgical workflow - tasks requiring extensive amounts of data. We chose to explore techniques of knowledge transfer to enrich available data and compensate for its deficiency.
3. Insufficient validation, poor analysis of errors and use of inappropriate metrics to assess the performance of recognition methods notably slow down integration of systems in the OR and hinders objective estimation of progress. We decided to demonstrate why the current validation methods were not effective and to give a sense of how the validation process should be approached.

The prevalent concentration of researchers on purely recognition challenges often obstructs global vision of the situation awareness in the OR. That is why this work was carried out to address the problems “around” the recognition in order to overcome barriers blocking its progress and to reconnect it with the operating room. For this purpose, we asked

ourselves several important questions. How to make the recognition of surgical activities easier and more effective? How to deal with existing quantity constraints of surgical data? How to assess recognition methods in a meaningful and objective way? Besides finding answers to these questions, we also wanted to propose a new perspective to the recognition problem and validation process.

7.2 Summary of contributions

7.2.1 Analysis of activity elements

We conducted a study examining the importance of each semantic element (e.g. verb, instrument and structure) regarding activity recognition to facilitate the recognition process. The work was based on a common approach of one-by-one recognition of activity elements and their posterior concatenation. We had the following hypothesis. Not all of the input elements have to be recognized in order to make an accurate recognition of the entire activity; it is sufficient to find and use the most informative elements only. This would help to minimize the number of OR sensors and to guide their choice. We proposed an original approach for analysis to prove our hypothesis. We first assumed that all necessary elements can be confidently recognized. Then, we modelled the surgical workflow using Long-Short Term Memory recurrent neural network. We performed multiple experiments to estimate the importance of semantic elements. In the experiments, different activity elements and their combinations were evaluated one after another by masking their information while measuring the recognition performance of the network. Additional experiments were also conducted to see how noise and temporal delay in certain elements impact the recognition.

During the experiments, several observations and conclusions were made. We, indeed, proved our initial hypothesis: two elements of three were enough to accurately recognize an activity. The best choice was the combination of the anatomical structure and instrument. Thus, the sensors recognizing these two elements would be the most effective. The experiments with noise also confirmed advantage of this combination on others which were shown to be insufficient for a correct recognition. In addition, some interesting observations about surgical practice were made as, for example, the difference between the hospitals of Leipzig and Rennes in terms of instruments use. We also observed that despite the coordination of the hands, one hand (neither right nor left) is not enough for an accurate recognition.

7.2.2 Knowledge transfer

In this work, the problem of lack of data for training was transformed into the problem of lack of knowledge. We decided to address this problem by proposing methods of knowledge transfer that aimed to improve the recognition performance. We took a task of next

activity prediction as an example for demonstrating different approaches of knowledge transfer within a deep learning architecture. The deep architectures were chosen because of their potential to improve themselves along with the growth of data amount and easiness of transfer. We used two complementary types of transfer. The first consisted in using word embedding, a very popular technique in natural language processing [Collobert 2011], to extract and encode the knowledge about the surgical domain in general from scientific and medical texts. We created several word corpora, using freely available on-line services and research engines, on which the embeddings were trained with two different algorithms. We then integrated the embeddings into the deep neural network and estimated their impact on the activity prediction. The second type was the transfer learning technique passing the knowledge about the surgical process between the surgeries of the same speciality. The knowledge from one surgery encoded in internal parameters of the neural network was extracted and used as a training base for other surgeries. We tested different combinations and splits of data, as well as various transfer configurations to find what worked the best.

Thanks to both methods of transfer, the prediction accuracy of the network increased by almost 22% in total (11.7% for word embeddings and 10.2% for transfer learning). We demonstrated that even relatively small medical texts were good enough to extract helpful information about surgical terms without enlarging the actual dataset. The best transfer learning was made transmitting the knowledge between different procedures. That demonstrated that the extracted process knowledge can be independent of the procedure itself and encode some general information common for all procedures. This gives a hope that surgeries from other specialities are also able to contribute in the training process and help to to overcome the problem of data deficiency. We also made multiple observations about surgical practice. The most important concerned the difference between the expertise levels, surgical sites and procedures in terms of workflow.

7.2.3 Validation metrics and approaches

By addressing the problem of insufficient validation, we made a first step towards improvement explaining why standard machine-learning metrics miss important aspects of performance and why they are not appropriate for surgical workflow. We also showed their disconnection from real clinical applications. Then, we suggested more adapted informative metrics for both phase and activity recognition, as well as new application-dependent approaches for error estimation. We demonstrated their use on concrete phase recognition methods from the MICCAI M2CAI 2016 challenge and on our method of activity prediction. Withal, we discussed interesting properties of each method that have come into view thanks to the proposed metrics. We finally shared our vision of advancement directions.

7.3 Limitations of the work and ways of improvement

7.3.1 Data

One of the major limitations of the current work is absence of any raw data. The workflows of all neurosurgical datasets were annotated on-line without recording any physical signal. That is why, only semantic descriptions of activities were used as input for deep neural networks. Yet, activities with the same semantic discription may be different from a quantitative point of view or actually have different meaning depending on the global context meaning that the chosen descriptions possibly miss some important activity indicators. Adding numerical values of physical signals to the descriptions (e.g., videos, sound, motion data) would provide more of contextual information for learning.

7.3.2 Experiments

Multiple experiments designed to study diverse relationships between data components were conducted in this work. However, some types of experiments were not covered due to small sizes of the datasets. The experiments could have include studying surgeon-specific behaviours, analysing procedures from a multi-actor perspective, splitting the data by level of surgical complexity (e.g., patient's age, medical antecedents, disease stage or patient-specific plan of surgery) and treating activities separately for each hand. These experiments would create more homogeneous clusters of data and better explain some observations. Similarly, if available dataset contained more procedures or surgical sites, the conclusions about the surgical practice would be more solid.

7.3.3 Method

The LSTM model, easily capturing distant temporal dependences, was chosen as a well appropriate method for sequential problems as surgical workflow. This method also enables an on-line analysis and a straightforward knowledge transfer. Despite that, one could argue that deep learning does not suit the problems with limited quantities of training data. We demonstrated the ways to overcome this obstacle. However, to confirm that our conclusions are independent from the method, other sequential approaches should also be tested.

7.3.4 Model complexity and performance

In this work, quite simple LSTM models were used. We have chosen basic architectures to be able to concentrate more effort on auxiliary aspects of training as preliminary data analysis and knowledge transfer that also contribute in high performance. However, an advanced research on deep architectures (i.e., type, number and order of layers, etc.)

and network parameters (i.e., number of neurons on each layer, optimizer, loss and activation functions, learning rate, etc.) is needed for two reasons. The first is to improve the performance even more. For input signals analysis, it will lead to further minimization of the number of sensors necessary for an accurate activity recognition. In case of transfer, it will minimize the needed amount of knowledge. The second reason is to eliminate uncertainties. In this works, multiple conclusions were made by observing the dependences between the input and output of the network relying on causality principle. However, some uncertainties may hide biases and distort true causal relationships. It means that in some cases a particular output can be caused by randomness of choice and not by an actual change in the input. Thus, to confidently prove our observations, recognition or prediction errors must be minimized.

7.3.5 Online vs. Offline

This thesis was focused on methods enabling on-line processing and analysis where the system makes its decision based on the past and present information only. The validation metrics and targeted applications were also mainly placed in the on-line context. Yet, the off-line methods and applications could have been considered. For example, LSTM could have bidirectional connections improving the performance. The separate metrics for on-line and off-line use would also be relevant as the requirements for recognition are often not the same.

7.4 Perspectives

7.4.1 Semantic distance

In the previous chapter, we mentioned the idea of using semantic distance as a validation metric to better assess activity recognition and get a deeper understanding of method's capacities. This distance can also be explicitly integrated in the method itself to give a better idea of how the activities and their elements should be interpreted. The distance can be provided as part of the input, or be integrated in the reasoning process. The semantic similarity can be defined as a value computed as cosine distance between word embeddings, or defined with ontology as a class in a hierarchical system of objects and concepts. Sure enough, the idea of semantic distance has to be discussed with medical experts in the first place.

7.4.2 Surgical practice analysis

In our opinion, the future of computer assistance is tightly connected to exploration and understanding of surgical processes and practices, since new intelligent ORs will be

centred on the surgeon and process. It is thus necessary to go further in their analysis. The observations about surgical practices made in this thesis have still to be confirmed in a rigorous study similar to [Riffaud 2012]. Yet, they provide a good base for further research. An analytical approach analogical to ours can also be tested in the context of other tasks as, for example, prediction of remaining operating time or surgical skill evaluation. This will help to discover aspects of the surgical process imperceptible during activity recognition and prediction. Nevertheless, other approaches of knowledge extraction and analysis must to be explored as well.

7.4.3 Knowledge extraction

In this thesis, the extracted knowledge about the surgical process, which got encoded in the network's internal parameters, was only a consequence of learning tasks and not their objective. It is thus possible the extracted information to be incomplete and inaccurate representation of the process. Making the knowledge extraction the main task could result in a more accurate and comprehensive representation. There is a class of deep learning approaches called auto-encoders specifically designed for optimal extraction of relevant and representative information [Bengio 2013, Kočiskỳ 2016]. They are often used in unsupervised learning when no labels for input data provided. The auto-encoders are trained to accurately reconstruct input signal after a certain distortion which forces them to extract and encode the most important information about the data in a very efficient way. The auto-encoders could be used to capture essential knowledge of the surgical process. Further, of course, an effort should be done to transform the output of the auto-encoder into a human-readable representation.

7.4.4 Knowledge validation

Before using extracted knowledge for analysis or any other manipulation on surgical workflow, it has to be validated first. It means that one has to ensure that the representation about the domain that the network learnt is valid. For instance, a network predicting next activities can also be used to generate new possible activity sequences using its own predictions about the past as input. The validity of these generated sequences can be verified with ontologies or gSPMs. Passing this knowledge validation step proves the ability of the network to accurately capture connections between different components of surgical process and correctly understand its idea.

7.4.5 Visualization and understanding

Transparency is a very important property for a medical system. Clinicians tend to be more willing to accept a new system if they can understand it functioning and follow its decision making process. Deep learning, however, has been criticized for the nature

of its learning process that is poorly understood. In spite that the neural networks were inspired by the functioning of the human brain, their decisions are hardly understandable by humans as they follow no specific set of rules that could explain them. This may create mistrust on behalf of the surgeons. An effort was done to visualize the learning process of deep networks in order to understand *what* exactly they learn [Yosinski 2015]. Although, the answer to the question "*why* exactly do they learn what they learn?" for now remains a mystery. In addition, the research on understanding and visualization of the learning process was mostly done for an image-based signal. The same is much more harder to achieve for non-visual sequential data. Thus, looking inside LSTM, understanding and visualizing its learning process would be a great step forward. Clear explanation of output values would give more information about the system reasoning, ensure its stability and provide convincing evidences of its efficiency to clinicians. All this would help to painlessly integrate the system in the OR and introduce it in an every-day clinical routine.

7.4.6 Community

This thesis represented a small step towards a well-rounded process of surgical workflow recognition and understanding. Nevertheless, the discussed problems can be truly resolved with a join effort of the entire research community only. Their solution requires a highly cohesive collaboration between clinicians and computer scientists, as well as between research centres. For a more productive work, the community needs to create a common platform for data sharing, which also proposes a complete and versatile processing pipeline handling tasks related to surgical workflow starting from data acquisition to tools for final evaluation of clinical effect. The creation of this platform implies conception of multiple solid standards, e.g. for data format and validation process. Several collaborative initiatives were already launched for computer assisted medical interventions (CAMI). In France, for example, CamiTK framework [Fouard 2012] "helps researchers and clinicians to easily and rapidly collaborate in order to prototype CAMI applications that feature medical images, surgical navigation and biomechanical simulations". A french work group M2CAMI also decided to create a platform for sharing data and code implementations for modelling and monitoring of CAMI. Only working all together on these important problems will make the domain truly evolve.

APPENDIX

A Résumé étendu de la thèse

Contexte de la thèse

La *chirurgie assistée par ordinateur* est un domaine de recherche dans lequel la technologie informatique sert à guider et aider les chirurgiens pendant tout le processus chirurgical. Cette assistance chirurgicale peut intervenir durant la formation, ainsi que durant les périodes pré, per- et post-opératoires sous la forme de simulateurs [Gallagher 2005], planification et aide à la décision [Garg 2005], assistance robotique [Lanfranco 2004], navigation et guidage par image [Peters 2006] et analyse et optimisation des procédures [Schumann 2015].

Au début des années 2000, le concept de bloc opératoire intelligent, plus communément appelé « *OR of the future* », a été développé par la communauté scientifique et est aujourd'hui l'axe central de recherche dans le domaine [Cleary 2005, Feussner 2003, Satava 2003]. La principale innovation d'un tel type de salle opératoire réside dans l'assistance chirurgicale de manière *contextuelle*. Dans ce cadre, le system contextuel doit intervenir en prenant en compte le déroulement actuel de la procédure, en observant constamment la scène chirurgicale avec ses différents acteurs, ainsi qu'être au courant d'événements importants, d'actions réalisées et du besoin des cliniciens.

La reconnaissance automatique du flux de travail chirurgical est probablement la fonctionnalité la plus utile du système contextuel d'un point de vue clinique. Dans ce contexte, *le flux de travail* décrit une séquence de tâches chirurgicales accomplies afin de réaliser une procédure suivant un schéma répétitif. Le flux de travail de chaque intervention est unique en raison des spécificités liées au patient et au chirurgien. Cependant, il est possible à partir de ces modèles spécifiques de créer un modèle abstrait représentant l'ensemble des chirurgies appelé Modèle de Processus Chirurgical [Lalys 2014], utile et nécessaire pour la reconnaissance du flux de travail.

Etat de l'art

Spécialité chirurgicale

Chaque spécialité est caractérisée par ses propres objectifs médicaux, ses techniques interventionnelles, configuration du bloc opératoire (ce qui implique les intruments et dispositifs chirurgicaux ainsi que les conditions de travail) et ses besoins en assistance informatique. Actuellement, la reconnaissance automatique flux de travail est principalement étudiée dans le contexte de la coelioscopie manuelle [Bouarfa 2011, Padoy 2012, Twinanda 2017] et également assistée par robot [Malpani 2016, Zia 2017]. Certines études sont également menées en neurochirurgie [Forestier 2015, Lalys 2011], enophtalmologie [Lalys 2012], oto-rhino-laryngologie [Meißner 2014, Unger 2014] et chirurgie dentaire [Katić 2010].

Données

Afin de créer un système capable de reconnaître le processus chirurgical, un ensemble de données est nécessaire pour la création du modèle. Pour se faire, plusieurs types de données cliniques sont utilisés dans la littérature : mise à part les données cliniques liées à une intervention sur patient, des données issues de simulation [Bardram 2011, Padoy 2009, Tran 2017], de modèles animaux [Thiemjarus 2012], de fantômes anatomiques [Ahmadi 2009, Meißner 2014] et des données générées de manière artificielles [Katić 2010, Weede 2012] sont également employées lors de l'entraînement de ces modèles.

Niveaux de granularité

Le niveau de granularité détermine la précision avec laquelle la procédure chirurgicale est modélisée et reconnue. De manière hiérarchique ascendante, les niveaux de granularités sont définis de la manière suivante : dexemes, surgemes, activités, étapes, phases et états. Les dexemes décrivent des gestes courts effectués d'une main et qui sont dépourvus d'objectif sémantique. Une séquence de dexemes permet alors d'accomplir un surgeme, décrit comme un geste chirurgical réalisé avec un but précis et ayant une sémantique explicite [DiPietro 2016]. Une activité est une action physique décrite principalement par le triplet structure anatomique opérée, outil chirurgical utilisé et action réalisée (autrement appelée un verbe) [Forestier 2017, Lalys 2013]. Une étape est définie comme un ensemble d'activités permettant d'accomplir un objectif chirurgical [Franke 2015a, Twinanda 2015]. Une phase représente un intervalle de temps plus long, comprenant plusieurs étapes et pouvant impliquer des interactions avec d'autres membres de l'équipe chirurgicale [Bouarfa 2011, Droueche 2014, Klank 2008, Nara 2017, Twinanda 2017]. Pour finir, un état

représente une période hors de la chirurgie elle-même, décrivant l'état du bloc opératoire [Bardram 2011, Padoy 2009].

Approches de modélisation

Les connaissances sur le processus chirurgical utilisées pour l'entraînement du système de reconnaissance peuvent être formalisées de différentes manières : en utilisant une ontologie dédiée [Katić 2015], une décomposition hiérarchique [Franke 2015a], un diagramme de transition d'état [Blum 2010], une liste séquentielle [Klank 2008] et non séquentielle [Ahmadi 2009] de tâches chirurgicales.

Approches de reconnaissance

Les approches d'apprentissage automatique sont généralement divisées en trois catégories : supervisée nécessitant une annotation complète des données d'entrée, non supervisée recherchant une structure pertinentes par ses propres moyens dans les données d'entrée, et semi-supervisée utilisée notamment lorsqu'une partie des données en entrée est étiquetée. Les approaches d'apprentissage utilisées pour la reconnaissance du flux de travail chirurgical peuvent être divisées en plusieurs groupes: les classifieurs probabilistes [Forestier 2015, Quellec 2014b, Thiemjarus 2012] et non probabilistes [Bhatia 2007, Lalys 2011], les modèles temporels [Blum 2010, Padoy 2009] et les algorithmes d'alignement de séquence [Droueche 2014, Forestier 2017], les arbres de décision [Quellec 2014a, Schreiter 2016], les approches basées règles [Katić 2010], les réseaux de neurones y compris l'apprentissage profond [Malpani 2016, Twinanda 2017], les approches par regroupement [Nara 2017, Zia 2017] et par renforcement [Dergachyova 2016], ainsi que les algorithmes génétiques [Katić 2016, Klank 2008].

Applications

La reconnaissance automatique du contexte chirurgical permet plusieurs applications cliniques telles que l'automatisation des tâches [Bardram 2011, Weede 2012], l'optimisation et la gestion du processus chirurgical [Bhatia 2007], l'entraînement et l'évaluation du praticien [Forestier 2017, Malpani 2016], l'assistance chirurgicale [Katić 2010, Lalys 2011] et l'aide à la décision [Quellec 2015].

Positionnement de la thèse

L'analyse de la littérature nous a permis de mettre en avant trois problématiques importantes qui ralentissent l'apparence et l'intégration des systèmes d'assistance contextuels. Tout d'abord, il y a très peu de recherches sur les activités chirurgicales. Pourtant, ces

dernières permettent une compréhension profonde de la situation et participent également à la détection des événements indésirables intervenant au bloc opératoire. Les activités sont toutefois difficiles à reconnaître uniquement à partir d'un signal brut. Une abstraction du signal physique en faveur d'une description sémantique de la scène fournit plus de sens à la situation. Un bon choix de description permet alors de faciliter et améliorer la reconnaissance. Cependant, aucune étude n'a encore été proposée pour analyser la pertinence des différentes informations sémantiques dans le contexte de reconnaissance d'activités chirurgicales. Dans ces travaux, nous proposons une nouvelle approche pour l'analyse des composantes sémantiques de l'activité afin de caractériser les éléments essentiels et nécessaires pour une reconnaissance de haute qualité, ainsi que des recommandations pour un choix de capteurs judicieux au bloc opératoire.

La seconde problématique s'oriente vers le manque de données d'entraînement. Aujourd'hui, l'apprentissage profond dépasse fortement les méthodes d'apprentissage classiques dans de nombreux domaines. Cependant, pour un fonctionnement réussi, il nécessite d'avoir d'énormes quantités de données. Pourtant, toutes les études sur la reconnaissance d'activité que nous avons examinées contiennent en moyenne seulement 17 interventions enregistrées. Dans cette thèse, nous explorons des techniques alternatives d'augmentation de la base de données d'entrainement, appelées transfert de connaissances, qui aident à améliorer les résultats de la reconnaissance.

Enfin, la troisième problématique concerne le processus de validation. Les métriques standards fréquemment utilisées dans la littérature ne sont pas assez informatives, pas adaptées aux problèmes séquentiels et sont déconnectées des applications cliniques. Un processus de validation inadéquat ralentit l'intégration et l'utilisation de systèmes contextuels. Nous proposons donc une nouvelle vision sur le processus de validation ainsi que des stratégies et des métriques adaptées à la reconnaissance du flux de travail chirurgical.

Données

Ce travail repose sur l'utilisation de données acquises durant des chirurgies du rachis cervical par voie antérieure, (ACDF) [Forestier 2013], des hernies discale lombaire (LDH) [Riffaud 2010], des adénomes d'hypophyse (PA) [Lalys 2010] et des chirurgies de la cataracte (CS) [Lalys 2013] précédemment enregistrées (Figure A.1). Les données neurochirurgicales (c'est-à-dire ACDF, LDH et PA) ont été recueillies dans deux hôpitaux universitaires : Rennes (France) et Leipzig (Allemagne). La chirurgie de la cataracte a été enregistrée à l'hôpital universitaire de Munich (Allemagne). Au total, 154 interventions ont été acquises. Toutes les données ont été organisées par procédure et par hôpital dans 7 ensembles de données. Pour toutes les études de cette thèse, seules les annotations manuelles du flux de travail ont été utilisées, à savoir les phases et les activités chirurgicales.

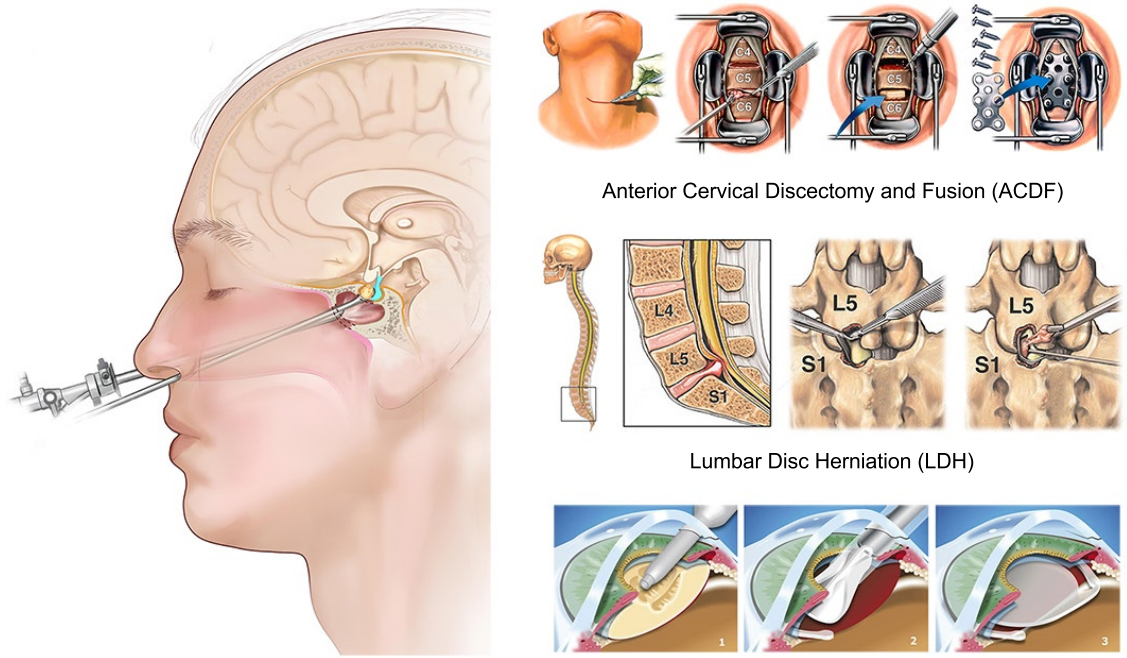

Figure A.1: Données

Analyse des signaux pour la reconnaissance d'activités

Contexte

Tenant compte de la complexité de la tâche pour la reconnaissance d'activités, les approches proposées dans la littérature décomposent l'activité en ses éléments principaux (c'est-à-dire le verbe, l'instrument et la structure anatomique) et procèdent à leur détection individuelle. L'activité est ensuite déduite à partir d'un élément ou de leur combinaison. Les éléments à détecter sont généralement choisis en fonction des signaux disponibles uniquement, sans aucune analyse de leur pertinence. L'instrument est souvent considéré comme un indicateur fiable de la tâche en cours [Bouarfa 2012, Kranzfelder 2011, Maktabi 2017], même s'il a été démontré qu'il est possible d'avoir plusieurs fonctions par instrument, qui varient en fonction de la situation et du chirurgien [Mehta 2002]. Le verbe quant à lui qui fournit des informations pertinentes sur le contexte de l'activité. Il est difficile à reconnaître en raison d'une forte variabilité de l'exécution d'une action [Meißner 2014] et nécessite souvent des capteurs supplémentaires. Enfin, la structure anatomiquepeut être reconnue à partir des signaux visuels généralement disponibles [Lalys 2013], sans avoir besoin d'installer de capteurs supplémentaires.

Cependant, aucune étude n'a permis de justifier le choix des éléments à détecter. Nous avons réalisé ce travail pour découvrir quels capteurs et signaux permettent de faciliter la reconnaissance automatique des activités chirurgicales, en nous basant sur l'hypothèse suivante : la reconnaissance de l'activité ne nécessite pas obligatoirement de capteurs pour la détection de chaque élément composant l'activité.

Méthodologie d'analyse

Afin de valider notre hypothèse et d'évaluer l'impact de chaque élément sémantique sur le processus de reconnaissance, nous proposons l'analyse suivante. Imaginons qu'à chaque instant, nous avons l'information sur les éléments et nous devons donc inférer l'activité en cours. Pour cela, nous pouvons également regarder les activités précédemment réalisées et prendre en compte leur information. Le problème consiste à faire correspondre une séquence d'activités partiellement cachées à une activité complète. Pour résoudre ce problème, nous utilisons les réseaux neuronaux profonds. Pour chaque élément, ou une combinaison d'éléments examinés, nous suggérons de masquer ses informations dans les données d'entrée et de réaliser la reconnaissance sans cette modalité pour évaluer son importance.

Dans cette étude, nous utilisons des réseaux de neurones récurrents, à savoir LSTM (Long Short-Term Memory), en supposant que les éléments cachés d'une activité dépendent non seulement des éléments actuellement connus mais aussi du contexte temporel. LSTM permet d'analyser de longues séquences avec des dépendances temporelles complexes. Dans ces travaux, nous employons une variante de LSTM classique [Graves 2012] comprenant trois portails (input, forget, output), une fonction d'activation en sortie et une technique d'abandon (Figure A.2). Afin de recréer les mêmes conditions d'analyse pour tous les éléments d'activités, le même modèle décrit ci-dessous est utilisé pour toutes les configurations et les expériences. Le modèle possède deux couches temporelles empilées avec un taux d'abandon de 0,2, chacune contenant 256 neurones cachés. Il a été entraîné pendant 50 époques avec un taux d'apprentissage de 0,001 par paquet de taille 128. L'entropie croisée est utilisée en tant que fonction de propagation de perte avec l'optimisateur Adam.

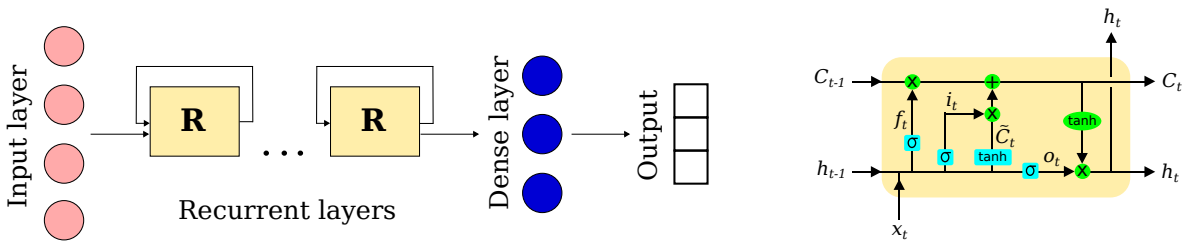

Figure A.2: Long Short-Term Memory Recurrent Neural Network

Expériences

Nous avons testé trois types de configuration : configurations à un élément (V - uniquement l'information sur le verbe était disponible, I - uniquement l'instrument, et S - uniquement la structure), configurations à deux éléments (VI - le verbe et l'instrument étaient connus, VS - le verbe et la structure, et IS - l'instrument et la structure), et la configuration à trois éléments qui a joué le rôle d'une référence (VIS - tous les éléments étaient connus).

Selon la définition utilisée dans cette thèse, l'activité est représentée par trois éléments principaux séparément définis pour les deux mains du chirurgien, ce qui donne un 6-uplet. Nous avons donc également réalisé une expérience pour estimer l'influence de chacune des mains. En outre, nous avons également évalué l'importance de connaître la durée de l'activité. Enfin, des expériences complémentaires estimant l'influence du bruit et du retard temporel dans les données ont été réalisées.

Résultats

Les expériences ont démontré qu'un seul élément n'était pas suffisant pour reconnaître l'activité avec confiance. La combinaison VI a généré des résultats faibles, insuffisants pour une reconnaissance stable et correcte (Table A.1). La combinaison VS a produit des résultats relativement bons d'environ 85% de précision, ce qui serait probablement acceptable pour certaines applications. Pour tous les procédures et les sites, la combinaison IS suffisait pour reconnaître les activités avec confiance et produisait un score supérieur à 95% (Figure A.3).

Table A.1: Précision moyenne de reconnaissance (en %) pour chaque élément de l'activité

	ACDF.L	ACDF.R	LDH.L	LDH.R	PA.L	PA.R	CS
V	49.72	66.29	52.64	62.32	48.63	68.23	**92.79**
I	**59.08**	**79.06**	59.69	**74.89**	58.17	**80.25**	90.40
S	54.50	63.27	**64.91**	68.29	**60.52**	60.08	90.18

La configuration VIS de la main droite a donné une précision d'environ 66% - le maximum parmi d'autres configurations d'une main mais toujours insuffisant pour une reconnaissance précise. La configuration IS de la main droite a généré des résultats comparables. L'expérience a aussi révélé un écart de 15 à 20% entre les performances des éléments de la main gauche et droite. L'expérience examinant l'importance de la durée de l'activité a montré que son usage en tant qu'information complémentaire d'entrée n'a que légèrement amélioré les résultats de la reconnaissance (une augmentation de 1,3 à 4,3%). Néanmoins, cette expérience a permis d'attendre une précision de reconnaissance de 98,5% pour la combinaison IS.

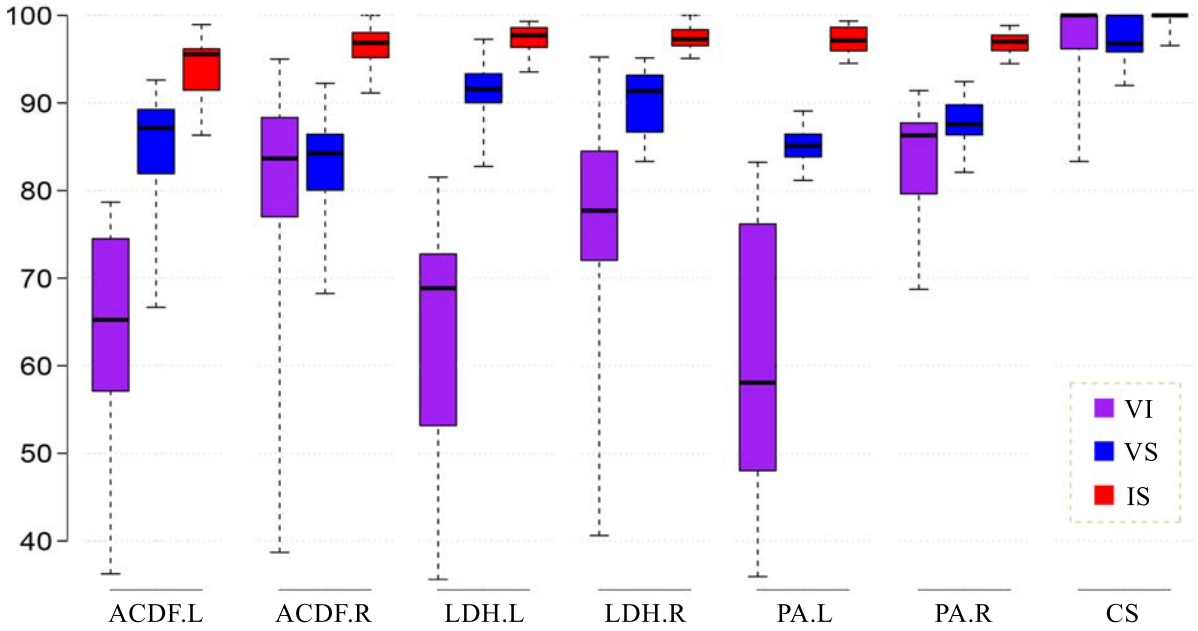

Figure A.3: Boîtes à moustaches décrivant la précision de reconnaissance pour les combinaisons d'éléments (en %)

Les expériences menées avec le bruit et le retard temporel ont montré que toutes les configurations avaient une capacité réduite à prédire l'activité en cours. Cependant, la combinaison IS est restée la plus informative. Sa précision de reconnaissance variait de 79 à 84,4% avec 5% de bruit, et a diminué à une moyenne de 6,6% avec 75% de bruit. La même combinaison a retenu des scores très élevés pour un retard d'une seconde, allant de 91 à 97,9%, avec une moyenne de 94,1%.

Conclusion

Nous avons validé notre hypothèse en démontrant que la combinaison de la structure anatomique et de l'instrument (avec les capteurs correspondants) est suffisante pour une reconnaissance précise d'une activité chirurgicale. Toutefois, le suivi des deux mains est nécessaire. La structure anatomique est un élément crucial. L'instrument et le verbe, au contraire, contiennent des informations similaires. Cette connaissance facilitera le choix des capteurs appropriés à installer dans les futurs blocs opératoires. En outre, nous avons également fait des observations intéressantes sur la pratique chirurgicale. Par exemple, dans toutes les procédures réalisées à Rennes, il existait une liaison plus forte entre l'instrument et la structure, ainsi qu'entre le verbe et la structure. Notre hypothèse est que, contrairement à Rennes, les instruments chirurgicaux à Leipzig sont plus souvent utilisés pour de nouvelles fonctions plutôt que pour leur application initialement prévue. De telles observations améliorent la compréhension du processus chirurgical.

Transfert de connaissance pour la prédiction d'activités

Contexte

Aujourd'hui, dans une nouvelle ère d'apprentissage profond, la quantité de données devient un facteur prépondérant. Malheureusement, il arrive parfois qu'une tâche d'apprentissage soit effectuée sur un petit jeu de données, comme, par exemple, la reconnaissance ou la prédiction d'activités chirurgicales avec un déficit de données cliniques. Afin de résoudre ce problème, nous proposons d'utiliser le *transfert de connaissances*. Ce transfert comprend toutes les méthodes qui utilisent des ressources provenant d'autres domaines d'intérêt pour améliorer l'apprentissage d'une tâche ciblée. Toutefois, le transfert de connaissance du processus chirurgical n'a jamais été réalisé auparavant. Pour l'appliquer, nous avons choisi d'étudier la tâche de prédiction de la prochaine l'activité.

Méthodologie

Nous proposons deux méthodes de transfert de connaissances. La première méthode dite de « word embedding » (plongement de mots en français) sert à extraire de la connaissance sémantique des termes chirurgicaux à partir de textes médicaux. La seconde méthode représente l'apprentissage par transfert qui permet de capturer des informations importantes sur le processus chirurgical et de les transférer d'une chirurgie à l'autre. Ainsi, notre travail se compose de deux parties interconnectées, car la deuxième méthode est basée sur la première et utilisent ses résultats (Figure A.4).

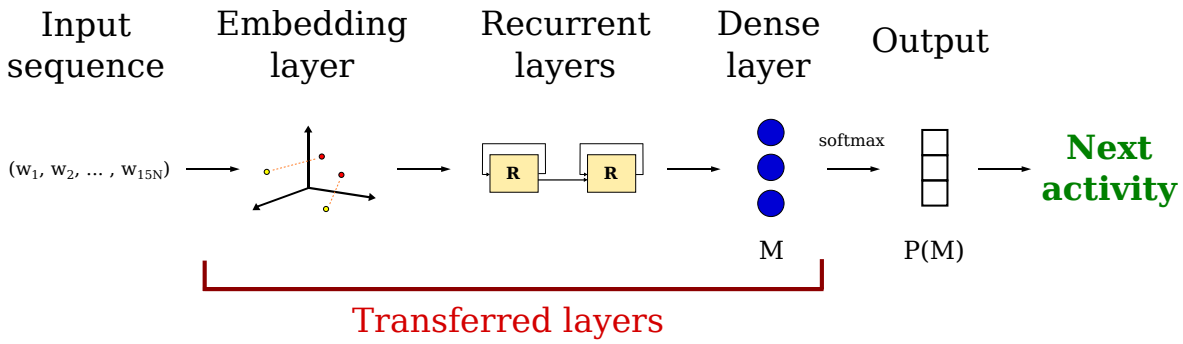

Figure A.4: Le modèle utilisé pour le transfer

Word embedding

Le « *word embedding* » est une famille de méthodes provenant du traitement du langage naturel qui cherchent à exprimer la signification sémantique des mots dans un espace géométrique. Pour cela, il est nécessaire de réaliser l'association d'un vecteur de nombres

réels à chaque mot dans le dictionnaire afin que la distance entre les vecteurs capture les relations sémantiques entre les mots correspondants. Les valeurs des vecteurs sont trouvées en fonction de l'information sur la co-occurrence des mots, c'est-à-dire la fréquence de leur apparition mutuelle dans un grand corpus de texte. Le *corpus de texte* est une collection de textes sur un sujet donné, à la forme d'une simple séquence de mots séparés par des espaces uniques. Pour cette étude, nous avons créé trois corpus de texte à l'aide des jeux de données disponibles sur internet et en obtenant des données à l'aide des moteurs de recherche scientifiques. Nous avons également utilisé deux méthodes de « word embedding » word2vec et GloVe que nous avons intégrés dans l'architecture LSTM utilisée pour prédire la prochaine activité.

Apprentissage par transfert

Nous émettons l'hypothèse que les séquences d'activités représentant un processus chirurgical encodent une certaine forme de connaissance abstraite sur la procédure donnée, la pratique chirurgicale et le processus en général. Ces connaissances peuvent être extraites et exploitées pour améliorer toutes sortes d'opérations sur les données de processus chirurgicaux, y compris l'analyse, la reconnaissance et la prédiction. On a particulièrement supposé que les connaissances obtenues d'une procédure peuvent améliorer la prédiction des activités pour une autre procédure. L'apprentissage par transfert consiste à entraîner un modèle sur un jeu de données, dit « *source* », pour extraire ses connaissances sous la forme de paramètres internes du réseau (c'est-à-dire les poids des couches cachées) et les appliquer pendant le processus d'entraînement sur un autre ensemble de données, dit « *destination* »(c'est-à-dire initialiser le nouveau modèle avec des poids extraits).

Expériences

Nous avons utilisé les procédures neurochirurgicales pour démontrer l'effet du transfert de connaissances. Le facteur principal observé à travers l'étude entière est Δ - la quantité d'amélioration de la prédiction entre différentes configurations (c'est-à-dire entre l'entraînement sans et avec « embeddings », sans apprentissage par transfert et avec différents types de transfert).

Pour les « word embeddings », nous avons établi la référence en utilisant le modèle initial sans « embedding ». Ensuite, nous avons évalué le modèle qui entraîne des « embeddings » à partir de zéro en se basant sur les procédures neurochirurgicales. Enfin, nous avons estimé l'impact des « embeddings » pré-entraînés contenant des informations sémantiques extraites des textes médicaux.

Pour l'apprentissage par transfert, nous avons cherché à estimer la quantité de connaissances pouvant être transférées à l'intérieur du domaine de neurochirurgie et à trouver les meilleures paires d'ensembles de données source-destination afin d'améliorer la capacité de prédiction du modèle. D'abord nous avons effectué une expérience pour établir

la référence. Nous avons ensuite effectué une deuxième expérience pour découvrir comment un mélange de différentes données dans un seul jeu de données d'entrée sans transfert modifie la précision de la prédiction. Dans la troisième expérience, les données initiales ont été divisées par le niveau d'expertise du chirurgien pour tenter d'améliorer la prédiction en séparant les chirurgiens novices des experts. La dernière expérience a permis de mesurer l'amélioration apportée par le vrai apprentissage par transfert. Pour cela, le modèle source a été entraîné en amont sur l'ensemble des données source. Ses poids internes ont été extraits et sauvegardés à part. Ils ont ensuite été utilisés comme initialisation pour le modèle destinataire, qui a ensuite été entraîné et testé sur l'ensemble des données de destination. Nous avons effectué trois types de transfert en utilisant différentes combinaisons d'ensembles de données source et de destination: inter-site, inter-procédure et inter-expertise.

Résultats

Dans les expériences avec les « word embeddings », le modèle de base a atteint 67,4 $\pm$ 12,5% de précision. L'entraînement des « embeddings » à partir de zéro sur les procédures neurochirurgicales a permis une augmentation de précision de 0,8%. Les résultats les plus élevés pour les « embeddings » pré-entraînés ont atteint 79,1 $\pm$ 7,5% avec une augmentation moyenne de 11,7% (Figure A.5). Ces derniers résultats ont été obtenus en utilisant la technique GloVe sur un corpus représentant une combinaison de tous les corpus créés (vecteurs à 500 dimensions) avec un entraînement supplémentaire sur les procédures neurochirurgicales.

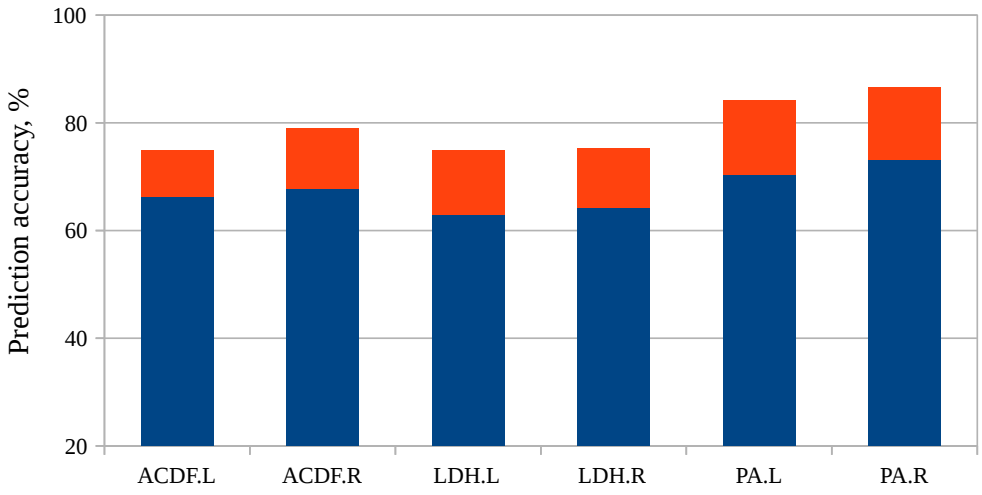

Figure A.5: L'amélioration de performance apportée par « word embedding ». Les scores de référence sont en bleu et les scores des « embeddings » pré-entraînés sont en rouge

Dans les expériences d'apprentissage par transfert, le modèle de base a obtenu un score de 78,91 $\pm$ 7,53% de précision. Le mélange des interventions des deux sites appartenant à une procédure a donné 81,42% (Δ = 2,51%), de toutes les procédures dans un site 75,89% (Δ = 2,72%), de différentes procédures entre elles 79,96% (Δ = 1,05% par rapport au modèle de base et Δ = -1,46% par rapport au mélange des sites). Le mélange de tous les ensembles de données initiaux a produit 77,54% (Δ = -1,37%) de précision.

La division des interventions par niveau d'expertise au sein d'une procédure a généré 80,16% (Δ = 1,15% du modèle de base et Δ = -1,26% du mélange des sites), au sein d'un site chirurgical 76,72% (Δ = -2,19%). La division de toutes les données a produit le résultat de 76,54% (Δ = -2,37%).

Dans le transfert inter-site (Figure A.6), en moyenne pour toutes les combinaisons source-destination, la précision a atteint 85,97% (Δ = 7,06%). Dans le transfert inter-procédure (Figure A.7) elle a atteint 86,49% (Δ = 7,58% du modèle de base et Δ = 5,07% du mélange des sites). En revanche, si nous choisissons uniquement les procédures les plus appropriées pour le transfert, la moyenne de la précision peut être recalculée à 89,09% (Δ = 10,18% du au modèle de base et Δ = 7,67% du mélange des sites). Le transfert inter-expertise à l'intérieur d'une procédure a montré une précision de 80,34% (Δ = 1,43% du modèle de base et Δ = 0,28% du modèle de division), d'un site 76,09% (Δ = -2,82% du au modèle de base et Δ = -0,63% du modèle de division), de toutes les donnée 76,23% (Δ = -2,68% du modèle de base et Δ = -0,31% du modèle de division).

Conclusion

Deux méthodes de transfert de connaissance utilisées dans ce travail ont permis d'augmenter la précision de la prédiction de près de 22% au total. Pour les « word embeddings », nous avons montré que même l'utilisation d'un corpus relativement restreint peut considérablement améliorer les résultats de la prédiction, et qu'une plus grande taille du vecteur offre de meilleurs résultats. Pour l'apprentissage par transfert, des résultats positifs ont été obtenus lors du transfert d'une procédure à l'autre, et entre les sites chirurgicaux. Plusieurs observations concernant la pratique chirurgicale ont été faites en même temps que les expériences. Nous avons observé que la différence entre les niveaux d'expertise est inférieure à la différence entre les sites chirurgicaux, et que cette dernière est plus petite que la différence entre les procédures. Entre outre, les procédures menées à Rennes semblent être plus similaire en termes de flux de travail que celles effectuées à Leipzig.

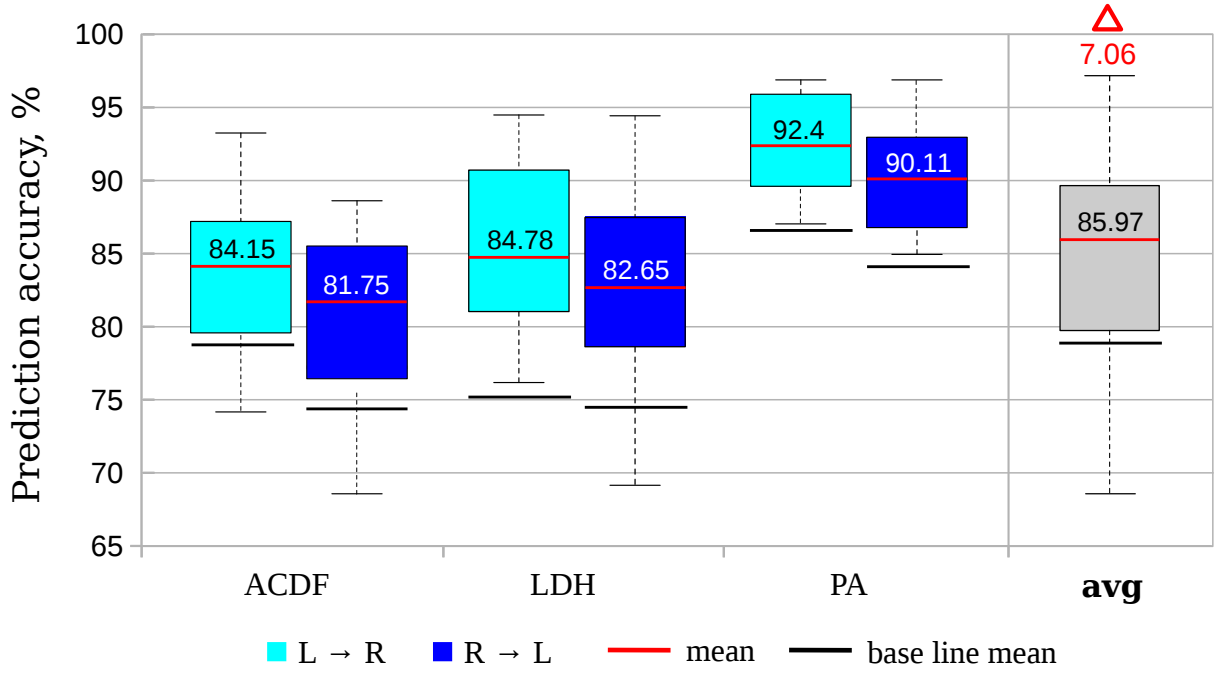

Figure A.6: Résultats du transfert inter-site

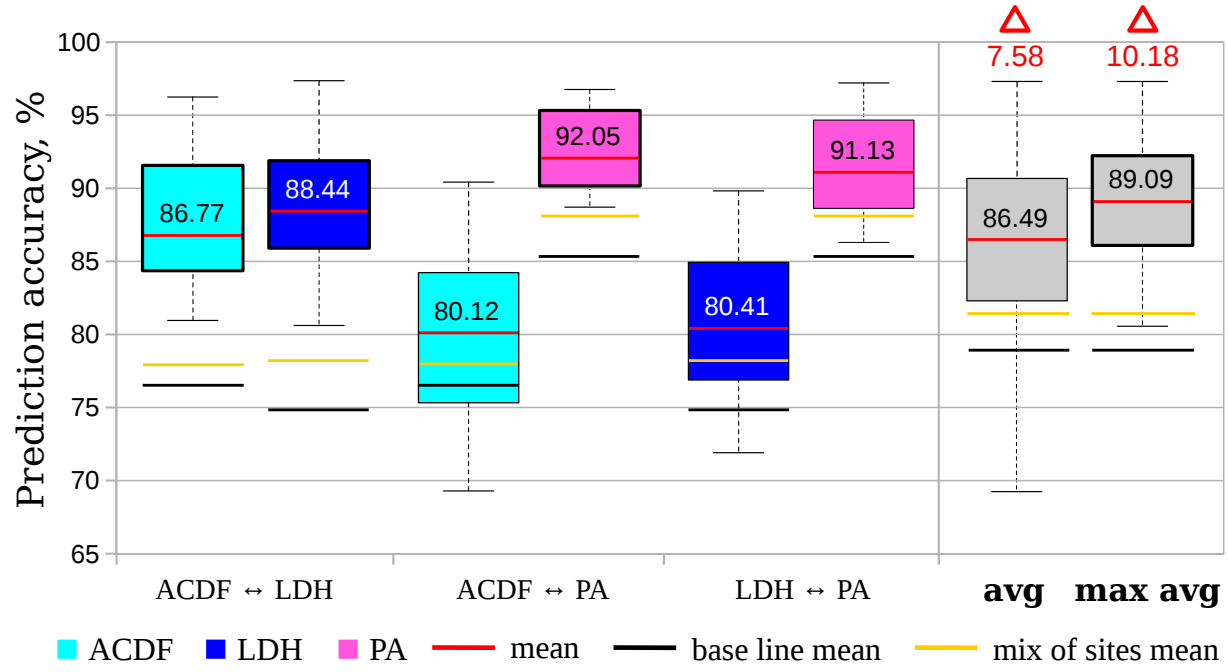

Figure A.7: Résultats du transfert inter-procédure

Métriques de validation

Contexte

Dans la plupart des publications sur la reconnaissance du flux de travail chirurgical, le processus de validation est réduit à une mesure de quelques scores de performance standards qui ne fournissent que des résultats généraux et une vague appréciation des capacités et des défauts de la méthode. Cela limite la comparaison objective des approches car plusieurs aspects de leur performance restent voilés du lecteur. Ces scores sont également très déconnectés des applications cliniques. Ils restent trop stricts et pas assez informatifs pour capturer la préparation de la méthode à son utilisation au bloc opératoire, ni son utilité pour le processus chirurgical.

Liste des métriques et des approches d'estimation proposées

Pour la reconnaissance de phase

1. *Délai moyen de transition* - pour mesurer le délai de reconnaissance entre les transitions de phases réelles et détectées,
2. *Taux de transition* - pour mesurer le nombre de transitions détectées reflétant la consistance de la reconnaissance,
3. *Taux d'échec réel* - pour mesurer la quantité de vraies défaillances de reconnaissance hors délai,
4. *Les scores spécifiques à l'application* - approche pour la ré-estimer les scores standards afin de relier la performance à l'application clinique, en prenant en compte le délai de reconnaissance acceptable.

Pour la reconnaissance et la prédiction d'activité

1. *Précision rationnelle* - pour mesurer la précision par élément pour une compréhension plus complète de la performance,
2. *Mains inversées* - métrique ré-estimant la précision pour les applications où l'inversion de mains dans la reconnaissance est acceptable,
3. *Eléments non ordonnés* - pour les applications où l'ordre des éléments reconnus est peu important, seule leur présence est essentielle,
4. *Eléments significatifs* - pour les applications où la reconnaissance de différents éléments a une importance différente pour l'assistance fournie.

Conclusion

Nous avons proposé de nouvelles métriques adaptées qui aident à extraire plus d'informations pertinentes sur la méthode et ré-évaluer les scores de performance standard pour une application ciblée. Les métriques pour la reconnaissance de phase ont été testées sur les méthodes soumises au défi MICCAI M2CAI 2016 sur la reconnaissance du flux de travail pour la chirurgie endoscopique [1]. Les métriques pour les activités ont été testées sur notre méthode de prédiction de prochaine(s) activité(s). Les résultats ont révélé des propriétés intéressantes et démontré comment les méthodes pourraient être liées à une application clinique.

Conclusion générale

Dans ce travail, nous avons abordé trois problématiques importantes de reconnaissance du flux de travail chirurgical: 1) la difficulté de reconnaissance des activités chirurgicales, 2) le manque de données cliniques pour l'entraînement, et 3) le processus de validation inadéquat. Pour résoudre le premier problème, nous avons mené une étude portant sur l'importance de chaque élément sémantique de l'activité (c'est-à-dire le verbe, l'instrument et la structure) pour faciliter et améliorer sa reconnaissance ainsi que guider le choix des capteurs appropriés à installer au bloc opératoire. Pour le deuxième problème, nous avons appliqué le transfert des connaissances en utilisant deux techniques différentes d'extraction et de transfert afin de compenser le manque de données et d'améliorer les performances d'un réseau de neurones profond. Enfin, dans le troisième problème, nous avons démontré l'insuffisance de la méthodologie de validation utilisée couramment et proposé de nouvelles métriques et approches d'estimation de performance adaptées à la reconnaissance du flux de travail chirurgical. Tout au long de ce travail, nous avons également partagé des observations pertinentes sur la pratique chirurgicale qui permettent de mieux comprendre le processus chirurgical et le comportement des chirurgiens.

Perspectives

A notre avis, outre la reconnaissance elle-même, l'attention des chercheurs doit être également portée aux problèmes « autour » de ce processus qui bloquent l'émergence de systèmes contextuels. Ce travail, par exemple, a représenté un effort de résolution des problèmes importants tels que le choix des signaux et des capteurs, la quantité de données d'entraînement et la validation. Néanmoins, d'autres aspects importants doivent également être abordés. Tout d'abord, dans nos études, nous avons appliqué

1. http://camma.u-strasbg.fr/m2cai2016/index.php

l'apprentissage profond à la résolution de plusieurs problèmes. Actuellement, les modèles profonds représentent une « boîte noire » dont le fonctionnement n'a pas encore été complètement expliqué. La visualisation et la meilleure compréhension de leur processus d'apprentissage seront très pertinentes pour les systèmes médicaux. Deuxièmement, le processus chirurgical lui-même doit être mieux étudié pour améliorer sa propre modélisation. Enfin, un standard de validation commun doit être établi pour l'évaluation et la comparaison objective des méthodes de reconnaissance développées.

Bibliography

[Abu-El-Haija 2016] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan et Sudheendra Vijayanarasimhan. *YouTube-8M: A large-scale video classification benchmark*. arXiv preprint arXiv:1609.08675, 2016. Cited page 81.

[Agarwal 2006] Sheetal K Agarwal, Anupam Joshi et Tim Finin. *Context-Aware System to Create Electronic Medical Encounter Records*. PhD thesis, University of Maryland, Baltimore County, 2006. Cited page 10.

[Ahmadi 2009] Seyed-Ahmad Ahmadi, Nicolas Padoy, Kateryna Rybachuk, Hubertus Feussner, SM Heinin et Nassir Navab. *Motif discovery in OR sensor data with application to surgical workflow analysis and activity detection*. In M2CAI workshop, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2009. Cited pages 146 and 147.

[Arora 2010] Sonal Arora, Louise Hull, Nick Sevdalis, Tanya Tierney, Debra Nestel, Maria Woloshynowych, Ara Darzi et Roger Kneebone. *Factors compromising safety in surgery: stressful events in the operating room*. The American Journal of Surgery, vol. 199, no. 1, pages 60–65, 2010. Cited page 4.

[Auer 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak et Zachary Ives. *Dbpedia: A nucleus for a web of open data*. The semantic web, pages 722–735, 2007. Cited pages 80 and 81.

[Bardram 2011] Jakob E Bardram, Afsaneh Doryab, Rune M Jensen, Poul M Lange, Kristian LG Nielsen et Søren T Petersen. *Phase recognition during surgical procedures*

using embedded and body-worn sensors. In IEEE International Conference on Pervasive Computing and Communications (PerCom), pages 45–53, 2011. Cited pages 58, 146, and 147.

[Béjar Haro 2012] Benjamín Béjar Haro, Luca Zappella et René Vidal. *Surgical gesture classification from video data*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 34–41, 2012. Cited pages 13 and 21.

[Bengio 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent et Christian Jauvin. *A neural probabilistic language model*. Journal of machine learning research, vol. 3, no. Feb, pages 1137–1155, 2003. Cited pages 84 and 87.

[Bengio 2013] Yoshua Bengio, Aaron Courville et Pascal Vincent. *Representation learning: A review and new perspectives*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pages 1798–1828, 2013. Cited page 143.

[Bharathan 2013] Rasiah Bharathan, Rajesh Aggarwal et Ara Darzi. *Operating room of the future*. Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 27, no. 3, pages 311–322, 2013. Cited page 4.

[Bhatia 2007] Beenish Bhatia, Tim Oates, Yan Xiao et Peter Hu. *Real-time identification of operating room state from video*. In Proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI), volume 2, pages 1761–1766, 2007. Cited pages 9 and 147.

[Blum 2010] Tobias Blum, Hubertus Feußner et Nassir Navab. *Modeling and segmentation of surgical workflow from laparoscopic video*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 400–407, 2010. Cited page 147.

[Bodenstedt 2017] Sebastian Bodenstedt, Martin Wagner, Darko Katić, Patrick Mietkowski, Benjamin Mayer, Hannes Kenngott, Beat Müller-Stich, Rüdiger Dillmann et Stefanie Speidel. *Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis*. arXiv preprint arXiv:1702.03684, 2017. Cited page 58.

[Bouarfa 2011] Loubna Bouarfa, Pieter P Jonker et Jenny Dankelman. *Discovery of high-level tasks in the operating room*. Journal of Biomedical Informatics, vol. 44, no. 3, pages 455–462, 2011. Cited pages 131 and 146.

[Bouarfa 2012] Loubna Bouarfa et Jenny Dankelman. *Workflow mining and outlier detection from clinical activity logs*. Journal of Biomedical Informatics, vol. 45, no. 6, pages 1185–1190, 2012. Cited pages 10, 58, 59, and 149.

[Cadene 2016] Remi Cadene, Thomas Robert, Nicolas Thome et Matthieu Cord. *M2CAI Workflow Challenge: Convolutional Neural Networks with Time Smoothing and Hidden Markov Model for Video Frames Classification.* arXiv preprint arXiv:1610.05541, 2016. Cited page 125.

[Chakraborty 2013] Ishani Chakraborty, Ahmed Elgammal et Randall S Burd. *Video based activity recognition in trauma resuscitation.* In 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pages 1–8, 2013. Cited page 13.

[Cho 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk et Yoshua Bengio. *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078, 2014. Cited page 63.

[Cleary 2005] Kevin Cleary et A Kinsella. *OR 2020: the operating room of the future.* Journal of laparoendoscopic & advanced surgical techniques. Part A, vol. 15, no. 5, pages 495–497, 2005. Cited pages 4 and 145.

[Collobert 2008] Ronan Collobert et Jason Weston. *A unified architecture for natural language processing: Deep neural networks with multitask learning.* In Proceedings of the 25th international conference on Machine learning, pages 160–167, 2008. Cited page 84.

[Collobert 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu et Pavel Kuksa. *Natural language processing (almost) from scratch.* Journal of Machine Learning Research, vol. 12, no. Aug, pages 2493–2537, 2011. Cited page 140.

[Cruz-Roa 2013] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi et Fabio Augusto González Osorio. *A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection.* Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 403–410, 2013. Cited page 35.

[Dai 2007] Wenyuan Dai, Qiang Yang, Gui-Rong Xue et Yong Yu. *Boosting for transfer learning.* In Proceedings of the 24th international conference on Machine learning, pages 193–200, 2007. Cited page 81.

[Deng 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei. *Imagenet: A large-scale hierarchical image database.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255, 2009. Cited page 81.

[Dergachyova 2016] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi et Pierre Jannin. *Automatic data-driven real-time segmentation and recognition of surgical workflow*. International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. 6, pages 1081–1089, 2016. Cited pages 77, 118, 125, 131, and 147.

[Despinoy 2016] Fabien Despinoy, David Bouget, Germain Forestier, Cédric Penet, Nabil Zemiti, Philippe Poignet et Pierre Jannin. *Unsupervised trajectory segmentation for surgical gesture recognition in robotic training*. IEEE Transactions on Biomedical Engineering, vol. 63, no. 6, pages 1280–1291, 2016. Cited page 25.

[DiPietro 2016] Robert DiPietro, Colin Lea, Anand Malpani, Narges Ahmidi, S Swaroop Vedula, Gyusung I Lee, Mija R Lee et Gregory D Hager. *Recognizing surgical activities with recurrent neural networks*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 551–558, 2016. Cited pages 13, 21, 58, and 146.

[Droueche 2014] Zakarya Droueche, Gwenole Quellec, Mathieu Lamard, Guy Cazuguel, Beatrice Cochener et Christian Roux. *Computer-Aided Retinal Surgery using Data from the Video Compressed Stream*. International Journal of Image and Video Processing: Theory and Application, vol. 1, no. 1, pages 1–10, 2014. Cited pages 146 and 147.

[Fard 2016] Mahtab J Fard, Abhilash K Pandya, Ratna B Chinnam, Michael D Klein et R Darin Ellis. *Distance-based time series classification approach for task recognition with application in surgical robot autonomy*. The International Journal of Medical Robotics and Computer Assisted Surgery, 2016. Cited page 9.

[Feussner 2003] Hubertus Feussner. *The operating room of the future: a view from Europe*. In Seminars in Laparoscopic surgery, volume 10, pages 149–156, 2003. Cited pages 4 and 145.

[Flin 2013] Rhona Flin. *Non-technical skills for anaesthetists, surgeons and scrub practitioners (ANTS, NOTSS and SPLINTS)*. The Healthcare Foundation, pages 1–9, 2013. Cited page 9.

[Forestier 2012] Germain Forestier, Florent Lalys, Laurent Riffaud, Brivael Trelhu et Pierre Jannin. *Classification of surgical processes using dynamic time warping*. Journal of Biomedical Informatics, vol. 45, no. 2, pages 255–264, 2012. Cited pages 58 and 114.

[Forestier 2013] Germain Forestier, Florent Lalys, Laurent Riffaud, D Louis Collins, Jurgen Meixensberger, Shafik N Wassef, Thomas Neumuth, Benoit Goulet et Pierre Jannin. *Multi-site study of surgical practice in neurosurgery based on surgical process models*. Journal of Biomedical Informatics, vol. 46, no. 5, pages 822–829, 2013. Cited pages 40, 58, 114, and 148.

[Forestier 2015] Germain Forestier, Laurent Riffaud et Pierre Jannin. *Automatic phase prediction from low-level surgical activities*. International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 6, pages 833–841, 2015. Cited pages 58, 60, 131, 146, and 147.

[Forestier 2017] Germain Forestier, François Petitjean, Laurent Riffaud et Pierre Jannin. *Automatic matching of surgeries to predict surgeons' next actions*. Artificial Intelligence in Medicine, 2017. Cited pages 10, 110, 118, 131, 146, and 147.

[Fouard 2012] Céline Fouard, Aurélien Deram, Yannick Keraval et Emmanuel Promayon. *CamiTK: a Modular Framework Integrating Visualization, Image Processing and Biomechanical Modeling*. In Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, pages 323–354. 2012. Cited page 144.

[Franke 2013] Stefan Franke, Jürgen Meixensberger et Thomas Neumuth. *Intervention time prediction from surgical low-level tasks*. Journal of Biomedical Informatics, vol. 46, no. 1, pages 152–159, 2013. Cited pages 9 and 10.

[Franke 2015a] Stefan Franke, Jürgen Meixensberger et Thomas Neumuth. *Multi-perspective workflow modeling for online surgical situation models*. Journal of Biomedical Informatics, vol. 54, pages 158–166, 2015. Cited pages 131, 146, and 147.

[Franke 2015b] Stefan Franke et Thomas Neumuth. *Rule-based medical device adaptation for the digital operating room*. In 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1733–1736, 2015. Cited page 10.

[Gallagher 2005] Anthony G Gallagher, E Matt Ritter, Howard Champion, Gerald Higgins, Marvin P Fried, Gerald Moses, C Daniel Smith et Richard M Satava. *Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training*. Annals of surgery, vol. 241, no. 2, page 364, 2005. Cited pages 4 and 145.

[Gao 2016a] Yixin Gao, S Swaroop Vedula, Gyusung I Lee, Mija R Lee, Sanjeev Khudanpur et Gregory D Hager. *Query-by-example surgical activity detection*. International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. 6, pages 987–996, 2016. Cited pages 13, 21, and 58.

[Gao 2016b] Yixin Gao, S Swaroop Vedula, Gyusung I Lee, Mija R Lee, Sanjeev Khudanpur et Gregory D Hager. *Unsupervised surgical data alignment with application to automatic activity annotation*. In IEEE International Conference on Robotics and Automation (ICRA), pages 4158–4163, 2016. Cited page 21.

[Garg 2005] Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, PJ Devereaux, Joseph Beyene, Justina Sam et R Brian Haynes. *Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review*. Jama, vol. 293, no. 10, pages 1223–1238, 2005. Cited pages 4 and 145.

[Glaser 2015] Bernhard Glaser, Stefan Dänzer et Thomas Neumuth. *Intra-operative surgical instrument usage detection on a multi-sensor table*. International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 3, pages 351–362, 2015. Cited page 14.

[Glorot 2011] Xavier Glorot, Antoine Bordes et Yoshua Bengio. *Domain adaptation for large-scale sentiment classification: A deep learning approach*. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 513–520, 2011. Cited page 81.

[Graves 2012] Alex Graves. Supervised sequence labelling with recurrent neural networks. Springer Berlin Heidelberg, 2012. Cited pages 63 and 150.

[Greenhalgh 2001] Trisha Greenhalgh. *Computer assisted learning in undergraduate medical education*. British Medical Journal, vol. 322, no. 7277, page 40, 2001. Cited page 10.

[Greff 2017] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink et Jürgen Schmidhuber. *LSTM: A search space odyssey*. IEEE Transactions on Neural Networks and Learning Systems, 2017. Cited page 63.

[Guédon 2016] Annetje CP Guédon, M Paalvast, FC Meeuwsen, David MJ Tax, AP van Dijke, LSGL Wauben, M van der Elst, Jenny Dankelman et JJ van den Dobbelsteen. *'It is Time to Prepare the Next patient' Real-Time Prediction of Procedure Duration in Laparoscopic Cholecystectomies*. Journal of medical systems, vol. 40, no. 12, 2016. Cited page 9.

[Haro 2012] Benjamín Béjar Haro, Luca Zappella et René Vidal. *Surgical gesture classification from video data*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 34–41, 2012. Cited page 58.

[Havaei 2017] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin et Hugo Larochelle. *Brain tumor segmentation with deep neural networks*. Medical image analysis, vol. 35, pages 18–31, 2017. Cited page 35.

[Hochreiter 1997] Sepp Hochreiter et Jürgen Schmidhuber. *Long short-term memory*. Neural computation, vol. 9, no. 8, pages 1735–1780, 1997. Cited page 60.

[Houliston 2011] Bryan Houliston, David Parry et Alan Merry. *TADAA: Towards automated detection of anaesthetic activity*. Methods of information in medicine, vol. 50, no. 5, page 464, 2011. Cited page 13.

[Huang 2013] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng et Yifan Gong. *Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7304–7308, 2013. Cited page 81.

[Huang 2016] Jian Huang, Keyang Xu et VG Vinod Vydiswaran. *Analyzing Multiple Medical Corpora Using Word Embedding*. In IEEE International Conference on Healthcare Informatics (ICHI), pages 527–533, 2016. Cited page 84.

[Huaulmé 2017a] Arnaud Huaulmé. *Détection automatique de déviations chirurgicales et identification de comportements chirurgicaux par modélisation et analyse des processus chirurgicaux*. PhD thesis, Université Grenoble-Alpes, 2017. Cited pages 10 and 25.

[Huaulmé 2017b] Arnaud Huaulmé, Sandrine Voros, Laurent Riffaud, Germain Forestier, Alexandre Moreau-Gaudry et Pierre Jannin. *Distinguishing surgical behavior by sequential pattern discovery*. Journal of Biomedical Informatics, vol. 67, pages 34–41, 2017. Cited pages 12, 13, 58, 130, and 131.

[Hübler 2014] Antje Hübler, Christian Hansen, Oliver Beuing, Martin Skalej et Bernhard Preim. *Workflow Analysis for Interventional Neuroradiology using Frequent Pattern Mining*. In Proceedings of the annual meeting of the German Society of Computer- and Robot-Assisted Surgery, pages 165–168, 2014. Cited pages 4 and 9.

[Hull 2012] Louise Hull, Sonal Arora, Rajesh Aggarwal, Ara Darzi, Charles Vincent et Nick Sevdalis. *The impact of nontechnical skills on technical performance in surgery: a systematic review*. Journal of the American College of Surgeons, vol. 214, no. 2, pages 214–230, 2012. Cited page 5.

[Jannin 2001] Pierre Jannin, Mélanie Raimbault, Xavier Morandi, E Seigneuret et Bernard Gibaud. *Design of a neurosurgical procedure model for multimodal image-guided surgery*. In International Congress Series, volume 1230, pages 102–106. Elsevier, 2001. Cited page 12.

[Jannin 2006a] Pierre Jannin, Christophe Grova et Calvin R Maurer. *Model for defining and reporting reference-based validation protocols in medical image processing*. International Journal of Computer Assisted Radiology and Surgery, vol. 1, no. 2, pages 63–73, 2006. Cited page 34.

[Jannin 2006b] Pierre Jannin, Elizabeth Krupinski et Simon K Warfield. *Validation in medical image processing*. IEEE Transactions on Medical Imaging, vol. 25, no. 11, pages 1405–9, 2006. Cited page 132.

[Jannin 2007] Pierre Jannin et Xavier Morandi. *Surgical models for computer-assisted neurosurgery*. Neuroimage, vol. 37, no. 3, pages 783–791, 2007. Cited page 12.

[Jannin 2008] Pierre Jannin et Werner Korb. *Assessment of image-guided interventions*. Image Guided Interventions Technology and Applications, pages 531–549, 2008. Cited page 132.

[Jozefowicz 2015] Rafal Jozefowicz, Wojciech Zaremba et Ilya Sutskever. *An empirical exploration of recurrent network architectures*. Journal of Machine Learning Research, 2015. Cited page 63.

[Karpathy 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar et Li Fei-Fei. *Large-scale video classification with convolutional neural networks*. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pages 1725–1732, 2014. Cited page 81.

[Katić 2010] Darko Katić, Gunther Sudra, Stefanie Speidel, Gregor Castrillon-Oberndorfer, Georg Eggers et Rüdiger Dillmann. *Knowledge-based situation interpretation for context-aware augmented reality in dental implant surgery*. Medical Imaging and Augmented Reality, pages 531–540, 2010. Cited pages 9, 146, and 147.

[Katić 2015] Darko Katić, Chantal Julliard, Anna-Laura Wekerle, Hannes Kenngott, Beat Peter Müller-Stich, Rüdiger Dillmann, Stefanie Speidel, Pierre Jannin et Bernard Gibaud. *LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase recognition*. International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 9, pages 1427–1434, 2015. Cited page 147.

[Katić 2016] Darko Katić, Jürgen Schuck, Anna-Laura Wekerle, Hannes Kenngott, Beat Peter Müller-Stich, Rüdiger Dillmann et Stefanie Speidel. *Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy*. International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. 6, pages 881–888, 2016. Cited pages 58 and 147.

[Kersten-Oertel 2013] Marta Kersten-Oertel, Pierre Jannin et D Louis Collins. *The state of the art of visualization in mixed reality image guided surgery*. Computerized Medical Imaging and Graphics, vol. 37, no. 2, pages 98–112, 2013. Cited page 4.

[Klank 2008] Ulrich Klank, Nicolas Padoy, Hubertus Feussner et Nassir Navab. *Automatic feature generation in endoscopic images*. International Journal of Computer Assisted Radiology and Surgery, vol. 3, no. 3, pages 331–339, 2008. Cited pages 146 and 147.

[Kočiskỳ 2016] Tomáš Kočiskỳ, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom et Karl Moritz Hermann. *Semantic parsing with semi-supervised sequential autoencoders*. arXiv preprint arXiv:1609.09315, 2016. Cited page 143.

[Kranzfelder 2011] Michael Kranzfelder, Armin Schneider, Sonja Gillen et Hubertus Feussner. *New technologies for information retrieval to achieve situational awareness and higher patient safety in the surgical operating room: the MRI institutional approach and review of the literature*. Surgical endoscopy, vol. 25, no. 3, pages 696–705, 2011. Cited pages 59 and 149.

[Krasin 2017] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan et Kevin Murphy. *OpenImages: A public dataset for large-scale multi-label and multi-class image classification*. Dataset available from https://github.com/openimages, 2017. Cited page 81.

[Lalys 2010] Florent Lalys, Laurent Riffaud, Xavier Morandi et Pierre Jannin. *Automatic phases recognition in pituitary surgeries by microscope images classification*. In International Conference on Information Processing in Computer-Assisted Interventions, pages 34–44, 2010. Cited pages 40 and 148.

[Lalys 2011] Florent Lalys, Laurent Riffaud, Xavier Morandi et Pierre Jannin. *Surgical phases detection from microscope videos by combining SVM and HMM*. International MICCAI Workshop on Medical Computer Vision, pages 54–62, 2011. Cited pages 118, 131, 146, and 147.

[Lalys 2012] Florent Lalys, Laurent Riffaud, David Bouget et Pierre Jannin. *A framework for the recognition of high-level surgical tasks from video images for cataract surgeries*. IEEE Transactions on Biomedical Engineering, vol. 59, no. 4, pages 966–976, 2012. Cited pages 131 and 146.

[Lalys 2013] Florent Lalys, David Bouget, Laurent Riffaud et Pierre Jannin. *Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures*. International Journal of Computer Assisted Radiology and Surgery, vol. 8, no. 1, pages 39–49, 2013. Cited pages 40, 58, 59, 146, 148, and 149.

[Lalys 2014] Florent Lalys et Pierre Jannin. *Surgical process modelling: a review*. International Journal of Computer Assisted Radiology and Surgery, vol. 9, no. 3, pages 495–511, 2014. Cited pages 25, 27, and 145.

[Lanfranco 2004] Anthony R Lanfranco, Andres E Castellanos, Jaydev P Desai et William C Meyers. *Robotic surgery: a current perspective*. Annals of surgery, vol. 239, no. 1, page 14, 2004. Cited pages 4 and 145.

[Lea 2012] Colin S Lea, James C Fackler, Gregory D Hager et Russell H Taylor. *Towards automated activity recognition in an intensive care unit.* In MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions, pages 19–28, 2012. Cited page 13.

[LeCun 1999] Yann LeCun, Patrick Haffner, Léon Bottou et Yoshua Bengio. *Object recognition with gradient-based learning.* Shape, contour and grouping in computer vision, pages 823–823, 1999. Cited page 82.

[Lemke 2006] Heinz U Lemke et Michael W Vannier. *The operating room and the need for an IT infrastructure and standards.* International Journal of Computer Assisted Radiology and Surgery, vol. 1, no. 3, pages 117–121, 2006. Cited page 5.

[Leskovec 2014] Jure Leskovec et Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection.* http://snap.stanford.edu/data, Juin 2014. Cited pages 80 and 81.

[Lin 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick. *Microsoft coco: Common objects in context.* In European conference on computer vision, pages 740–755, 2014. Cited page 81.

[Macario 2010] Alex Macario. *What does one minute of operating room time cost?* Journal of clinical anesthesia, vol. 22, no. 4, pages 233–236, 2010. Cited page 4.

[MacKenzie 2001] CL MacKenzie, JA Ibbotson, C Cao et AJ Lomax. *Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment.* Minimally Invasive Therapy & Allied Technologies, vol. 10, no. 3, pages 121–127, 2001. Cited page 12.

[Maier-Hein 2017] Lena Maier-Hein, Swaroop Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou*et al.* *Surgical data science: enabling next-generation surgery.* arXiv preprint arXiv:1701.06482, 2017. Cited pages 5, 7, and 8.

[Maktabi 2017] Marianne Maktabi et Thomas Neumuth. *Online time and resource management based on surgical workflow time series analysis.* International Journal of Computer Assisted Radiology and Surgery, vol. 12, no. 2, pages 325–338, 2017. Cited pages 9, 58, 59, 60, and 149.

[Malpani 2016] Anand Malpani, Colin Lea, Chi Chiung Grace Chen et Gregory D Hager. *System events: readily accessible features for surgical phase detection.* International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. 6, pages 1201–1209, 2016. Cited pages 118, 131, 146, and 147.

[Marszałek 2009] Marcin Marszałek, Ivan Laptev et Cordelia Schmid. *Actions in Context*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. Cited page 34.

[McAuley 2013] Julian McAuley et Jure Leskovec. *Hidden factors and hidden topics: understanding rating dimensions with review text*. In Proceedings of the 7th ACM conference on Recommender systems, pages 165–172, 2013. Cited page 81.

[Mehta 2002] NY Mehta, RS Haluck, MI Frecker et AJ Snyder. *Sequence and task analysis of instrument use in common laparoscopic procedures*. Surgical endoscopy, vol. 16, no. 2, pages 280–285, 2002. Cited pages 59 and 149.

[Meißner 2014] Christian Meißner, Jürgen Meixensberger, Andreas Pretschner et Thomas Neumuth. *Sensor-based surgical activity recognition in unconstrained environments*. Minimally Invasive Therapy & Allied Technologies, vol. 23, no. 4, pages 198–205, 2014. Cited pages 58, 59, 146, and 149.

[Mikolov 2013] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013. Cited pages 84, 86, 87, and 111.

[Miyawaki 2005] Fujio Miyawaki, Ken Masamune, Satoshi Suzuki, Kitaro Yoshimitsu et Juri Vain. *Scrub nurse robot system-intraoperative motion analysis of a scrub nurse and timed-automata-based model for surgery*. IEEE Transactions on Industrial Electronics, vol. 52, no. 5, pages 1227–1235, 2005. Cited page 9.

[Muneeb 2015] TH Muneeb, Sunil Kumar Sahu et Ashish Anand. *Evaluating distributed word representations for capturing semantics of biomedical concepts*. In Proceedings of BioNLP, pages 158–163, 2015. Cited page 84.

[Nakawala 2017] Hirenkumar Nakawala, Giancarlo Ferrigno et Elena De Momi. *Toward a Knowledge-Driven Context-Aware System for Surgical Assistance*. Journal of Medical Robotics Research, page 1740007, 2017. Cited page 13.

[Nara 2017] Atsushi Nara, Chris Allen et Kiyoshi Izumi. *Surgical Phase Recognition using Movement Data from Video Imagery and Location Sensor Data*. In Advances in Geocomputation, pages 229–237. Springer, 2017. Cited pages 146 and 147.

[Nessi 2015] Federico Nessi, Elisa Beretta, Giancarlo Ferrigno et Elena De Momi. *Recognition of user's activity for adaptive cooperative assistance in robotic surgery*. In 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 5276–5279, 2015. Cited page 9.

[Neumuth 2006a] T Neumuth, N Durstewitz, M Fischer, G Strauss, A Dietz, J Meixensberger, P Jannin, K Cleary, H Lemke et O Burgert. *Structured recording of intraoperative surgical workflows*. In SPIE medical imaging, volume 6145, page 61450A, 2006. Cited page 43.

[Neumuth 2006b] Thomas Neumuth, Gero Strauß, Jürgen Meixensberger, Heinz U Lemke et Oliver Burgert. *Acquisition of process descriptions from surgical interventions*. In DEXA, pages 602–611. Springer, 2006. Cited page 12.

[Neumuth 2007] T Neumuth, R Mudunuri, Pierre Jannin, J Meixensberger et O Burgert. *The tool landscap for surgical workflow analysis*. In Computer Assisted Medical and Srugical Interventions (SURGETICA), pages 199–204, 2007. Cited page 43.

[Neumuth 2011] Thomas Neumuth, Pierre Jannin, Juliane Schlomberg, Jürgen Meixensberger, Peter Wiedemann et Oliver Burgert. *Analysis of surgical intervention populations using generic surgical process models*. International Journal of Computer Assisted Radiology and Surgery, vol. 6, no. 1, pages 59–71, 2011. Cited page 12.

[Neumuth 2012] Thomas Neumuth et Christian Meißner. *Online recognition of surgical instruments by information fusion*. International Journal of Computer Assisted Radiology and Surgery, vol. 7, no. 2, pages 297–304, 2012. Cited page 14.

[Oquab 2014] Maxime Oquab, Leon Bottou, Ivan Laptev et Josef Sivic. *Learning and transferring mid-level image representations using convolutional neural networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1717–1724, 2014. Cited page 81.

[Padoy 2009] Nicolas Padoy, Diana Mateus, Daniel Weinland, Marie-Odile Berger et Nassir Navab. *Workflow monitoring based on 3d motion features*. In Proceedings of the International Conference on Computer Vision Workshop on Video-oriented Object and Event Classification, pages 585–592, 2009. Cited pages 146 and 147.

[Padoy 2010] Nicolas Padoy. *Workflow and activity modeling for monitoring surgical procedures*. PhD thesis, Université Henri Poincaré-Nancy I, 2010. Cited page 12.

[Padoy 2012] Nicolas Padoy, Tobias Blum, Seyed-Ahmad Ahmadi, Hubertus Feussner, Marie-Odile Berger et Nassir Navab. *Statistical modeling and recognition of surgical workflow*. Medical image analysis, vol. 16, no. 3, pages 632–641, 2012. Cited page 146.

[Pan 2010] Sinno Jialin Pan et Qiang Yang. *A survey on transfer learning*. IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pages 1345–1359, 2010. Cited pages 81 and 110.

[Parlak 2011] Siddika Parlak, Ivan Marsic et Randall S Burd. *Activity recognition for emergency care using RFID*. In Proceedings of the 6th International Conference on Body Area Networks, pages 40–46, 2011. Cited page 13.

[Pennington 2014] Jeffrey Pennington, Richard Socher et Christopher D Manning. *Glove: Global vectors for word representation*. In EMNLP, volume 14, pages 1532–1543, 2014. Cited pages 84 and 88.

[Peters 2006] Terry M Peters. *Image-guidance for surgical procedures*. Physics in medicine and biology, vol. 51, no. 14, page R505, 2006. Cited pages 4 and 145.

[Primus 2016] Manfred Jürgen Primus, Klaus Schoeffmann et Laszlo Böszörmenyi. *Temporal segmentation of laparoscopic videos into surgical phases*. In 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pages 1–6. IEEE, 2016. Cited pages 118 and 119.

[Quellec 2014a] Gwénolé Quellec, Katia Charriere, Mathieu Lamard, Zakarya Droueche, Christian Roux, Béatrice Cochener et Guy Cazuguel. *Real-time recognition of surgical tasks in eye surgery videos*. Medical image analysis, vol. 18, no. 3, pages 579–590, 2014. Cited page 147.

[Quellec 2014b] Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener et Guy Cazuguel. *Real-time segmentation and recognition of surgical tasks in cataract surgery videos*. IEEE Transactions on Medical Imaging, vol. 33, no. 12, pages 2352–2360, 2014. Cited page 147.

[Quellec 2015] Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener et Guy Cazuguel. *Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials*. IEEE Transactions on Medical Imaging, vol. 34, no. 4, pages 877–887, 2015. Cited page 147.

[Reiley 2011] Carol E Reiley, Henry C Lin, David D Yuh et Gregory D Hager. *Review of methods for objective surgical skill evaluation*. Surgical endoscopy, vol. 25, no. 2, pages 356–366, 2011. Cited page 4.

[Riffaud 2010] Laurent Riffaud, Thomas Neumuth, Xavier Morandi, Christos Trantakis, Jürgen Meixensberger, Oliver Burgert, Brivael Trelhu et Pierre Jannin. *Recording of surgical processes: a study comparing senior and junior neurosurgeons during lumbar disc herniation surgery*. Neurosurgery, vol. 67, pages 325–332, 2010. Cited pages 40, 44, 58, 114, and 148.

[Riffaud 2012] Laurent Riffaud. *Recording of cranial and spinal surgical procedures, and analysis of workflows for study of the surgical processes*. PhD thesis, Université Rennes 1, 2012. Cited page 143.

[Rodas 2017] Nicolas Loy Rodas, Fernando Barrera et Nicolas Padoy. *See It With Your Own Eyes: Markerless Mobile Augmented Reality for Radiation Awareness in the Hybrid Room*. IEEE Transactions on Biomedical Engineering, vol. 64, no. 2, pages 429–440, 2017. Cited page 10.

[Satava 2001] Richard M Satava. *Surgical education and surgical simulation*. World journal of surgery, vol. 25, no. 11, pages 1484–1489, 2001. Cited page 4.

[Satava 2003] Richard M Satava. *Disruptive vision. The operating room of the future*. Surgical endoscopy, vol. 17, pages 104–107, 2003. Cited pages 4 and 145.

[Schreiter 2016] L Schreiter, P Philipp, J Giehl, Y Fischer, J Raczkowsky, M Schwarz, J Beyerer et H Woern. *Situation detection for an interactive assistance in surgical interventions based on dynamic bayesian networks*. International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. Suppl 1, pages S115–S116, 2016. Cited page 147.

[Schumann 2015] Sandra Schumann, Ulf Bühligen et Thomas Neumuth. *Outcome quality assessment by surgical process compliance measures in laparoscopic surgery*. Artificial intelligence in medicine, vol. 63, no. 2, pages 85–90, 2015. Cited pages 4 and 145.

[Shin 2016] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura et Ronald M Summers. *Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning*. IEEE Transactions on Medical Imaging, vol. 35, no. 5, pages 1285–1298, 2016. Cited page 81.

[Shuhaiber 2004] Jeffrey H Shuhaiber. *Augmented reality in surgery*. Archives of surgery, vol. 139, no. 2, pages 170–174, 2004. Cited page 4.

[Stauder 2016] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner et Nassir Navab. *The TUM LapChole dataset for the M2CAI 2016 workflow challenge*. arXiv preprint arXiv:1610.09278, 2016. Cited pages 34, 119, and 133.

[Tao 2013] Lingling Tao, Luca Zappella, Gregory D Hager et René Vidal. *Surgical gesture segmentation and recognition*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 339–346, 2013. Cited page 21.

[Thiemjarus 2012] Surapa Thiemjarus, Adam James et G-Z Yang. *An eye–hand data fusion framework for pervasive sensing of surgical activities*. Pattern Recognition, vol. 45, no. 8, pages 2855–2867, 2012. Cited pages 146 and 147.

[Tran 2017] Dinh Tuan Tran, Ryuhei Sakurai, Hirotake Yamazoe et Joo-Ho Lee. *Phase segmentation methods for an automatic surgical workflow analysis*. International journal of biomedical imaging, vol. 2017, 2017. Cited pages 131 and 146.

[Twinanda 2015] Andru P Twinanda, Emre O Alkan, Afshin Gangi, Michel de Mathelin et Nicolas Padoy. *Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms*. International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 6, pages 737–747, 2015. Cited pages 58 and 146.

[Twinanda 2016] Andru P Twinanda, Didier Mutter, Jacques Marescaux, Michel de Mathelin et Nicolas Padoy. *Single-and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016*. arXiv preprint arXiv:1610.08844, 2016. Cited page 125.

[Twinanda 2017] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin et Nicolas Padoy. *Endonet: A deep architecture for recognition tasks on laparoscopic videos*. IEEE Transactions on Medical Imaging, vol. 36, no. 1, pages 86–97, 2017. Cited pages 34, 58, 81, 118, 119, 131, 133, 146, and 147.

[Unger 2014] Michael Unger, Claire Chalopin et Thomas Neumuth. *Vision-based online recognition of surgical activities*. International Journal of Computer Assisted Radiology and Surgery, vol. 9, no. 6, pages 979–986, 2014. Cited page 146.

[Wang 2016] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad et Andrew H Beck. *Deep learning for identifying metastatic breast cancer*. arXiv preprint arXiv:1606.05718, 2016. Cited page 35.

[Weede 2012] Oliver Weede, Frank Dittrich, Heinz Wörn, Brian Jensen, Alois Knoll, Dirk Wilhelm, Michael Kranzfelder, Armin Schneider et Hubertus Feussner. *Workflow analysis and surgical phase recognition in minimally invasive surgery*. In IEEE International Conference on Robotics and Biomimetics (ROBIO), pages 1080–1074. IEEE, 2012. Cited pages 24, 146, and 147.

[Werbos 1990] Paul J Werbos. *Backpropagation through time: what it does and how to do it*. Proceedings of the IEEE, vol. 78, no. 10, pages 1550–1560, 1990. Cited page 62.

[Yoo 2000] Terry S Yoo, Michael J Ackerman et Michael Vannier. *Toward a common validation methodology for segmentation and registration algorithms*. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 422–431, 2000. Cited page 132.

[Yosinski 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio et Hod Lipson. *How transferable are features in deep neural networks?* In Advances in neural information processing systems, pages 3320–3328, 2014. Cited page 82.

[Yosinski 2015] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs et Hod Lipson. *Understanding neural networks through deep visualization*. arXiv preprint arXiv:1506.06579, 2015. Cited page 144.

[Zhang 2015] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji et Dinggang Shen. *Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.* NeuroImage, vol. 108, pages 214–224, 2015. Cited page 35.

[Zia 2017] Aneeq Zia, Chi Zhang, Xiaobin Xiong et Anthony M Jarc. *Temporal clustering of surgical activities in robot-assisted surgery.* International Journal of Computer Assisted Radiology and Surgery, pages 1–8, 2017. Cited pages 131, 146, and 147.

Résumé

L'assistance informatique est devenue une partie indispensable pour la réalisation de procédures chirurgicales modernes. Le désir de créer une nouvelle génération de blocs opératoires intelligents a incité les chercheurs à explorer les problèmes de perception et de compréhension automatique de la situation chirurgicale. Dans ce contexte de prise de conscience de la situation, un domaine de recherche en plein essor adresse la reconnaissance automatique du flux chirurgical. De grands progrès ont été réalisés pour la reconnaissance des phases et des gestes chirurgicaux. Pourtant, il existe encore un vide entre ces deux niveaux de granularité dans la hiérarchie du processus chirurgical. Très peu de recherche se concentre sur les activités chirurgicales portant des informations sémantiques vitales pour la compréhension de la situation. Deux facteurs importants entravent la progression. Tout d'abord, la reconnaissance et la prédiction automatique des activités chirurgicales sont des tâches très difficiles en raison de la courte durée d'une activité, de leur grand nombre et d'un flux de travail très complexe et une large variabilité. Deuxièmement, une quantité très limitée de données cliniques ne fournit pas suffisamment d'informations pour un apprentissage réussi et une reconnaissance précise. À notre avis, avant de reconnaître les activités chirurgicales, une analyse soigneuse des éléments qui composent l'activité est nécessaire pour choisir les bons signaux et les capteurs qui faciliteront la reconnaissance. Nous avons utilisé une approche d'apprentissage profond pour évaluer l'impact de différents éléments sémantiques de l'activité sur sa reconnaissance. Grâce à une étude approfondie, nous avons déterminé un ensemble minimum d'éléments suffisants pour une reconnaissance précise. Les informations sur la structure anatomique et l'instrument chirurgical sont de première importance. Nous avons également abordé le problème de la carence en matière de données en proposant des méthodes de transfert de connaissances à partir d'autres domaines ou chirurgies. Les méthodes de « word embedding » et d'apprentissage par transfert ont été proposées. Ils ont démontré leur efficacité sur la tâche de prédiction d'activité suivante offrant une augmentation de précision de 22%. De plus, des observations pertinentes concernant la pratique chirurgicale ont été faites au cours de l'étude. Dans ce travail, nous avons également abordé le problème de la validation insuffisante et incorrecte des méthodes de reconnaissance. Nous avons proposé de nouvelles métriques et méthodes de validation pour évaluer les performances, afin de mieux relier les méthodes aux applications ciblées et de mieux caractériser leurs capacités. Le travail décrit dans cette thèse vise à éliminer les obstacles entravant l'avancement du domaine et à proposer une nouvelle perspective sur le problème de la reconnaissance du flux chirurgical.

Mots clefs : *Activités chirurgicales de bas niveau, Reconnaissance d'activités chirurgicales, Analyse sémantique, Word embedding, Apprentissage par transfert, Métriques de validation*

Abstract

Computer assistance became indispensable part of modern surgical procedures. Desire of creating new generation of intelligent operating rooms incited researchers to explore problems of automatic perception and understanding of surgical situations. Situation awareness includes automatic recognition of surgical workflow. A great progress was achieved in recognition of surgical phases and gestures. Yet, there is still a blank between these two granularity levels in the hierarchy of surgical process. Very few research is focused on surgical activities carrying important semantic information vital for situation understanding. Two important factors impede the progress. First, automatic recognition and prediction of surgical activities is a highly challenging task due to short duration of activities, their great number and a very complex workflow with multitude of possible execution and sequencing ways. Secondly, very limited amount of clinical data provides not enough information for successful learning and accurate recognition. In our opinion, before recognizing surgical activities a careful analysis of elements that compose activity is necessary in order to chose right signals and sensors that will facilitate recognition. We used a deep learning approach to assess the impact of different semantic elements of activity on its recognition. Through an in-depth study we determined a minimal set of elements sufficient for an accurate recognition. Information about operated anatomical structure and surgical instrument was shown to be the most important. We also addressed the problem of data deficiency proposing methods for transfer of knowledge from other domains or surgeries. The methods of word embedding and transfer learning were proposed. They demonstrated their effectiveness on the task of next activity prediction offering 22% increase in accuracy. In addition, pertinent observations about the surgical practice were made during the study. In this work, we also addressed the problem of insufficient and improper validation of recognition methods. We proposed new validation metrics and approaches for assessing the performance that connect methods to targeted applications and better characterize capacities of the method. The work described in this these aims at clearing obstacles blocking the progress of the domain and proposes a new perspective on the problem of surgical workflow recognition.

Keywords: *Low-level surgical activities, Surgical activity recognition, Semantic analysis, Word embedding, Transfer learning, Validation metrics*