UNIVERSITÉ DE STRASBOURG

ICube Laboratory (UMR 7357)
Research Group CAMMA on
Computational Analysis and Modeling of Medical Activities

3D Detection and Pose Estimation of Medical Staff in Operating Rooms using RGB-D Images

Thesis presented by

Abdolrahim Kadkhodamohammadi

on

1 December 2016

Submitted to the

University of Strasbourg

for obtaining the degree of

Doctor of Philosophy

delivered by the doctoral school MSII

Thesis Director:	
Mr. Michel de Mathelin	Professor, Université de Strasbourg
Thesis Supervisor:	
Mr. Nicolas Padoy	Assistant Professor on a Chair of Excellence, Université de Strasbourg
Chair of the Committee:	
Mr. Nicholas Ayache	Research Director, Inria, Sophia Antipolis
Examiners:	
Mr. Slobodan Ilic	Privatdozent, Technische Universität München
Mr. Danail Stoyanov	Senior Lecturer, University College London

Abstract

In this thesis, we address the problems of person detection and pose estimation in Operating Rooms (ORs), which are key ingredients needed to develop many applications in such environments, like surgical activity recognition, surgical skill analysis and radiation safety monitoring. Because of the strict sterilization requirements of the OR and of the fact that the surgical workflow should not be disrupted, cameras are currently one of the least intrusive options that can be conveniently installed in the room to sense the environment. Even though recent vision-based human detection and pose estimation methods have achieved fairly promising results on standard computer vision datasets, we show that they do not necessarily generalize well to challenging OR environments. The main challenges are the presence of many visually similar surfaces, loose and textureless clinical clothes, clutter, occlusions and the fact that the environment is crowded. To address these challenges, we propose to use a set of compact RGB-D cameras installed on the ceiling of the OR. Such cameras capture the environment by using two inherently different sensors and therefore provide complementary information about the surfaces present in the scene, namely their visual appearance and their distances to the camera.

In this dissertation, we propose novel approaches that take into account depth, multi-view and temporal information to perform human detection and pose estimation. Firstly, we introduce an energy optimization approach to consistently track body poses over entire RGB-D sequences. Secondly, we present a novel approach to estimate the body poses directly in 3D by relying on both color and depth images. The approach also uses a new RGB-D body part detector. Finally, we present a multi-view approach for 3D human pose estimation, which relies on depth data to reliably incorporate information across all views. We also present a method to automatically model a priori information about the OR environment for obtaining a more robust human detection model. To evaluate our approaches, we generate several single- and multi-view datasets in operating rooms. We demonstrate very promising results on these datasets and show that our approaches outperform state-of-the-art methods on data acquired during real surgeries.

Acknowledgments

I would like to acknowledge everyone who has assisted me during my PhD studies.

First and foremost, I want to thank my thesis supervisor, Dr. **Nicolas Padoy**, for the continuous support of my research, for his patience and for his enthusiasm. It has been an honor to be his first PhD student at the University of Strasbourg. I am grateful for his involvement and for always pushing me to the right direction. He always provided timely and high quality feedback on my work. I have always been fascinated by his unrivaled ability to spot weak points in my formulations, which has saved me from later embarrassment many times. I would like to also express my sincere gratitude to my thesis director, Prof. **Michel de Mathelin**, for his confidence in my abilities and his encouragement during my work. I am very thankful for his quick responses to my requests and for following up on both administrative and academic matters even though he is very busy as the director of the ICube laboratory.

Besides my thesis supervisor and director, I would like to thank other members of my thesis committee: Prof. **Nicholas Ayache**, Dr. **Slobodan Ilic** and Dr. **Danail Stoyanov** for their involvement in the evaluation of my work and for their insightful comments and questions. I am also grateful to Slobodan and Danail for reviewing my manuscript.

This work would not have been possible without close collaboration with our medical partners. I would like to thank the interventional radiology department of the University Hospital of Strasbourg and especially the head of the department, Prof. **Afshin Gangi**, for allowing us to install our camera recording systems in several operating rooms and to record data during real surgeries. Many thanks to **Nicolas Loy Rodas**, **Andru Putra Twinanda**, **Antonio De Donno** and **Fernando Barrera Campo** for their help during data acquisitions.

During my work on this thesis, I was a member of the research group CAMMA (Computational Analysis and Modeling of Medical Activities). I would like to thank all members of CAMMA for numerous fruitful discussions and friendly collaborations that we had during this project. A special thanks goes to two colleagues of mine: **Nicolas Loy Rodas** and **Andru Putra Twinanda**. I would like to thank Nicolas and Andru for always being available to help and discuss on scientific as well as nonscientific matters

and for their help in proofreading this manuscript. Profound gratitude goes to Nicolas for translating the long summary into French. I would like to thank **Laurent Goffin** for developing the core library for working with RGB-D cameras. I am grateful to **Antonio De Donno** who developed our multi-camera recording application and also the versatile annotation tool that was essential to generate high quality 2D and 3D ground-truth annotations. I would like to thank **Fernando Barrera Campo** for developing the calibration tool that allows us to easily and quickly compute camera extrinsic parameters. Also, I would like to thank the students I was working with: **Abinash Pant** who performed the experiments to evaluate the flexible mixtures of parts approach on OR data; **Kaveh Khorshidian** and **Omid Mahmoudi** who labeled hundreds of human body poses.

The research group CAMMA is a part of the research team AVR (Equipe Automatique, Vision, Robotique) within ICube laboratory at the University of Strasbourg. I would like to thank all my colleagues for creating an excellent research environment. Thank you to **Riad Khelifi**, **Paolo Cabras**, **Nitish Kumar**, **Arnaud Bruyas**, **Nadège Corbin**, **Markus Neumann**, **Laure-Anais Chanel**, **Gauthier Hentz**, **Nicole Lepoutre**, **François Schmitt** and **Cédric Girerd** for creating the atmosphere of fun in our group. I am also grateful to **Magali Darrieumerlou** and **Christelle Charles** for taking care of administrative issues.

I would like to thank my friends **Alain Yahagi**, **Javad Rassouli**, **Hoda Sheibani** and **Mohammad Mehdipour** for providing support and friendship that I needed.

Last but not the least; I would like to thanks my parents, my sisters and my brothers for their unconditional support throughout this thesis and my life in general. My dear wife, **Serveh Karimi**, deserves special thanks and gratitude for her patience, assistance, support and faith in me. This work and dissertation would not have come about without her sacrifices. I cannot thank her enough. I also want to thank my darling son, **Diyako**, whose beautiful smile gave a new meaning to our family life and brought so much joy and fulfillment in my life.

I could not have completed my PhD studies without the support of all these wonderful people!

Table of Contents

List of Figures

List of Tables

1 Introduction

Chapter Summary

A core part of the patient care system in hospitals is the surgical department. This department is responsible for preoperative consultations, for performing surgeries and for following up on patient treatments. In the surgical department, clinicians and clinical staff are the main actors. They are collaborating, making decisions and performing actions to fulfill all these responsibilities. Their main working environment is the operating room, where they perform actions based on preoperative plans to treat patients. The actions and the way the actions are performed directly impact the outcome of the treatments.

Thus, it is an important research objective to perceive, model and study the activities occurring in Operating Rooms (ORs) in order to better understand, potentially improve and enhance the patient care system, *e.g.* by analyzing more objectively surgical workflows and by providing more accurate and relevant feedback during training.

In order to achieve the aforementioned goals, localizing clinical staff as well as their body parts is of fundamental importance. As an example, let us consider surgical workflow models. They are needed in order to extract and analyze the statistical properties of a surgery. These models can then be used for monitoring the progress of a surgery, detecting anomalies and adapting the schedule of the operating room and personnel. To build such a model, it is essential to collect a substantial amount of information from different sources, such as tool usage data and patient's vital signals. Automatic localization of medical staff and of their body parts is a complementary source of information needed to provide workflow models with crucial information about the main actors and their interactions. Without approaches to automatically extract high level information from the OR, workflow analysis can only be performed through tedious manual annotations.

In order to enable clinician detection and pose estimation in operating rooms, we need to perceive the environment with a system that is not only capable of providing suitable data for performing such tasks, but that can also be conveniently integrated in such rooms with minimal invasiveness. In the following, we present several possible sensing technologies available to capture operating rooms. We then describe how clinician detection and pose estimation can contribute to the improvement of the activities taking place in the surgical department and of the health care system in general. We also present three state-of-the-art human pose estimation and detection methods as well as their qualitative performance on clinical data to illustrate their limitations in the OR environment. Finally, we discuss the contributions of this work and conclude the chapter by providing the outline of this thesis.

1.1 Perceiving Operating Room Environments

In this section, we start by briefly describing the operating room environment. We then present sensing technologies that could be used for human detection and/or human pose estimation in general. We also point out some key challenges and advantages of using these sensing technologies in operating rooms. Finally, we present the camera recording system used in our project.

1.1.1 Description of the Operating Room Environment

An operating room, also known as an operating theater, is the unit of a hospital where surgical operations are carried out. Operating rooms are designed and equipped to provide safe care to patients.

Figure 1.1 shows two operating rooms. Central to the OR is the operating table, where the patient lies. It is surrounded by the respiratory tower, a table for the surgical instruments, and the devices monitoring the patient's vital signs. Ceiling-mounted rails

(a) *Technical University of Munich, Klinikum rechts der Isar.* Courtesy of [Belagiannis 2016].

(b) *University hospital of Strasbourg, Interventional Radiology Department.*

Figure 1.1: Sample operating rooms: (a) an operating room and (b) an operating room equipped with an intra-operative CT scanner device.

or ceiling-mounted articulated arms are often used to facilitate the positioning of medical equipment, such as the surgical lamp and screens, in proximity of the operating table without hampering access to the patient. In addition, some operating rooms are equipped with intra-operative imaging devices such as an angiographic C-arm, an ultrasound device or a computed tomography scanner.

The operating room is a sterile environment. All personnel wear similar sterilized clothes called scrubs. In order to keep the environment free of germs, they also use masks over their faces, surgical caps and shoe covers.

Due to the delicate process of performing physical intervention on humans and to the safety as well as the sterilization requirements of the OR, introducing new devices or sensors requires special care and attention to avoid adding difficulties to the existing daily routine of the room.

1.1.2 Sensor Options

Overall, operating rooms can be sensed by using either wearable markers or cameras. On the one hand, with wearable markers, one needs to attach active or passive markers to the object of interest and focus on sensing these markers. Marker sensing is performed in two ways: 1) by using regular or specially designed cameras able to capture the marker with a contrast higher than the one of the background, *e.g.* infrared cameras with infrared retro-reflective markers, and 2) by using other sensors, *e.g.* using ultrasonic receivers in case of active ultrasonic markers. On the other hand, cameras can be used to perceive the whole environment that includes both the objects of interest and the background. Camera data however requires more sophisticated methods to analyze the scene.

Wearable markers. Radio Frequency Identification (RFID) tags can be used to identify

the presence or absence of people in the room. Such systems cannot however be used for localizing people. Instead, body-worn *ultra-wideband* transmitters are needed for localization. In practice, the localization system based on ultra-wideband is however very inaccurate [Bardram 2011]. More accurate person localization is possible using an ultrasonic location system. Such a system uses an array of ceiling-mounted ultrasonic receivers to measure the time of flight of an ultrasonic signal emitted by the transmitter carried by the person to be localized [Medina 2013]. Estimating the 3D location of the person requires however line of sight between a transmitter and at least three receivers. This condition is not always easy to fulfill, especially in operating rooms with many objects attached to the ceiling. In general, if the marker is worn over the scrub, it should also be sterilized according to OR's regulations. Therefore, each person has to wear a new sterilized marker upon entrance in the room. This makes it cumbersome to regularly use such systems in the OR. In addition, people may forget to wear the markers. Moreover, the aforementioned localization systems cannot be used for body part localization due to signal interferences or the line of sight requirement.

In order to estimate human poses in controlled environments, systems based on multiple inertial sensors can be used [Wong 2015]. These sensors should be placed at different places on the body, *e.g.* all body parts, to measure accelerations and orientations. However, these measurements are subject to drift, which has limited the reliability of such systems for capturing human poses over long durations [Wong 2015]. Infrared retro-reflective markers, which are tightly and precisely placed all over the body, can be used to reliably discover body part configurations. But, due to strong requirements such as the need for a high number of infrared sensors, *e.g.* 16 sensors, and for line of sight between the markers and the infrared light projectors, such systems can only be used in controlled environments with no clutter. In general, marker-based pose estimation systems require many more markers compared to marker-based people detection systems. Wearing a multitude of markers upon entrance would be a very demanding and tedious task unless a cheap and sterile surgical gown can be designed with the markers already integrated. Furthermore, the marker positions on the body often need to be manually registered with a 3D human body model, which is a delicate and time-consuming task to perform.

Cameras. Cameras allow to capture the entire scene and also do not require attaching any marker to the objects of interest. They however require robust detection methods. They are often installed in modern operating rooms for archiving and interactive teaching (*e.g.* see Figure 1.4). Soon, one can even imagine that cameras may become mandatory in the OR, in the same way that black boxes are mandatory in airplanes [Sutherland 2006, Kohn 2000]. One advantage of cameras is that they can be easily installed if an OR does not possess any. Moreover, they do not affect the daily operating room workflow.

Wearable markers vs. cameras. In general, wearable markers are intrusive. This makes them difficult to use in operating rooms. Furthermore, marker-based systems require line of sight and manual registration between a 3D body model and the marker

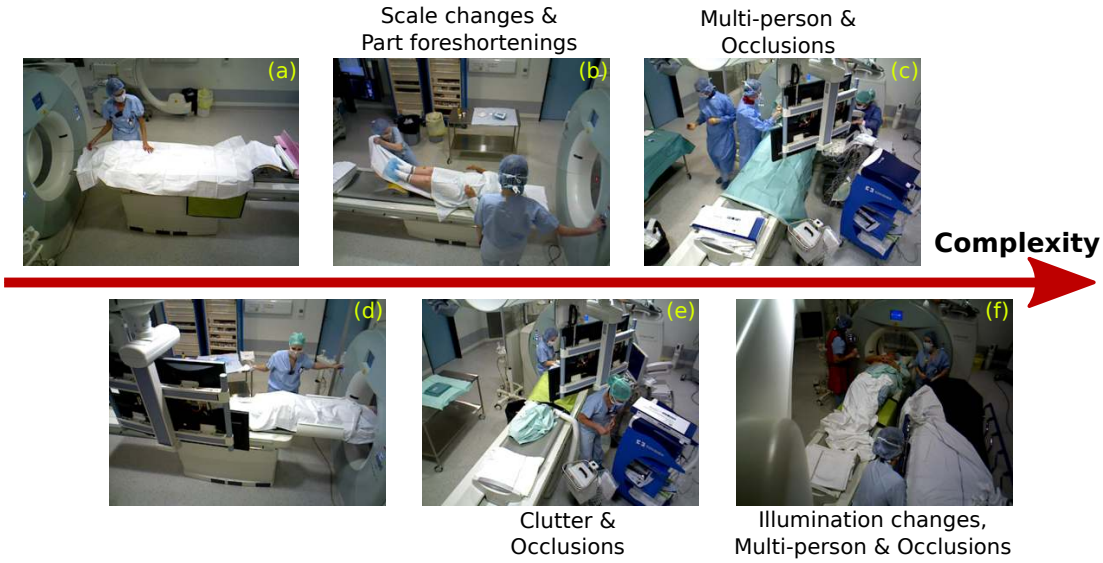


Figure 1.2: Sample images of an operating room showing persons during real surgeries and illustrating some of the challenges for clinician detection and pose estimation. The images are ordered from left to right based on visual complexity.

positions. These requirements make such systems unsuitable for localizing body parts of medical staff in the OR because they can not only be unreliable, but also disrupt the clinical workflow. Cameras are currently the most practical option for perceiving operating rooms during *real surgeries*, mainly due to two reasons. First, a camera recording system is one of the least intrusive sensing systems that can be integrated into an operating room. Second, such a system needs to be installed only once. Furthermore, these sensors provide a rich source of information going beyond the localization of humans. They offer the possibility to detect not only clinicians but also their activities. This information can be used for many applications, such as post-operative video review. This can permit to develop smart video browsing applications, for example by listing the time intervals where a specific number of persons is present in the room, and by identifying the time intervals during which clinicians or staff are working on the patient.

1.1.3 Visual Challenges in the Operating Room

The interpretation of the data provided by cameras is however not straightforward. In general, visual human detection and pose estimation are challenging tasks that become even more challenging in the OR. To describe some of the challenges, let us look at Figure 1.2 that shows sample images recorded during real surgeries. These images are ordered based on their complexity from left to right. In these images, clinicians and clinical staff are wearing **loose and textureless clothes** (a-f) that are very similar to the materials used to cover the other surfaces in the room. People can also appear in **various poses** (a-f). In addition, body parts might be occluded due to **self or object occlusions** (e). Furthermore, the appearance and dimension of a body part severely

Figure 1.3: A pair of color and depth images captured using an *Asus Xtion Pro* camera.

change depending on the body shape, imaging condition, viewpoint and pose. Due to **large appearance changes** (c and f) and **clutter in the environment** (e and f), it is very difficult to determine which image region belongs to the body parts of a person and which region belongs to the background. Finally, it is even more challenging to determine body part regions in **multi-person scenarios** (b, c and f) where the sizes of the persons in the image may be significantly different due to **perspective projections** (c, b and f). To perform visual clinician detection and pose estimation in unconstrained OR environments, we need to get many things right.

1.1.4 Our Camera Setup

In order to cope with the aforementioned challenges proper to the operating room, we propose to capture the environment using recently introduced low-cost RGB-D cameras. This type of camera allows to record an environment using both color and depth sensors simultaneously. A color image represents the color intensity on the objects' surfaces while a depth image encodes surface distances with respect to the sensor. On the one hand, the color sensor is sensitive to light in the visible wavelength range. On the other hand, the depth sensor works in the infrared light range that is invisible to human eyes. In cameras like the *Asus Xtion Pro*, the depth sensor decodes a predefined pattern that has been projected onto a scene in infrared light to compute distances between the surfaces in the scene and the sensor itself. Therefore, the lack of texture on the surfaces, color similarities between the surfaces or illumination changes do not affect the depth image computation. It is worth noting that due to technical limitations the depth resolution degrades dramatically above 5 meter distance from the sensor and the sensor fails on reflective surfaces. Therefore, the sensors have been mainly used in indoor environments.

In Figure 1.3, we have shown a pair of color and depth images that have been recorded using an RGB-D sensor in an operating room. All personnel wear sterilized clothes and gowns that cover the whole body. These clothes are often in green or blue colors without any visible texture. In this figure, one can notice that the color image provides

a good overall representation of the room. But, it is difficult to distinguish different surfaces with similar colors that are close to each other in the projected image, *e.g.* the arms. In contrast, the corresponding depth image provides a more visually discriminative representation for these surfaces.

In order to collect data during real surgeries, we have installed camera recording systems in several operating rooms in the course of this project. To benefit from the complementary information provided by the two inherently different sensors, we have chosen to use RGB-D cameras. We have selected *Asus Xtion Pro* cameras due to their compact size, lightweight, and connection with a single cable (powered USB). These camera systems have been used for recording all activities taking place in the room. This enabled us to collect data covering a wide range of potential visual challenges that exist in such environments. Figure 1.4 shows our camera setup in an operating room from the hospital of Strasbourg. We have indicated two sets of cameras with red and yellow boxes. The red box indicates an RGB camera that has been installed by the hospital to capture activities taking place on the bed close to the CT scanner tunnel. The other set of cameras shows our multi-view system, installed to capture the working environment in the room from three different viewpoints. Our multi-view camera system is fully calibrated using a method similar to [Ladikos 2010, Loy Rodas 2015].

In general, to install such a system in the OR, we need to take few elements into account. First, because of the sterilization process, we cannot pass cables over the floor and walls of the OR. Second, the cameras should be mounted on the ceiling to have a better coverage of the environment and also to reduce the risk that one or more views become occluded due to displacements of devices in the room. Finally, the cameras should be placed in such a way that they do not collide with ceiling-mounted articulated arms.

1.2 Applications of Clinician Detection and Pose Estimation

People detection and pose estimation in the OR can benefit various existing applications by providing them with location information about the persons in the room. This information is also required in order to develop and integrate new applications into the surgical workflow. In the following, we present two existing applications as well as two new potential ones and explain how clinician detection and pose estimation could help.

1.2.1 Context-aware Systems

Performing a surgery is a detailed and complicated process that proceeds in progressive stages. Context-aware systems aim at retrieving the OR context to build a full model of the process and at using this model to track and analyze the progress of an ongoing surgical operation. This awareness of the working context inside the operating room could be used to automatically provide the surgical team with the important clinical information, *e.g.* medical images and medical records, at the appropriate moment. The context model

Figure 1.4: A panoramic view of an operating room at *NHC Strasbourg, Interventional Radiology Department*. Camera positions are highlighted in yellow and red. Yellow boxes indicate the positions of the RGB-D cameras that we have installed in the room to capture the working environment from three viewpoints. The other camera is an RGB camera that has been installed in the room for archiving. It focuses on the bed close to the CT scanner device.

can also be used to generate a detailed medical transcription of the procedure. More importantly, such a system can enable the identification of critical situations in order to automatically give warnings or provide extra information for reducing medical errors. For example, according to the Institute of Medicine's report on human error, the occurrences of preventable medical errors in operating rooms result in the loss of tens of thousands of human lives in the USA per year [Makary 2016, Kohn 2000]. According to [Makary 2016] medical errors is the third leading cause of death in the USA, which means that the likelihood of an American to die from a medical error is higher than a death due to car accidents or HIV infections.

During the last decade, a significant effort in the Computer Assisted Intervention (CAI) community has focused on designing methods and algorithms in order to equip the OR with context-aware computer-assisted systems [Lalys 2014]. Proposed methods are targeting different components required for such systems, such as surgical tool detection and tracking [Bouget 2015, Kumar 2015], surgical phase and activity recognition [Padoy 2009, Padoy 2012, Twinanda 2015, Tran 2016] and a combination thereof [Twinanda 2016]. Recently, [Nara 2011, Agarwal 2007, Meißner 2014, Bardram 2011] have investigated the problem of reconstructing the operating room's context by using different types of data, *e.g.* patient vital signals, tool detection information and the location information of the personnel. These studies have shown that since the main

actors in the OR environment are the clinicians and clinical staff, monitoring their postures and the interactions among them are highly important for building the OR context. Clinician detection and pose estimation are necessary to supply context-aware systems with this essential information during live surgeries.

1.2.2 Surgical Skill Assessment

Designing effective methods for teaching and assessing surgical skills is crucial for hospitals to guarantee the quality of care [Vedula 2016]. Trainees are generally learning the skills in the classrooms and acquiring them through observing experts and reproducing the gestures and movements on the limited number of cases available for resident training. In [Wanzel 2002], a study on junior surgical resident training confirms that direct supervision and personalized feedback by experts significantly improve the performance of the residents. This study suggests that it is highly important to supervise every gesture and movement performed by the novice in the course of a surgery in order to boost safety and efficacy. On the one hand, following such a procedure is not practical because of the huge costs that would be imposed to hospitals and the huge increase in the expert surgeons' workloads. On the other hand, considering the specific characteristics of the problem such as repeatability and therefore the presence of many examples performed by experts, it is clear that the problem is well-suited for computer-based analysis. Automatic skill analysis can provide a great tool in the hands of the health care system to ensure the optimal utilization of resources and reduce costs by eliminating the need for extensive monitoring by expert surgeons. Such systems can also facilitate surgical skill training by providing trainees with more opportunities to practice.

As a result, many computer-based approaches have been proposed for surgical skill evaluation by comparing experts and trainees while performing similar tasks. Surgical skill evaluation can be performed by comparing trajectories obtained by marker-based tool and body part tracking [Meißner 2014] or using time to task completion during robotic surgeries [Judkins 2008, Vedula 2016]. Reiley et al. [Reiley 2011] presented a comparative review on surgical skill assessment methods and confirmed the common hypothesis that skill lies in the interrelation and arrangement of body movements. In general, current automatic skill assessment approaches have several limitations: 1) since the tracking systems require attaching markers to the bodies of the subjects, especially to the hands, it could affect their performance; 2) even though expert surgeons are performing many real procedures which cover all real case scenarios, current skill analysis systems cannot benefit from these examples due to the difficulty of installing tracking systems in the OR; and 3) more importantly, such systems cannot be used to evaluate surgical skills during real surgeries. Human pose estimation in the OR can serve as a fundamental step towards addressing these limitations by removing the need for marker-based body part tracking and enabling data collection from real surgeries.

1.2.3 Radiation Safety Monitoring

In addition to the great benefit of clinician detection and pose estimation for a wide range existing applications, clinician detection and pose estimation could open up possibilities for new applications, such as radiation safety monitoring and intra-operative human-robot collaboration that are presented below.

One of the greatest medical breakthroughs of the modern era is Minimally Invasive Surgery (MIS). MIS is performed by accessing internal organs via small incisions. Several types of MIS are relying on intra-operative x-ray based imaging to guide, control and monitor the tools inserted in the patient during a procedure. But, a major disadvantage of x-ray based procedures is that ionizing radiation is delivered both to the patient and to the clinicians. Since the clinicians regularly use the device, the dose they receive can be high.

To protect operating room personnel during x-ray guided surgeries, the personnel should use lead protections. However, many parts of the body, such as the arms and the head, are usually left unprotected. In practice, the amount of doses received by the unprotected parts can sometimes exceed the predefined dose limit, which can lead to serious negative effects on the body, for example cancer in the extreme [Vanhavere 2008]. Even though the personnel are currently wearing a dosimeter at chest level to measure the radiation exposure, a recent large consortium-based study on radiation exposure of OR personnel shows that it can vary significantly at different locations of the body [Carinou 2011]. Therefore, to compute accurate estimations of the radiation exposure of the medical staff, the exposure should be measured at different body locations, *e.g.* the head, the arms and the hands. Since relying on dosimeters would require to put a multitude of dosimeters over the bodies of the clinicians, dosimeters are not a practical option. Therefore, a noninvasive system is required to monitor radiation exposure of the medical staff in the OR. One practical option would be to use a radiation simulation system such as [Ladikos 2010, Loy Rodas 2015] along with a vision-based body part localization model. We should note that such a body part localization model needs not only to localize body parts but also to consistently track the parts in order to correctly compute the accumulation of the exposure over time per person and per body part. In addition, such a system could be used to alert the medical staff during a procedure and also to perform postoperative studies for finding the correlation between the risk and different actions as well as poses performed in the course of a surgery.

1.2.4 Human-robot Collaboration

The current medical robotics research focuses more and more on designing semi-autonomous robotic systems. These systems therefore need to include contextual information in order to guarantee safe and efficient collaboration and to reduce human supervision as much as possible. In other words, the system should be as intuitive as possible to behave as expected, yet ensure a safe usage while used in parallel or serial with other robots or human users. In order to enable safe and effective human-robot collaborations, the robot

needs to not only build a precise model of its workspace, but also to localize persons and their body parts in the workspace [Lasota 2014, Michalos 2015]. Consequently, it is required to localize persons and their body parts in the OR in order to allow the usage of semi-autonomous robotic systems in the room [Beyl 2015, Ladikos 2008].

One example of such a system in the OR is the robotized C-arm. Robotized C-arms can be used to obtain either 2D images or 3D scans of different body organs. The device constructs the 3D images by automatically performing a series of rotation and angular movements around the body of the patient laid on the bed. Therefore, it is important to make sure that medical staff and any other equipment do not enter the working volume of the device to avoid potential collisions. A collision could hurt the colliding person and damage the device. To avoid such collisions, 3D localizations of OR personnel and their body parts are thus required for sending notifications and stopping the device if someone accidentally enters the device's working volume [Ladikos 2008].

1.3 Performance of State-of-the-art Methods in the Operating Room

As already discussed in Section 1.1, in order to be able to conveniently deploy our system in the OR, we need to rely only on camera sensors for localizing the medical staff and their body parts. Vision-based human detection and pose estimation are key problems in computer vision, which have generated extensive literature in recent years. In general, the proposed approaches are either *part-based* or *holistic*. In part-based approaches, the human body is represented by a set of body parts and pose estimation is performed in two steps: first, by detecting potential body parts in an image; then by finding a collection (or collections) of body parts via verifying mutual spatial constraints among the parts. On the other hand, holistic approaches are directly mapping an image representation into person and body part positions. By having a mapping function that can access the entire image, holistic approaches are capable of learning powerful mapping models that can exploit the whole image context. However, such mapping models have a lot of parameters, which requires a very large training set to learn them. They also do not model interpart dependencies explicitly, which can lead to inconsistency in the predicted body configuration. Instead, part-based approaches need a much smaller training set and also provide a powerful formalism to explicitly model dependencies between body parts. As it is difficult to obtain lots of annotated data in the OR, and also because modeling interpart dependencies is essential due to the challenges mentioned in Section 1.1.3, we decide to base our work on part-based approaches.

An elaborated review on vision-based human detection and pose estimation is presented in Chapter 2. In the following, we briefly introduce three dominant state-of-the-art human pose estimation approaches in order to show their limitations in an environment like the OR. These approaches are Flexible Mixtures of Parts (FMP) [Yang 2013], *DeeperCut* [Insafutdinov 2016] and Kinect skeleton tracker [Shotton 2012]. The first two are part-based, and the last one is holistic. We consider the FMP approach because of

two reasons. First, it is among the top-performing approaches for different challenging datasets. Second, the approach serves as a basis to develop our clinician detection and pose estimation approach. We evaluate DeeperCut since it is a state-of-the-art approach for multi-person pose estimation. In addition, the Kinect skeleton tracker is evaluated since it is currently the most successful commercial product for vision-based human pose estimation.

1.3.1 Evaluated Approaches

Flexible mixtures of parts. The Flexible Mixtures of Parts (FMP) approach [Yang 2013] is based on the Pictorial Structures (PSs) framework that is the dominant part-based approach. FMP preserves the efficiency of the PSs model and extends the model to be robust to the foreshortening of the body parts and also to changes in the appearance of the parts. The FMP approach learns all model parameters automatically and jointly. This permits the approach to elegantly use all the training data for obtaining a reliable pose estimation model. Because of the efficiency of the FMP model, this model can be used for both human detection and pose estimation.

DeeperCut. The DeeperCut method [Insafutdinov 2016] is a very recent part-based approach for articulated human pose estimation in scenes with multiple persons. A very deep Convolutional Network (ConvNet) is used to generate a set of body part detections. Then, an elegant objective function is defined to jointly estimate the poses of all people in the image. The objective function relies on a set of constraints defined between every pair of detection candidates in order to jointly partition and label these candidates into disjoint sets of body part configurations corresponding to individual persons in the scene. These constraints are defined based on the appearance and 2D location of the detected candidates. The objective function is optimized using a branch-and-cut algorithm.

Kinect skeleton tracker. With the introduction of low-cost RGB-D sensors, Shotton et al. [Shotton 2012] proposed a holistic approach to estimate human poses on depth images. The approach assumes that the foreground is segmented. Then, a Random Forest (RF) with deep trees is used to localize body joints. The Kinect skeleton tracker uses a commercial and extended version of this approach, which is provided with Microsoft Kinect One sensors. The skeleton tracker shows very promising results in indoor scenes such as living rooms.

1.3.2 Qualitative Results

In this section, we show the qualitative results of the aforementioned approaches on OR data. These results are generated using models trained on standard computer vision datasets in order to evaluate the generalization of these models to the OR.

To evaluate FMP, we have learned an FMP model using the public implementation of [Yang 2013] provided by the authors. The model is trained on the Buffy dataset [Eichner 2012b] that has been frequently used in the computer vision community and

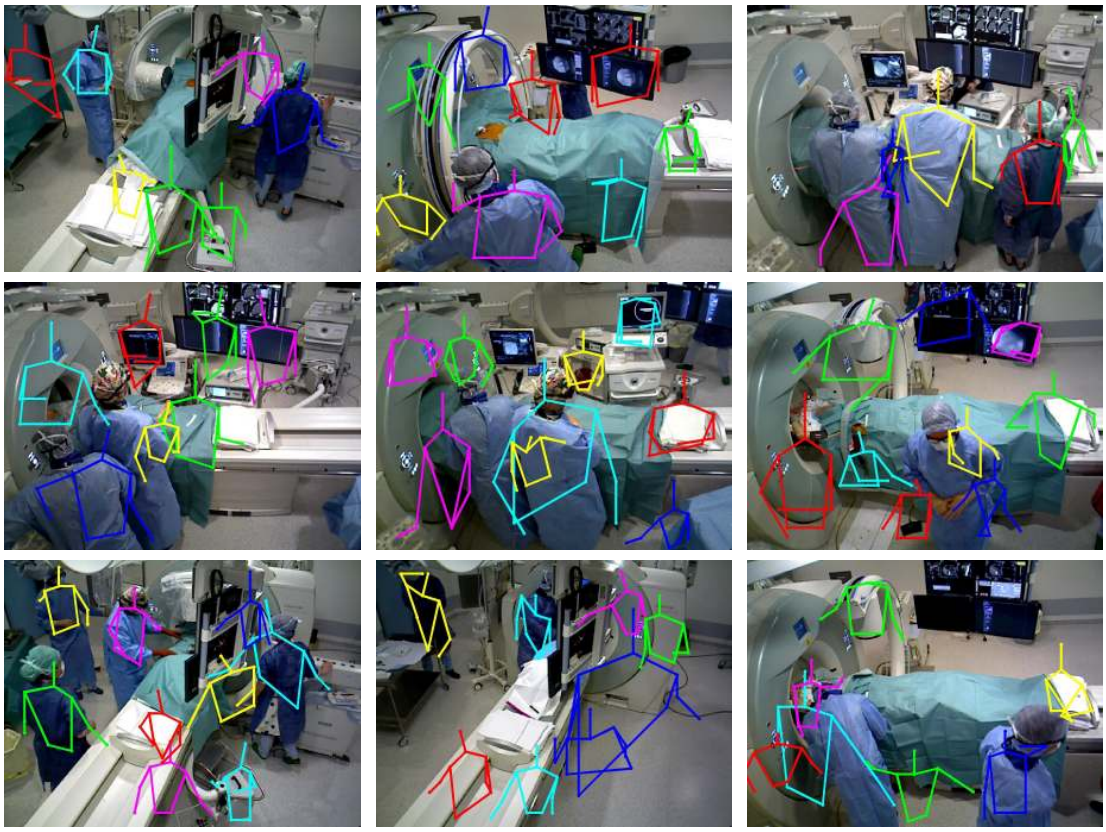


Figure 1.5: Qualitative evaluation results using flexible mixtures of parts [Yang 2013]. We have trained an FMP model on the Buffy dataset. We have randomly selected the test frames from several videos recorded during real surgeries.

in [Yang 2013] as well for both training and evaluating pose estimation methods. For DeeperCut, we use the publicly available model that has been provided by the authors [Insafutdinov 2016]. We use the Kinect skeleton tracker that is provided with the Microsoft Kinect cameras. Note that the Kinect tracker relies on depth and temporal information while the two other models rely on a color image and do not use any temporal information.

We randomly select a set of frames from several videos that have been recorded during real surgeries performed in the operating room shown in Figure 1.4. Qualitative results of the FMP and DeeperCut models on the selected frames are shown in Figure 1.5 and Figure 1.6, respectively. Since the Kinect skeleton tracker cannot be evaluated off-line, we have installed a Kinect camera in the same room and recorded the tracker during real surgeries. A set of frames extracted from these recordings are shown in Figure 1.7. These frames are extracted from video recordings after the first hundred frames to account for the time required by the tracker to be initialized[1].

[1]More examples are available at https://www.youtube.com/watch?v=iabbGSqRSgE

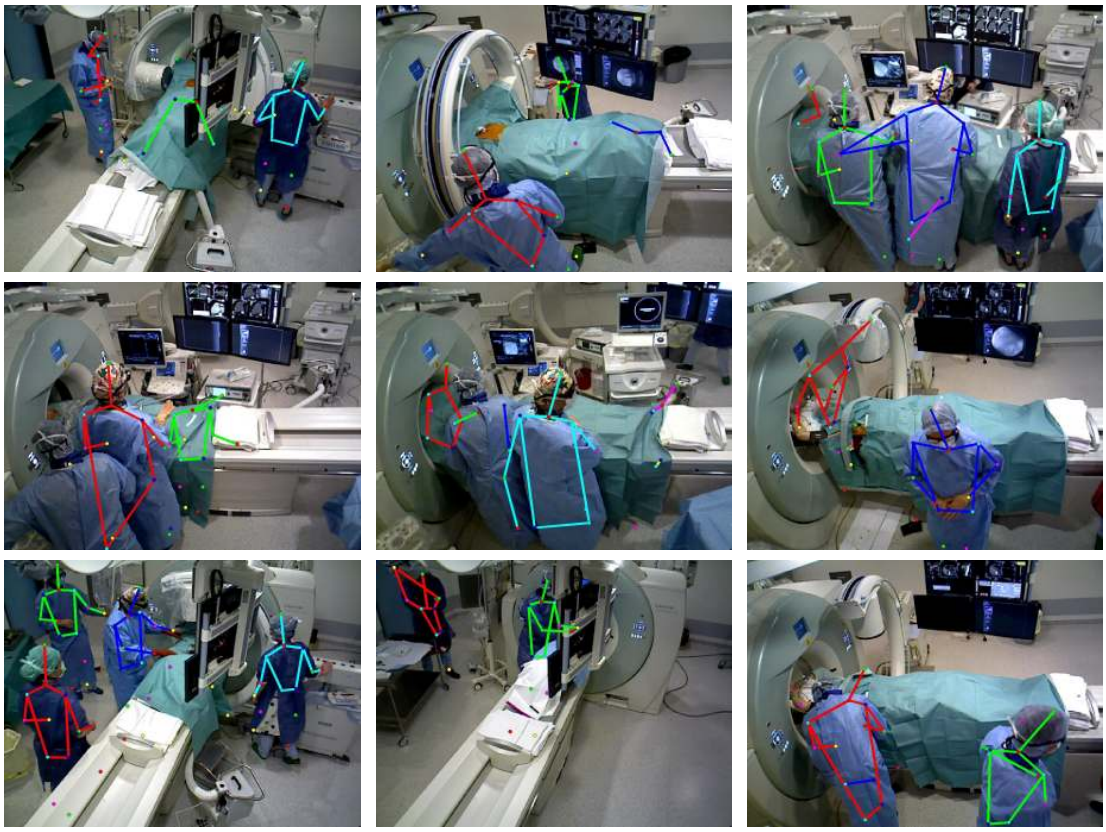


Figure 1.6: Qualitative evaluation results using the DeeperCut model that was made publicly available by the authors [Insafutdinov 2016].

1.3.3 Discussion

One can notice that the FMP model has generated many false positives and that in spite of all the false detections, there still exist persons without any detection in all the frames. In general, the estimated body poses are often inaccurate because body parts are confused among different persons and false body part detections are present on the background. The DeeperCut model performs much better than FMP and is less subjective to false positives. It can however be seen that there still exist undetected persons and also that the estimated skeletons are not always reliable. We believe that the low performance of FMP and DeeperCut is due to the presence of many surfaces with similar colors and to loose as well as textureless clinical clothes, which make the part detection and the appearance-based pairwise constraints unreliable. Even though, the FMP and DeeperCut methods have achieved impressive results on challenging dataset recorded in common indoor and outdoor scenes, these results on OR data indicate that they do not necessarily generalize well to such environments. Quantitative evaluation results for FMP and DeeperCut approaches are provided in the following chapters.

In general, the Kinect skeleton tracker performs reasonably well in cases where clinicians are facing the camera and located far from other persons and objects in the

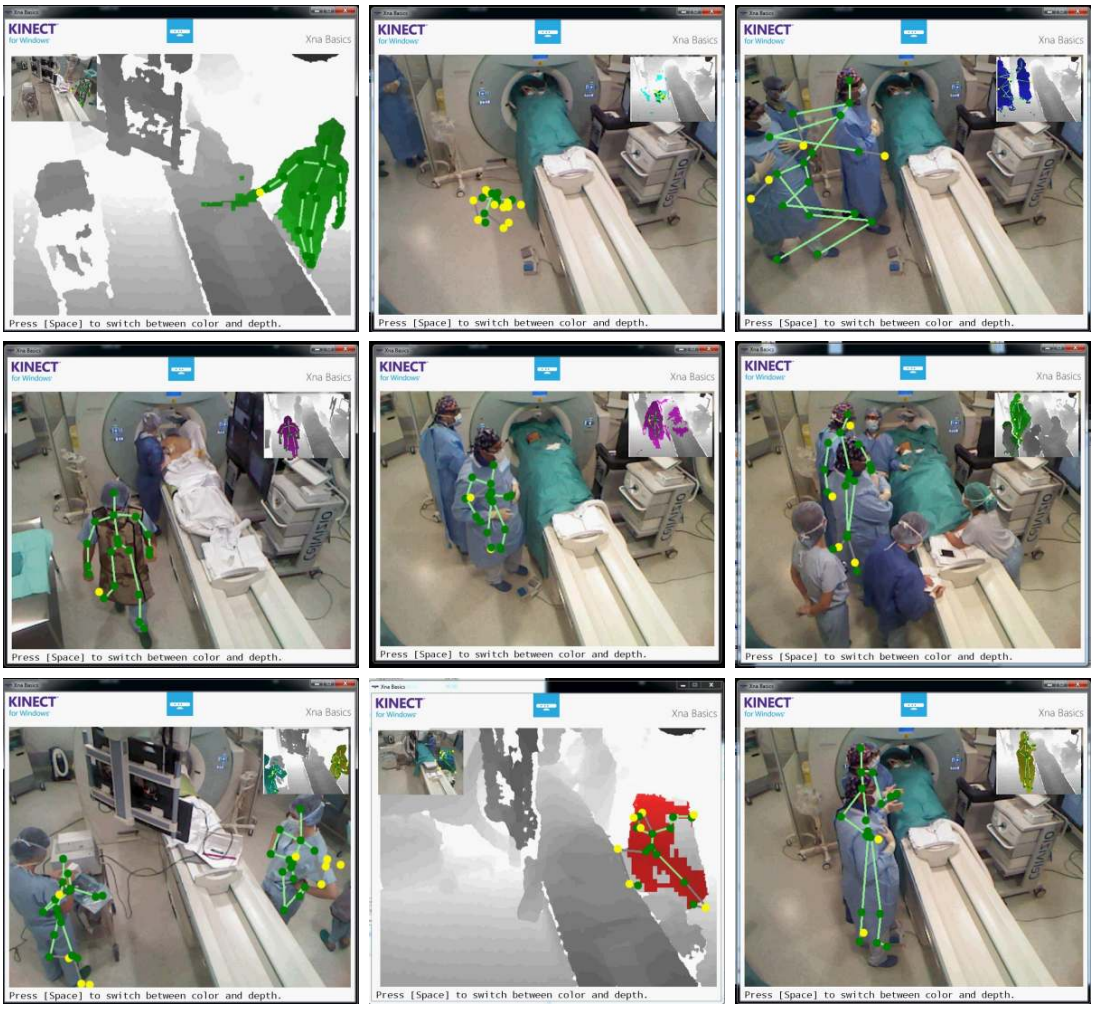


Figure 1.7: Examples of pose estimation results from Kinect skeleton tracker [Shotton 2012]. These frames are extracted from video recordings of the Kinect skeleton tracker evaluated in an operating room. We have extracted the frames after the first hundred frames to make sure that the tracker is initialized. The segmented foregrounds are indicated by different colors on the depth images. The estimated skeletons are overlaid on the color images. (Picture best seen in color)

room (the frames in the left-most column of Figure 1.7). But, it should be noted that even though arms are well localized in those cases, hip and lower body part localizations are inaccurate. This is due to loose clinical clothes that cover whole body, which make it difficult to discriminate between torso and lower body. The tracker often misses clinicians and mixes the body parts of different persons. We believe that this low performance is due to: *(1) cluttered scenes* that make the tracker as well as foreground estimation fail; and *(2) very different appearances of persons compared to the ones used during random forest training*, which are due to both the special clinical clothes and the top view of the camera that has to be mounted on the ceiling. In order to address the latter one, one

would need to retrain the forest, which is a challenging task due to the large training set that is required.

1.4 Contributions

This thesis makes several contributions in two main areas: 1) general human detection and pose estimation using RGB-D data; and 2) OR video analysis for applications such as smart post-operative review.

We propose three novel part-based approaches for human pose estimation. Two approaches are proposed to tackle the problem of human pose estimation using single-view RGB-D images. The third approach is a multi-view approach to recover 3D human poses by incorporating evidence across all views.

In the first approach, we propose a new method to estimate and track the upper-body poses of clinicians in RGB-D sequences. In order to provide a temporally consistent tracking of body parts over a sequence, which is required for applications such as radiation safety monitoring, we present a novel method that relies on discrete Markov Random Field (MRF) optimization. We propose an energy function that includes part detection likelihood through a unary potential and enforces dependency constraints between body parts using two pairwise potentials: kinematic and temporal. The kinematic potential enforces body physical constraints. The temporal potential is used to enforce temporal smoothness between body parts in every consecutive frame. In order to consistently track poses over the entire sequence and also to cope with the failures of the part detector in such a complex OR environment, we optimize the proposed energy function over the complete sequence.

Our second contribution is the design of a robust model to leverage a pair of registered color and depth images for performing human detection and pose estimation in visually challenging environments such as operating rooms. Our proposed approach, referred to as *3DPS*, extends pictorial structures to RGB-D data in three ways: 1) by building a more robust and discriminative part detector that relies on both color and depth images; 2) by constructing a more realistic deformation model that constellates parts based on their 3D distances in order to resolve the inherent ambiguity of projection from 3D to 2D; and 3) by proposing an efficient algorithm to reduce the size of the 3D state space to make exact inference tractable. We also present a novel feature descriptor for depth images. The descriptor encodes relative surface depth changes in a depth invariant representation. We have evaluated the approach on several datasets recorded during real surgeries and shown that the model generalizes well.

As a third contribution, we tackle the tasks of multiple human detection and pose estimation in multi-view setups. The approach relies on a set of images recorded synchronously from three RGB-D cameras to jointly detect and estimate body poses. The multi-view multi-person pose estimation is carried out in two steps by: firstly, detecting and generating skeleton candidates in each view; and secondly merging the skeletons across views and refining them in 3D. We also extend our 3DPS approach to use a deep ConvNet-

based body part detector on RGB-D images to be less subjective to false detections that can mislead the subsequent multi-view merging and optimization methods. In addition, we propose a random forest based prior to automatically model a priori information about the environment and to filter spurious skeleton candidates. A novel multi-view energy function is introduced to update 3D part positions based on multi-view cues, which has a number of appealing properties: (1) using depth information to efficiently and reliably estimate correspondences across views, (2) estimating reprojection costs based on depth instead of appearance similarity, which is unreliable in OR environments and (3) iteratively optimizing the multi-view energy function to efficiently explore a large 3D space.

As a fourth contribution, we have generated various challenging RGB-D datasets from several days of recordings during real surgeries (see Appendix A). These datasets have been manually annotated with ground-truth person bounding boxes and body part locations. These datasets allow us to truly evaluate and analyze the performance of our proposed approaches for the tasks of detecting and estimating poses of medical staff on real OR data. To the best of our knowledge, this is the first work that presents an evaluation on data recorded during real surgeries.

As another area of contribution of this thesis, we have used our approach to study the room usage and moving patterns of surgical team members during different procedures. We have also presented how our approach could be useful for several other applications such as estimating the accumulation of radiation exposure par body part.

1.5 Outline

In this section, we briefly outline the structure of this dissertation.

Chapter 2 reviews the literature on human detection and on human pose estimation. We describe the most prominent approaches related to our work and highlight commonalities and differences focusing on both methodology and application.

Chapter 3 presents an overview of graphical models that serve as bases to build the proposed method. We briefly describe the tree-structured as well as loopy graphical models and inference algorithms that can be used in each case. As graphical models are an abundant topic, we only focus on specific models that have been used to develop our approaches.

Chapter 4 presents our approach based on discrete Markov random field for tracking clinician poses on single-view RGB-D sequences. This work has been published in [Kadkhodamohammadi 2014].

Chapter 5 introduces our novel 3D pictorial structures approach on color and depth images for human pose estimation. We also present a new descriptor on depth images that encodes local depth level changes in a multi-scale representation. Work related to this chapter has been presented in [Kadkhodamohammadi 2015,Kadkhodamohammadi 2017a].

Chapter 6 presents a multi-view approach to estimate clinician poses in operating rooms. This approach is based on the single-view 3D pictorial structures method

presented in Chapter 5. Work related to this approach has been published in [Kadkhodamohammadi 2017b].

Chapter 7 investigates the use of our approach from Chapter 6 for several applications in operating rooms.

Chapter 8 finally concludes this work and describes possible directions for future work.

2 Related Work

Chapter Summary

Human detection and pose estimation are important research objectives that have been investigated in depth during the last few decades. This is mainly because of the valuable information that humans detection and pose estimation can provide for a wide range of applications, for instance, visual surveillance, activity recognition and human-computer interaction. *Human detection* is defined as the process of localizing people with 2D bounding boxes or 3D bounding cubes depending on the type of data that is available. *Human pose estimation* is the process of estimating the configuration of the body by recovering the body part positions in 2D or 3D. This process is also referred to as body configuration recovery.

From a modeling perspective, proposed approaches for human detection and pose estimation can be categorized into two main categories:

a. **Part-based approaches.** These approaches break down the problem of pose estimation into a set of body detectors along with an inter-part dependency model.

b. **Holistic approaches.** Holistic approaches attempt instead to learn models that

directly predict body part positions. In other words, these approaches are learning a direct mapping from image pixels values or image features into person and body part positions.

In terms of input data, existing vision-based methods can be divided into *single-view* or *multi-view* approaches. Multi-view approaches are based on a set of images that have been captured from different viewpoints at the same time, while single-view approaches are only relying on input data from a single camera. Multi-view approaches have fewer problems with occlusions, but this is achieved at the cost of a more complex system in order to synchronize and calibrate the cameras. From another perspective, the proposed methods can rely either on input data recorded at a single time step or on a sequence of images recorded over time. If the body motion model is known, temporal information can be extracted from the sequence of images to resolve detection ambiguities.

We begin this chapter by reviewing in Section 2.1 the state-of-the-art vision-based methods for human detection and pose estimation. We then proceed by presenting approaches that have been proposed for clinician detection and pose estimation in ORs. This review of OR approaches will also present methods relying on other types of sensors.

We follow the same terminology as above and highlight different aspects of the methods from both the modeling perspective and input data.

2.1 Computer Vision Methods for Human Detection and Pose Estimation

In order to address many challenging problems associated with visual human detection and pose estimation such as severe appearance changes (due to illumination variations, occlusions and cluttered background) and the many degrees of freedom of the body parts, human detection and pose estimation have been actively researched over the years. We review RGB methods in Sections 2.1.1 and 2.1.2 depending on the number of views that are used. We then present in Section 2.1.3 human detection and pose estimation methods relying on RGB-D images since our proposed approaches also rely on RGB-D images.

2.1.1 Single-view Approaches

This section investigates different single-view approaches that have been proposed for tackling human detection and human pose estimation jointly or separately.

2.1.1.1 Human Detection

In recent years, most of the approaches for human detection have focused on pedestrian detection. Pedestrian detection has the potential to greatly impact the quality of life through several applications in automotive safety and robotic navigation [Dollar 2012, Benenson 2014]. However, one should note that pedestrian detection is more constrained compared to person detection in general. Pedestrians are always appearing in upright orientation and exhibit more regularities in pose, which make their detection more

tractable. Pedestrian detection methods however serve as bases for most of the proposed human detection and pose estimation approaches. We therefore proceed by first reviewing the literature on pedestrian detection methods and then present approaches for generic people detection.

Pedestrian detection. Early progress on pedestrian detection was made in [Gavrila 1999, Gavrila 2000]. The authors propose exemplar-based methods that compute image edges and then compare the edges with a hierarchy of silhouette exemplars. But, these approaches are not applicable to realistic environments due to the highly cluttered backgrounds and a limited number of training examples that cannot cover all possible variations in pedestrian silhouettes. Over the last decade, great improvements have been achieved in pedestrian detection from single monocular images using different approaches that can be divided into three main groups:

- Sliding window detectors: These approaches scan the image using a detection window or multiple windows over all positions as well as scales. Then, each window is represented by a feature vector and classified independently in order to predict the absence or presence of a pedestrian. Different representations have been used to compute the feature vector, for example, Haar wavelet [Oren 1997, Papageorgiou 1999], Histogram of Oriented Gradients (HOG) [Dalal 2005] and HOG combined with motion features [Dalal 2006]. The prediction is often performed using the Support Vector Machine (SVM) classifier [Vapnik 1995].

- Part-based models: These methods represent the person as a set of rigid parts in combination with their spatial relationships. The parts are defined in different ways: by relying on body part annotations [Andriluka 2009] or by mining automatically parts based on their discriminative properties [Felzenszwalb 2010, Bourdev 2010];

- Holistic approaches: With the large increase in datasets' sizes and computing power, approaches are proposed to directly regress pedestrian positions from input images. These methods are based on deep convolutional neural networks [Sermanet 2013, Ouyang 2012, Ouyang 2013] and decision trees [Dollar 2009, Benenson 2013].

For a comparative study of recent pedestrian detection approaches on public datasets, the reader is referred to the reviews by [Dollar 2012] and [Benenson 2014].

Generic people detection. In contrast to the vast literature on pedestrian detection, fewer approaches are addressing the problem of generic people detection in unconstrained environments. This is mainly due to the large pose variabilities that people can go through compared to pedestrians. Since variations in poses change the appearance of a person dramatically, approaches have been proposed to tackle both human pose estimation and human detection tasks jointly. These approaches performing pose estimation are discussed in Section 2.1.1.2.

Recently, Felzenszwalb et al. [Felzenszwalb 2010] have proposed a part-based method to detect people in still images. The approach represents a human using a Deformable

Part Model (DPM) that is characterized by an ensemble of rigid parts[1] and their spatial displacements. Bounding box annotations are used to learn a set of filters for the main silhouette from HOG-based appearance representations. These filters are called root filters. Then, a discriminative approach is used to automatically discover parts based on the root filters. At test time, the root and part filters are deployed in a sliding window scheme to compute part scores for every possible image position. The person likelihood is then computed based on the scores and on an inter-part displacement model learned from training data. The automatic part selection paradigm makes the approach suitable for detecting different objects without the need for human experts to specify the parts. This part-based representation also enables the approach to be robust to pose changes. But, the approach does not explicitly consider occlusions and view-point changes that can make some of the parts invisible.

In order to automatically mine parts by considering not only their appearance but also their spatial configurations (*i.e.* body pose), an approach is proposed in [Bourdev 2009], which relies on 3D annotations of body keypoints to construct parts. Each body annotation includes 19 keypoints (full-body joints, ears, eyes and nose). The 3D annotations and the appearance of people in the training set are mapped into a manifold in which people with similar appearance and body configurations are close to each other. The parts, that are called *poselets*, are constructed by performing clustering in this manifold. Finally, the generalized Hough transform is used to combine poselets' scores and localize the person. Both DPM and poselets are using a discriminative approach to discover the parts. Both approaches are also relying on HOG-based representations and SVM to build part appearance models. DPM uses less than ten parts while poselets uses few hundred parts. The poselets approach encodes occlusions and view-point changes by relying on 3D keypoint annotations. However, the detailed annotation of 3D body pose is a very demanding task, especially on real monocular images. The poselets approach has been extended by using 2D body pose annotations on a large training set [Bourdev 2010], learning spatial relationship between poselets [Gkioxari 2014] and using deep convolutional networks [Bourdev 2014].

2.1.1.2 Human Pose Estimation

Here we start by reviewing single-view human pose estimation methods based on the part-based framework that is central to the work presented in this thesis. We also briefly discuss holistic approaches.

Part-based approaches. The basic idea behind part-based models is to look for an assembly of body parts that is feasible according to body physical constraints and that the best fits image observations. The key ideas of the part-based model were originally presented in [Fischler 1973]. Fischler and Elschlager proposed the Pictorial Structures (PSs) approach that represents an object using a set of rigid parts and their spatial layout constraints. They have discussed that the loopy dependency between parts makes exact

[1]Note that a part is an abstract representation and does not necessarily correspond to a specific body part. A part can represent a set of body parts or even the whole body.

inference intractable. In their case, where a loopy model is used, a heuristic approach is proposed, which does not guarantee to discover the optimal solution.

Felzenswalb and Huttenlocher presented in [Felzenszwalb 2005] a computationally efficient framework for the part-based modeling and detection of objects. This framework is based on pictorial structures and uses a tree-structured pairwise deformation model that encodes spatial displacement between parts. Both model learning and object detection are cast into a probabilistic interpretation. They proposed to discriminatively and separately learn part detectors, which are also often called appearance models. The interpart dependency model, *i.e.* the pairwise deformation model, was learned by a maximum likelihood formulation using ground-truth locations of body parts. The object matching is then performed based on *dynamic programing* and *belief propagation* [Bishop 2006, Felzenszwalb 2004]. In [Felzenszwalb 2004], a linear time inference algorithm based on the generalized distance transform is proposed to find the optimal solution for tree-structured pictorial structures. As a result, a wide variety of pictorial structures based methods have subsequently been developed.

In [Felzenszwalb 2005], the appearance model relies on simple part templates based on background subtraction. For that reason, the approach has only been evaluated on images recorded in a controlled laboratory environment with clean backgrounds. In order to address this shortcoming, Ramanan [Ramanan 2007] proposed to iteratively parse the input image for better features. In each iteration, a soft labeling of image pixels into region types such as background, torso, left arm, etc, is computed, which is referred to as a *parse* in the paper. The initial parse is obtained from an edge-based model. Then, color models are constructed for different regions and subsequently updated to learn better features tuned for each image. This approach has been extended by integrating temporal information and automatic segmentation [Ferrari 2008] and by learning appearance models that take inter-part dependencies into account [Eichner 2009]. However, occlusions and the presence of multiple persons in close proximity of each other can confuse these approaches.

Andriluka et al. [Andriluka 2009] have proposed to build appearance models based on a dense shape context representation [Mikolajczyk 2005] and an AdaBoost classifier. But, the approach does not explicitly model the occlusion or foreshortening of body parts. Yang and Ramanan [Yang 2013] proposed Flexible Mixtures of Parts (FMP) that encodes the body pose via a configuration of body joints instead of body parts to capture the foreshortening of the parts. The approach also learns multiple mixtures per part to handle appearance changes. The HOG appearance representation and SVM classifier are used to build part appearance models, which are commonly used in the literature as well [Yang 2013, Bourdev 2010, Gkioxari 2014].

Deep Convolutional Networks (ConvNets) have recently become popular for many vision-based tasks including human pose estimation [Insafutdinov 2016, Tompson 2015, Schmidhuber 2015]. Significant improvements are obtained by using deep ConvNet-based appearance models that are capable of learning strong feature representations and of including a wide image context through deep architectures [Chen 2014, Yang 2016]

In addition to the improvement of the appearance model, different approaches are presented to construct more robust deformation models using *fields of parts* that incorporates higher-order dependencies among body parts [Kiefel 2014], *adaptive pose priors* that automatically choose an image-dependent prior [Sapp 2010], *stretchable models* that incorporate motion, color as well as contours [Sapp 2011], repulsive factors between left and right body parts [Andriluka 2012a], or temporal consistency [Tokola 2013, Sapp 2011]. One should however note that these deformation models come with the penalty of approximate inference. Deep ConvNets are also used to construct image dependent pairwise terms between body parts for models with exact [Chen 2014] and approximate [Tompson 2014] inference. However, all these approaches are used in single person scenarios while approaches such as [Yang 2013] and [Andriluka 2012a] can be used to detect poses of several persons by sampling from the estimated pose posterior probability.

But, [Yang 2013] and [Andriluka 2012a] detect each person separately. The pictorial structures framework has recently been extended to incorporate context when several people in an image perform the same pose [Eichner 2012a] or complementary poses, for example in case of dancing couples [Andriluka 2012b]. In [Pishchulin 2016], a multi-person pose estimation approach is presented that makes no assumption about pose dependencies among different individuals in the image. The approach first generates a set of independent body part candidates and constructs a fully connected graph by connecting each part candidate with all the other candidates in the set. Then, the approach uses integer linear programing to jointly label each candidate with a unique body part label and uniquely assign them into different individuals. However, this optimization problem is NP-hard. The approximate inference algorithm takes about 72 hours to process a single image as reported in [Insafutdinov 2016]. [Insafutdinov 2016] builds on [Pishchulin 2016] and proposes to use a deep ConvNet-based body part detector and image dependent pairwise terms in conjunction with an incremental optimization method in order to speed-up the inference. For a recent survey on part-based human pose estimation, the reader is referred to the review presented in [Liu 2015].

Holistic approaches. Most recently, holistic approaches have become popular with the availability of large training set and of more computing power. The approaches are often using random forests [Shotton 2012] or ConvNets [Toshev 2014, Jain 2015, Wei 2016] to regress for body part locations in a holistic manner. These approaches are capable of learning strong feature representations, but, spatial body joint constraints are not explicitly modeled. These constraints are however crucial to guarantee joint consistency in the predicted body configuration. As a result, they may generate imperfect body pose predictions especially in multi-person scenarios. Therefore, the current trend in human pose estimation is to cast holistic-based approaches to the part-based paradigm to allow for the explicit modeling of dependency constraints between body joints [Buys 2013, Yang 2016, Tompson 2014, Chen 2014, Insafutdinov 2016].

2.1.2 Multi-view Approaches

Rather than using data captured from one viewpoint, multi-view approaches rely on a multi-camera system to capture a scene from different viewpoints. The multi-view systems offer a variety of benefits. For instance, they are less subject to occlusions and provide the possibility to recover the 3D layout of the scene. However, there still exist challenging problems that need to be addressed in order to fully benefit from these systems. For example, how can correspondences be established across views? How can occlusions and incorrect correspondences be distinguished? Furthermore, estimating a 3D pose in a 3D space can be an expensive task in terms of computation cost due to both the presence of more degrees of freedom for body parts in 3D and the larger size of the 3D state space. In the computer vision literature, various approaches have been explored to address the aforementioned problems in order to benefit from multi-view data.

Early work on multi-view human pose estimation focused on single person scenarios in controlled environments to reduce the ambiguity of data association [Sigal 2009]. [Gall 2010,Yao 2012,Stoll 2011] have addressed this task using 3D body models that require to perform complex inference in a high-dimensional space of 3D body configurations. Hofmann and Gavrila [Hofmann 2011] proposed an exemplar-based approach that recovers 3D pose exemplars per view and then relies on appearance consistency across views and temporal smoothness for predicting 3D body configurations. All these approaches are however relying on a constant background to reduce both false positives and the ambiguity of the 2D to 3D back-projection process.

In order to deal with dynamic backgrounds, [Burenius 2013] proposed to use 3D pictorial structures. But, to cope with the high complexity of exact inference in 3D and its memory requirements, the approach uses a very coarse discretization of the 3D space and binary pairwise constraints, which in turn limits the expressiveness of the model. Instead, Amin et al. [Amin 2013] proposed to use the 2D PS approach of [Andriluka 2009] per view and to perform inference in 2D using more informative multi-view pairwise constraints. In the end, the 3D pose is reconstructed by triangulation.

In contrast to the above-cited multi-view literature, which is evaluated on single person scenarios, multiple people pose estimation is addressed in [Mitchelson 2003] using a hierarchical stochastic sampling scheme. The samples are ordered based on a fitness function that relies on color, shape as well as temporal cues, and a body kinematic model. In another work, [Huo 2012] proposed an approach to estimate occlusions in each view based on the 2D distances among the projections of 3D silhouettes of all persons in the scene. Then, the estimated occlusion labels and multi-view cues are used to estimate the 3D body poses. Recently, [Belagiannis 2014a, Belagiannis 2014b] have proposed a part-based approach that identifies part hypotheses in the views, which are then back-projected into 3D. Finally, loopy belief propagation is used to estimate 3D body configurations. [Amin 2014] builds on [Amin 2013] and extends the approach to perform pose estimation in two stages. In the first stage, the approach generates multi-view body pose hypotheses using [Amin 2013] and selects the most promising hypotheses, denoted as *key-frames*. The model in [Amin 2013] is then extended to incorporate appearance

similarities with these key-frames as an additional cue. The extended model is finally used to predict body poses.

However, due to the inherent difficulties of acquiring multi-view input, current work on multi-view human pose estimation is generally evaluated on scenarios recorded in controlled laboratory environments that include people in upright poses only. Furthermore, these approaches are proposed for multi-person scenarios in which the number of persons is known in advance. An exception is [Joo 2015], which tackles the problems of multi-person detection and pose tracking during social interactions over multi-view sequences in which people exhibit a larger pose variability compared to the previously mentioned scenarios. A *panoptic studio* is used to capture people. The panoptic studio requires a massive multi-view system that consists of 480 color cameras. In this approach, FMP [Yang 2013] is used as pose detector in each views. Body skeleton trajectories are then estimated in 3D by relying on a multi-view likelihood, body kinematic constraints and motion cues.

2.1.3 Approaches for RGB-D Data

In the last few years, the introduction of affordable RGB-D cameras (*e.g.* the *Microsoft Kinect One* and *Asus Xtion Pro*) has led to many new approaches for human detection and pose estimation. These cameras permit to simultaneously record an environment using both color and depth images. The depth sensor captures the distance of the object surfaces in the scene from the camera viewpoint by decoding a known pattern projected onto the scene in infrared light. The computed depth map can serve as an important source of information to deal with visually similar surfaces in cluttered and crowded environments. Moreover, the depth map can be used to reconstruct the 3D layout of the scene.

In the last few years, some methods have examined how to leverage RGB-D data to improve pedestrian detections. In [Spinello 2011], two SVMs have been trained separately on color and depth images. Then, the detection scores have been combined in order to detect pedestrians. Munaro and Menegatti [Munaro 2014] have proposed a two-stage framework to first cluster a 3D point cloud and then run an RGB-based person detector on the clusters. Regions of interests extracted from 3D point clouds, color and depth features as well as temporal information have also been used for pedestrian detection [Liu 2013, Jafari 2014].

In [Shotton 2012], Shotton et al. have proposed a holistic approach for estimating human body poses on depth images. The approach uses a Random Forest (RF) with deep decision trees to regress for body joint locations. The approach has been trained on a huge number of training images and successfully evaluated on foreground images of human bodies. An extended and commercial version of this approach has been used by the Kinect skeleton tracker. It has shown very promising results in indoor scenes such as living rooms.

Inspired by [Shotton 2012] and its successful application in a commercial product, Buys et al. [Buys 2013] proposed a two-step system for pose estimation. In the first step, a RF-based body part detector is used to scan the whole depth image. Color and

spatial consistencies between body parts are used to compute the foreground. In the last step, the forest combined with a spatial deformation constraint is used to discover body configurations. But, the presence of surfaces with similar color and of clutter can lead to incorrect foreground estimation. Moreover, the approach does not take occlusions and multi-person into account. Baak et al. [Baak 2011] proposed an exemplar-based method to deal with self-occlusion. This approach retrieves the closest exemplar from the database and enforces temporal smoothness to recover 3D body poses. Similarly, Ye et al. [Ye 2011] estimate the 3D body pose from single camera in single person scenarios by using database lookup and nonrigid registration between the exemplar and the observation. Recently, a shallow convolutional neural network and a deep ConvNet have been used to jointly estimate the occlusion and configuration of body parts from a single depth image [Haque 2016]. However, these approaches are only evaluated on single person scenarios captured in controlled laboratory environments.

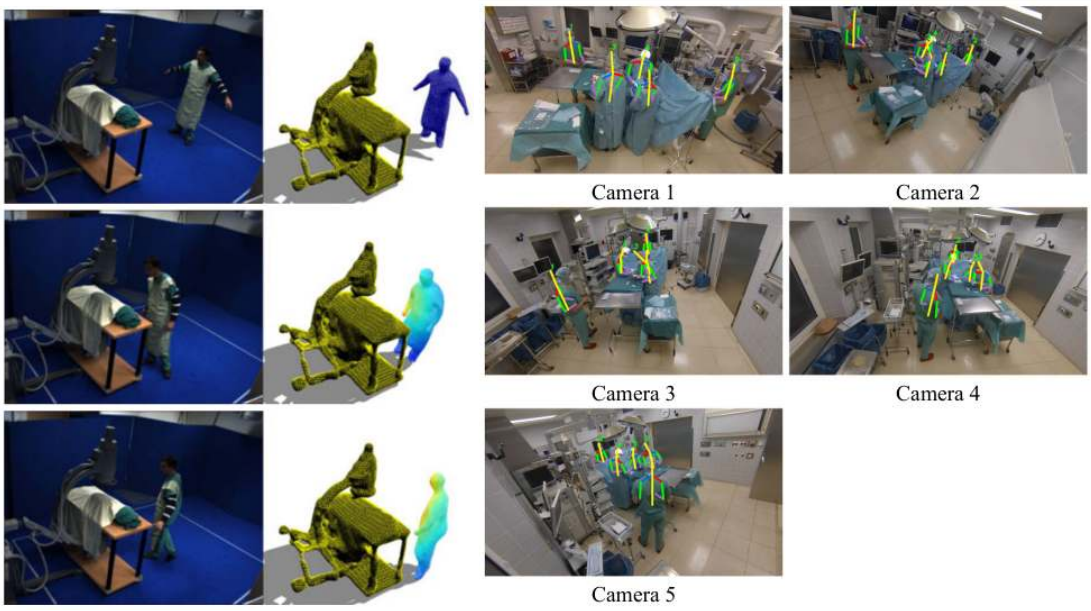
Multi-view RGB-D approaches have also been proposed for human pose estimation in single-person scenarios [Beyl 2015, Xu 2016]. The approaches have relied on multiple RGB-D cameras to capture a person from complementary views in order to reduce the risk of self-occlusion. However, these approaches address human pose estimation in simple scenarios in laboratory setups and also rely on background subtraction.

2.2 Methods for the Operating Room

In previous sections, we have presented an overview on visual human detection and pose estimation. In this section, instead of only focusing on vision-based approaches, we review approaches that are aiming at human detection and pose estimation in operating rooms using any kind of sensors. These approaches are generally using visual sensors, non-visual sensors or a combination thereof.

Recently, multi-view human pose estimation approaches have been proposed for various applications in the operating room [Ladikos 2010, Beyl 2015, Belagiannis 2016]. In [Ladikos 2010], an approach is proposed to reconstruct and track the 3D body mesh of a physician. The approach relies on background subtraction and shape from silhouette in a 16 RGB camera multi-view system. The reconstructed 3D meshes are used to compute the accumulation of the radiation exposure. As stated in the paper, this work does not aim at applying the proposed system in the OR, see Figure 2.1(a). But, this work is mainly motivated by the need for the radiation exposure monitoring system and wants to identify necessary components required to tackle such an application in the OR. One of the main components is body part localization that is required both for computing the radiation risk that varies depending on body parts' positions and for accumulating the risk during an intervention.

In [Beyl 2015], a four-view RGB-D camera system is used to recover 3D body configurations to enable safe human-robot cooperation in operating rooms. The body pose is separately computed per view by using the OpenNI skeleton tracker, called NiTE [OpenNI 2016]. The OpenNI skeleton tracker uses a RF-based method similar



(a) Courtesy of [Ladikos 2010] (b) Courtesy of [Belagiannis 2016]

Figure 2.1: (a) Results on a sequence recorded in a lab using a 16 camera multi-view system. Images recorded from one of the viewpoints are shown in the left-most column. Corresponding radiation exposure estimations are shown in the color-coded 3D meshes next to the images. (b) Human pose estimation results of [Belagiannis 2016] on a dataset recorded using five RGB cameras.

to [Shotton 2012]. However, both [Ladikos 2010] and the OpenNI tracker used in [Beyl 2015] are relying on foreground segmentation, which is difficult to compute in real operating rooms as they have complex dynamic backgrounds. Moreover, these approaches have been only evaluated on single person scenarios recorded in controlled laboratory environments.

A multi-person pose estimation approach is presented in [Belagiannis 2016]. The approach uses a five-view RGB camera system and estimate the body poses in three steps. First, the multi-view tracking system of [Berclaz 2011] and foreground segmentation are used to generate person trajectories over a complete sequence. Second, given the detection bounding boxes provided by the tracker, a deep neural network is used to generate part hypotheses in each view for each person separately. Third, the 3D body part pose of each individual is computed by using the 3D pictorial structures method of [Belagiannis 2014a]. Figure 2.1 displays the five viewpoints used to record a dataset and the 3D poses projected to all views. One should note that this dataset has been recorded during two simulated medical procedures, which does not include all the visual challenges occurring during live surgeries. Moreover, foreground segmentation is not a trivial task to perform in real operating rooms.

Instead of camera sensors, [Agarwal 2007] uses RFID tags to detect the presence/absence of clinical staff and tools. This data is combined with patient monitoring signals as

well as patient history to construct the context of surgical procedures and automatically detect medically significant events. However, in practice, since many RFID tags are present in the room and can be in close proximity of each other, data may get lost due to interferences among the signals returned by the tags. Moreover, people locations cannot be obtained using RFIDs, even though they can provide crucial information for understanding the OR context [Meißner 2014]. In addition to the technical limitations of RFID, it is also difficult to use this technology in operating rooms because of several critical issues such as interferences with other signals, the tedious task of tagging all tools ranging from pretty big to very tiny ones and more importantly the strict infection control regulations in ORs.

In order to track 3D positions of clinical staff, [Bardram 2011] uses the Ubisense real-time location tracking system that is capable of tracking persons who are wearing a tag. In addition, RFID tags are used to detect tools that are used during a procedure. The paper underscores the importance of tool detection and 3D people tracking in order to study the workflow in the operating room. However, in addition to the difficulties implied by the use of the RFID tags, the experimental results show that the tracking system is very noisy and unreliable even during experiments performed in a mock-up operating room.

In another work, an ultrasound-based location system is used to track the 3D positions of the medical staff during neurosurgical operations in order to recognize surgical stages and monitor surgical workflow [Nara 2011]. The 3D location estimation system consists of ultrasonic transmitters worn by staff and an array of receivers mounted on the ceiling of the OR. The staff's trajectory patterns are analyzed to monitor the progress of a procedure and also to automatically detect risky situations. This work has recently been extended in [Nara 2015] by incorporating optical flow computed from single-view color video recordings. The system has been evaluated in an OR at Tokyo's Women's Medical University, Tokyo, Japan. The results show that the video data and 3D trajectories obtained by the Ultrasound-based system are providing complementary information for improving surgical phase recognition. However, to properly track people, a line of sight between the transmitter tag and at least three receivers is required, which is challenging to fulfill in operating rooms with low ceiling or with many articulated arms mounted on the ceiling. Moreover, by using this type of 3D tracker, the system is limited to only rely on the 3D locations of people, while localizing their body parts can provide a much richer source of information as demonstrated in [Wong 2015].

2.3 Thesis Positioning

As discussed in Chapter 1, localizing people and estimating their poses are essential for many applications in the operating room. Due to the OR's requirements, cameras are currently the only practical option to sense operating rooms during real surgeries. The operating room is however a very visually challenging environment due to clutter, similar color of clothes and equipment and occlusions. We propose to use RGB-D cameras

to capture the environment using two complementary color and depth sensors. We investigate different directions to make use of different types of data, namely single-view images, temporal sequences and multi-view images for designing robust human pose estimation models.

Our approaches are built upon the part-based framework that enables us to explicitly model body part articulations and also to learn model parameters using a small dataset.

In this thesis, we introduce a novel single-view RGB-D approach for human detection and pose estimation. Our approach is based on the pictorial structures framework [Fischler 1973, Felzenszwalb 2005], which is the dominant part-based approach, and extends pictorial structures in two ways:

- We extend appearance models to construct part detectors based on both color and depth images. In contrast to [Jafari 2014, Spinello 2011] where HOG, which was originally proposed for color images, is used on depth images, a new descriptor is proposed to encode surface depth changes of objects on the scene. Color and depth detectors are separately learned in [Haque 2016, Munaro 2014]. In contrast, our appearance model jointly relies on color and depth images to benefit from the complementary information coming from the two inherently different sensors.

- The deformation model is extended to rely on 3D constraints instead of 2D constraints [Insafutdinov 2016, Yang 2016, Amin 2013, Yang 2013, Felzenszwalb 2005]. [Burenius 2013] allows for exact inference in 3D. But, to make the exact inference tractable, this approach uses a coarse discretization of the space, which leads to the loss of appearance information, and also uses simple binary pairwise terms. Instead, our approach keeps all appearance information and does not enforce any constraints on pairwise terms.

To enable smooth tracking, temporal information is used in different ways in the literature to enforce pose smoothness across frames [Baak 2011, Amin 2014, Hofmann 2011, Sapp 2011, Tokola 2013]. In this thesis, we investigate a novel approach to estimate 3D pose over an entire sequence. The approach encourages long-term temporal consistency across the entire sequence, which is in contrast to [Sapp 2011, Baak 2011, Hofmann 2011] that are only enforcing temporal consistency between consecutive frames. [Belagiannis 2016] relies on the entire sequence to determine person trajectories, but body poses are estimated for each frame separately. [Tokola 2013] also enforces temporal consistency over the entire sequence by generating a set of trajectory hypotheses for each body part. A greedy algorithm is however used to build these trajectories, which does not deal with multi-person scenarios. By defining the smoothness terms in 3D, our approach reduces the ambiguity of 3D to 2D projection.

In this thesis, we also propose a multi-view human pose estimation method for scenes with multiple persons. Current multi-view approaches are proposed either for single-person scenarios [Burenius 2013, Hofmann 2011, Amin 2014] or for multi-persons scenarios in which the number of persons are known a priori [Luo 2010, Belagiannis 2014a]. Our approach allows to jointly detect and estimate the poses of multiple persons in multi-view

RGB-D setups. In order to reliably detect people and provide good candidates per view, we propose to automatically learn a priori information about the environment and human body kinematic constraints, which is in contrast to current approaches relying on a body kinematic prior alone [Belagiannis 2016, Amin 2013, Burenius 2013], and also propose to use a ConvNet-based body part detector with more informative 3D pairwise constraints instead of 2D pairwise constraints [Amin 2014, Yang 2016, Pishchulin 2016] or image dependent pairwise constraints [Chen 2014, Insafutdinov 2016]. Similarly to [Belagiannis 2014b, Belagiannis 2016], our multi-view energy function drives the body parts towards their optimal 3D locations by jointly optimizing over all views. But, we use depth data to compute reprojection costs and to establish correspondences across multiple views instead of relying on appearance similarity and on triangulation [Gall 2010, Luo 2010, Amin 2014, Belagiannis 2014b, Belagiannis 2016]. To the best of our knowledge, this is the first approach that makes use of joint RGB-D data to address human detection and pose estimation in multi-view setups.

To the best of our knowledge, this is also the first work that addresses clinician detection and pose estimation in real operating room using data from real surgeries.

3 Probabilistic Graphical Models

Chapter Summary

Probabilistic Graphical Models (PGMs), which are also called graphical models, are a unifying framework that combines probability theory and graph theory to elegantly represent various real-world problems. Graphical models provide a powerful formalism to jointly capture uncertainty and dependency constraints to model and attack many real-world problems in different scientific and engineering fields. Probabilistic graphical modes have also become an extremely popular tool for solving computer vision problems in the last two decades due to their flexibility as well as modeling power and to significant improvement in inference methods for such models.

In this chapter, we present a brief overview on graphical models, which are used in the following chapters. We assume a basic background on probability theory and begin by describing two common types of graphical models: *Bayesian Networks* and *Markov Networks*. Then, probabilistic inference techniques are described, which are used to compute the probability of outcomes given observed data.

3.1 Bayesian Networks

A Bayesian Network (BN) is represented by a directed acyclic graph. Hence, it is also called a directed graphical model. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ denotes a graph consisting of a set of nodes $\mathbf{V}$ and a set of edges $\mathbf{E}$.

A Bayesian network is defined by a pair $(\mathbf{G}, \mathbf{\Theta})$, where $\mathbf{G}$ is a directed graph in this case and $\mathbf{\Theta}$ is the set of model parameters. In the graph, the nodes correspond to the variables that we want to model and the edges indicate dependencies among the variables. More specifically, the graph illustrates the conditional independence assumptions in the probability distribution[1] represented by the BN, which are called the *Markov independence assumptions*. In order to define these assumptions, we need to introduce a few definitions on directed acyclic graphs. The *parents* of a node i are all nodes that have an arrow ending at the node. The *descendants* of a node i are all nodes that have a directed path in $\mathbf{G}$, which begins at node i.

Definition 3.1.1 *Let* $\mathbf{X}, \mathbf{Y}$ *and* $\mathbf{Z}$ *be three sets of random variables. We say that* $\mathbf{X}$ *is* **conditionally independent** *of* $\mathbf{Y}$ *given* $\mathbf{Z}$ *in a probability distribution* P *if*

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z}),$$

which is denoted by $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$. □

Definition 3.1.2 *Given a BN defined over the variables* $X_1, ..., X_n$*, the BN graph* $\mathbf{G}$ *encodes the following set of conditional independence assumptions among these variables, called* ***Markov independence assumptions****:*

$$\forall X_i \big(X_i \perp NonDescendants(X_i)|Parents(X_i)\big). \quad \square$$

In other words, the Markov independence assumptions state that each node i (*i.e.* X_i)[2] is conditionally independent of its non-descendants given its parents.

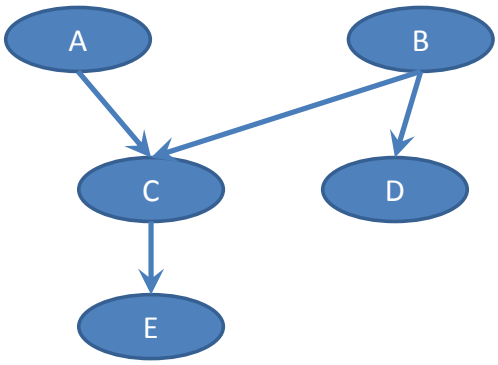
The other component of a Bayesian network $\mathbf{\Theta}$ is the set of conditional probability distributions for all variables in the graph. The probability distribution for each variable is conditioned on its parent nodes in the graph. Considering the set of Markov independence assumptions, the probability distribution over all random variables in the network is factorized and computed by the chain rule:

$$P(\mathbf{X}) = \prod_i P\big(X_i|Parents(X_i)\big). \tag{3.1}$$

To illustrate above mentioned concepts, let us look at the example Bayesian network shown in Figure 3.1. The graph indicates that we have five variables. The BN describes a set of Markov independence assumptions, for examples, $(A \perp B|\emptyset)$, $(C \perp D|\{A, B\})$

[1] Since in the following chapters we only consider discrete variables, throughout this chapter we assume that the variables are discrete.

[2] Since there is a one-to-one mapping between the graph nodes and the corresponding random variables, we use *node* i and variable V_i interchangeably to refer to the random variable.



(a)

A	P(A)
0	0.4
1	0.6

B	D	P(D|B)
0	0	0.2
0	1	0.8
1	0	0.6
1	1	0.4

E	C	P(E|C)
0	0	0.3
0	1	0.7
1	0	0.9
1	1	0.1

B	P(B)
0	0.1
1	0.9

A	B	C	P(C|A,B)
0	0	0	0.2
0	0	1	0.8
0	1	0	0.6
0	1	1	0.4
1	0	0	0.7
1	0	1	0.3
1	1	0	0.8
1	1	1	0.2

(b)

Figure 3.1: A Bayesian Network: (a) the directed graph encoding dependencies among variables, (b) probability tables corresponding to the factors in the graph.

and $(D \perp \{A, C, E\}|B)$[3]. The network gives us the factorization $P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B)P(E|C)$. Note that since the variables are binary, the full joint distribution of $P(A, B, C, D, E)$ requires $2^5 - 1 = 31$ independent parameters. The BN representation, however, needs $1 + 1 + 4 + 2 + 2 = 10$ parameters. In general, this compact representation results in fewer parameters and therefore requires less data to learn. More importantly, this representation is crucial to make inference tractable, which is described in Section 3.3.

3.2 Markov Networks

There exists many real-world phenomena where we cannot naturally attribute directionality to the interaction between the variables. Markov Networks are the other class of probabilistic graphical models that offer a powerful tool to model and learn those phenomena. Markov networks are defined over undirected graphs and are also known as *Markov Random Fields* (MRFs).

Similarly to the Bayesian network, an MRF is defined by a pair $(\mathbf{G}, \mathbf{\Psi})$, where $\mathbf{G}$ is an undirected graph in this case and $\mathbf{\Psi}$ is a set of *potential functions* specifying network parameters. The edges in the graph encode *affinities* among variables, but do not enforce any directionality (*i.e.* no parent or descendant relationships). Therefore, the potential functions cannot be straightforwardly interpreted as probabilities or conditional probabilities.

In order to present Markov random fields, we need to establish some notations and introduce some concepts as well as definitions that will be useful throughout the chapter.

[3]Please note that these are not the complete set of Markov independence assumptions and $\emptyset$ denotes empty set.

Let X_i denotes the random variable corresponding to the i$^{\text{th}}$ node in $\mathbf{V}$. We denote by x_i the instantiation of X_i, where x_i takes its value from the set of all possible values $\mathcal{X}_i$ called the state space of X_i (*i.e.* $x_i \in \mathcal{X}_i$).

Definition 3.2.1 *Two nodes in a graph are* **adjacent** *if both of them are the endpoints of the same edge.* □

Definition 3.2.2 *A* **clique** *is a subgraph of an undirected graph such that every two distinct vertices in the subgraph are adjacent.* □

Definition 3.2.3 *Let us define the domain* $dom(\mathbf{X})$ *of* $\mathbf{X} = \{X_1, ..., X_n\}$ *to be the set of all values* $\mathbf{x}$ *that can be taken from the joint state space of the variables (*i.e. $\mathbf{x} \in \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_n$*). A* **potential function** *or* **factor** $\psi(\mathbf{X})$ *over a set of random variables is defined as a mapping from* $dom(\mathbf{X})$ *to a non-negative real number:*

$$\psi(\mathbf{X}) : dom(\mathbf{X}) \rightarrow \mathbb{R}^{+}. \qquad \square$$

In MRFs, each potential function ψ is defined over a clique. It should be noted that it is not required to define potential functions for all cliques in the graph. But, every potential function should be defined over a clique.

In a similar way to BN, the connectivity in the graph can be used to extract the set of Markov independence assumptions in an MRF.

Definition 3.2.4 *Given a Markov network defined over the variables* $\mathbf{X} = \{X_1, ..., X_n\}$*, we define the* **Markov blanket** *of a variable* X_i *as the set of its adjacent variables in the graph* $\mathbf{G}$*, denoted* $\mathcal{N}_G(X_i)$*. The graph encodes the following* **Markov independence assumptions***:* X_i *is independent of the rest of the variables given its adjacent variables. Formally:*

$$\forall X_i \big(X_i \perp \mathbf{X} - \{X_i\} - \mathcal{N}_G(X_i) | \mathcal{N}_G(X_i) \big). \qquad \square$$

The set of potential functions can be used to compute the joint probability distribution over the random variables (nodes) in an MRF:

$$P(\mathbf{X}) = \frac{1}{Z} \prod_{c} \psi_c(X_c), \tag{3.2}$$

where c belongs to the set of cliques contained in the graph $\mathbf{G}$, ψ_c is the potential function defined over the set of variable X_c in the clique c and Z is the *partition function*. The partition function is defined as:

$$Z = \sum_{x \in \mathcal{X}} \prod_{c} \psi_c(X_c = x_c). \tag{3.3}$$

The partition function works as a normalization factor to make the probability sum to one.

In summary, the Markov independence assumptions imply that a distribution in a MRF factorizes according to the network structure. The interactions between a subset of variables that are in a clique, can be modeled via positive potential functions.

3.3 Inference

Both BN and MRF represent the full joint probability over a set of random variables, where dependencies among variables are encoded by the corresponding graph. Inference in these models corresponds to asking probabilistic queries about sets of variables. The two most common query types are *conditional probability* and *Maximum A-Posteriori* (MAP) queries.

In a conditional probability query, we are interested in estimating a distribution over a subset $\mathbf{Y}$ of random variables in the network, denoted as target variables, given the observed values for a subset $\mathbf{X}$ of the random variables in the network. In a mathematical notation, we want to compute $P(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{Y},\mathbf{x})}{P(\mathbf{x})}$, where $\mathbf{x}$ is an instantiation for the set of random variables $\mathbf{X}$.

In a MAP query, we want to compute the most likely assignment to a set of query/target variables $\mathbf{Y}$ conditioned on observed variables $\mathbf{X}$, where both are disjoint subsets of variables in the network. More formally, our task is to determine $\mathbf{y}^* = \text{argmax}_{\mathbf{y}}\, P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$. The MAP problem can be solved by performing a conditional probability query on the set of target variables and then by finding an assignment for the target variables, which has the highest conditional probability. However, since computing the distribution in case of a conditional probability query is computationally expensive [Cooper 1990, Koller 2007], this is not a very satisfactory approach to solve this inference problem. Moreover, in some cases, it is possible to directly target the MAP problem.

[Cooper 1990] shows that probabilistic inference in graphical models is in general NP-hard indicating that exact inference is intractable in time and in memory. As a result, many computationally tractable inference algorithms are proposed for special cases in order to perform exact inference [Schlesinger 2006, Felzenszwalb 2005] or approximate inference [Komodakis 2008, Wang 2013].

In computer vision, graphical models are often used in order to label pixels/regions of the observed image with different classes, *e.g.* objects, body parts and people. This corresponds to MAP inference in graphical models. It can also be shown that any Bayesian network can be converted into an MRF. We therefore continue this chapter by introducing algorithms for performing MAP inference on MRF.

For the sake of notation simplicity, let us write the maximum a-posteriori inference as:

$$\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\, P(\mathbf{Y} = \mathbf{y}) \tag{3.4}$$

where $\mathbf{Y}$ is a subset of the variables in the MRF graph and $\mathbf{y}$ is an instantiation of the variables from the joint state space $\mathcal{Y}$. The rest of the variables on the graph are clamped to the observed values. Since the potential functions are positive by definition, we can use a logarithmic transformation to define:

$$e_c(Y_c) = -\ln \psi_c(Y_c), \tag{3.5}$$

where c is a clique and $e_c(Y_c)$ is often called an energy function. The joint distribution $P(\mathbf{Y})$ can be represented by:

$$P(\mathbf{Y}) = \frac{1}{Z} exp(-E(\mathbf{Y})), \tag{3.6}$$

where

$$E(\mathbf{Y}) = \sum_{c} e_c(Y_c). \tag{3.7}$$

$E(\mathbf{Y})$ represents the energy of the MRF. MAP inference in Eq. (3.4) corresponds to the minimization of $E(\mathbf{Y})$ as follows:

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmin}} E(\mathbf{y}). \tag{3.8}$$

In general, this energy minimization can be solved exactly or approximately depending on the complexity of the potential functions in polynomial time [Barber 2012, Wang 2013]. The belief propagation algorithm is commonly used to solve this optimization, which is presented next.

3.3.1 Belief Propagation

Relying on *message passing* schemes, [Pearl 1988] presents the *Belief Propagation* (BP) algorithm to optimize the energy of an MRF. In this algorithm, each node receives *messages* from its direct neighbors and updates its current state, referred to as a *belief*, accordingly. The algorithm then proceeds by propagating messages around the graph until it converges to a consensus that all nodes agree on the messages they are sending.

In tree-structured networks, we can use BP to compute the optimal solution. The optimal solution is computed by using root node belief and back-tracing the messages. Therefore, to avoid repeated computations during back-tracing and also when a node needs to send a message several times, *e.g.* in case of more than one parent, exact inference algorithms have historically been obtained using dynamic programming. We will present a variant of BP that works on a subclass of MRFs called *pairwise Markov networks*. The pairwise MRF represents distributions where all potentials are over at most two variables. One can show that any non-pairwise MRF can be expressed using a pairwise MRF [Koller 2007].

In a pairwise Markov network, we can rewrite the energy minimization of Eq. (3.8) as:

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmin}} \sum_{i} e_i(y_i) + \sum_{i,j} e_{i,j}(y_i, y_j), \tag{3.9}$$

where $e_i(y_i)$ and $e_{i,j}(y_i, y_j)$ are the energy functions defined over unary and pairwise potentials, respectively. The optimal solution $\mathbf{y}^* = \{y_1, ..., y_n\}$ corresponds to a setting

of the variables that has the highest joint probability. We denote the message passed from node Y_j to node Y_i by $\mu_{j\to i}$. In the beginning, we initialize all messages to one and then use the following message update formula:

$$\mu_{j\to i}(y_i) = \underset{y_j\in\mathcal{Y}_j}{\operatorname{argmin}} \left(e_j(y_j) + e_{i,j}(y_i, y_j) + \sum_{y_k\in\mathcal{N}_{\mathbf{G}}(Y_j)\setminus Y_i} \mu_{k\to j}(y_j)\right) \tag{3.10}$$

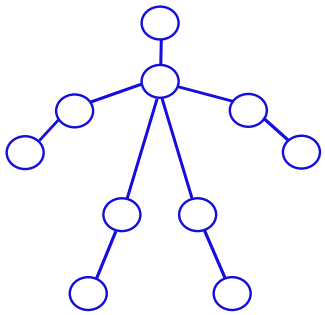
In fact, $\mu_{j\to i}(y_i)$ finds the best instantiation of Y_j as a function of y_i according to the information of its neighboring nodes. An estimation of current energy at node $Y_i = y_i$ is computed according to:

$$b_i(y_i) \propto e_i(y_i) + \sum_{y_j\in Neighbors(y_i)} \mu_{j\to i}(y_i), \tag{3.11}$$

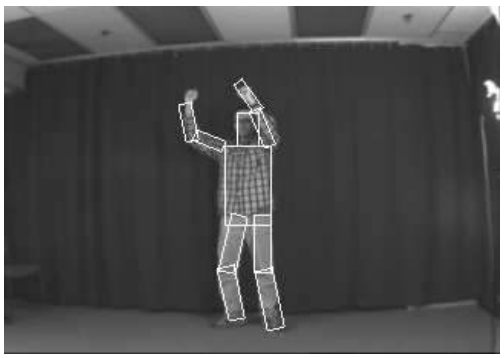
where $\propto$ denotes proportional relationship. The current energy of $b_i(y_i)$ reflects the node's and its neighboring node's information, hence called a *belief*. In this stage, the beliefs can be used to estimate the marginal probability distribution for each variable if needed, *e.g.* in case of a conditional probability query. The above scheme is used to iteratively update messages and beliefs. We stop updating when all messages stay unchanged from one iteration to the next one. This algorithm can be used for both loopy and tree-structured networks.

In case of loopy networks, BP does not guarantee to converge to the optimal solution. Different scheduling schemes are used to pass messages around the network to generate an approximate solution for the problem, which are known as *loopy belief propagation*. [Bishop 2006,Barber 2012]. In practice, loopy belief propagation gives a good approximation of the correct marginal if it converges. But, there might be oscillations in the messages in some cases, which will produce poor results [Murphy 1999]. Therefore, other approximation methods such as numerical stochastic sampling methods, known as *Markov Chain Monte Carlo* (MCMC) [Andrieu 2003] and *variational* methods that are based on deterministic approximations [Grimmer 2011] have also been proposed. Given infinite computational resources, MCMC methods can generate exact results. In practice, because sampling methods are extremely demanding in term of computation, these methods are often used for small-scale problems. Instead, variational methods analytically derive an approximation of the inference problem, which often scale well to large applications. In Section 3.3.1.2, we briefly describe a variational technique that is based on linear programming.

For tree-structured network, we can determine the optimal solution by performing one round of message passing. Note that in a tree, one can define parent-child relationships, when the root node is specified. We start by passing messages from leaf nodes upwards to the root node in the tree. When the root node receives all message from its child nodes, the nodes' beliefs will not be updated anymore. Hence, the optimal solution can be computed. This schemes can also be applied to compute marginal probability distribution for any variable or set of variables in the tree [Barber 2012].



(a)



(b) Courtesy of [Felzenszwalb 2005].

Figure 3.2: Pictorial structures: (a) MRF model used in PSs [Felzenszwalb 2005],(b) a body part configuration recovered using PSs.

In order to estimate the time complexity of the above mentioned inference algorithm, let us assume that the tree consists of n nodes and the size of the state space for a variable V_i is h (*i.e.* $h = |dom(V_i)|$, where $|.|$ denotes the *cardinality* of a set). Since the messages need to be propagated from every possible state in the child's state space to every state in the parent's state space, the time complexity of passing message from one node to its parent is $O(h^2)$. In total, the inference takes $O(h^2 n)$ as it is necessary to repeat the message passing step for every edge in the tree. [Felzenszwalb 2004] has presented a variation of belief propagation to perform exact inference in linear time in the size of the state space, which we describe below.

3.3.1.1 Generalized Distance Transform

Felzenszwalb and Huttenlocher [Felzenszwalb 2005, Felzenszwalb 2004] proposed an efficient approach for inference in tree-structured Pictorial Structures (PSs), which is a tree-structured pairwise MRF used in computer vision. Pictorial structures is a method used to detect articulated objects. Figure 3.2 shows an example of pictorial structures and its corresponding pairwise MRF model. The nodes in the graph represent different components of the object (here body parts). In a similar way to Eq. (3.8), the pictorial structures approach defines an energy $E(\mathbf{Y})$ over the pairwise MRF graph. The unary potentials capture the likelihood of parts being present at an image location. The pairwise potentials encode body kinematic constraints.

[Felzenszwalb 2004] proposed an efficient exact inference algorithm in tree-like pairwise MRFs using *Generalized Distance Transform* (GDT). The traditional distance transform associates the distance between every point on a grid $\mathcal{G}$ and the closest point in a given set $B \subseteq \mathcal{G}$. More formally:

$$\mathcal{D}_B(p) = \min_{q \in B} \big(d(p,q) + \mathbb{1}_B(q)\big), \tag{3.12}$$

where $d(p,q)$ is a measure of distance between p and q. The function $\mathbb{1}_B(q)$ indicates the

membership of an element q in the set B, and has value zero when $q \in B$, ∞ otherwise. The generalization of this distance transform was introduced in [Felzenszwalb 2004] by replacing the indicator function with an arbitrary soft function $f : \mathcal{G} \rightarrow \mathbb{R}$ over the grid $\mathcal{G}$:

$$\mathcal{D}_f(p) = \min_{q \in \mathcal{G}} \big(d(p,q) + f(q)\big). \tag{3.13}$$

Given $h = |\mathcal{G}|$, [Felzenszwalb 2004] proposed a linear time algorithm to compute the distance transform of f in $O(h)$ time under the following conditions:

- $\mathcal{G}$ should be a regular grid. For example, a two dimensional grid $\mathcal{G} = \{0, ..., m-1\} \times \{0, ..., n-1\}$ must be used in case of a 2D image of size $m \times n$.
- The grid should be fully connected. In other words, the closest grid location q for a point p could be at any location on the grid $\mathcal{G}$.
- The distance measure $d(.,.)$ is restricted to be the squared Euclidean distance.

In [Felzenszwalb 2005], GDT and dynamic programming have been used to solve efficiently the MAP problem of Eq. (3.10) where the MRF labels are the image locations on the grid. For the MRF of pictorial structures, the inference algorithm begins by finding the best grid location for each leaf node in the tree as a function of its parent

$$\mathcal{D}'_{e_i}(y_j) = \operatorname*{argmin}_{y_j} \big(e_j(y_j) + e_{i,j}(y_i, y_j)\big), \tag{3.14}$$

where $\mathcal{D}'$ is GDT where min is replaced by argmin. Similarly to Eq. (3.10), the tree is traversed towards its root by updating the belief for any node by $\mathcal{D}'_{e_i}$. The function e'_i is simply the sum of the unary energy and the beliefs delivered by child nodes, which are computed by dynamic programing. A detailed derivation of inference for PSs can be found in [Felzenszwalb 2005]. One can notice that in this case, the inference runs in $O(hn)$ instead of $O(h^2 n)$ time. It is worth mentioning that since the run time is mainly driven by the size of the state space (*i.e.* h), this algorithm dramatically reduces the computation time.

Therefore, the pictorial structures framework serves as a basis for the development of many articulated object detection approaches, some of which have already been discussed in Chapter 2. However, to benefit from the linear time inference algorithm, one should use tree-structured MRF models, where the domain of the variables are defined over a fully-connected regular grid and the pairwise potentials are only relying on Euclidean distances between the nodes. These constraints limit, in turn, the expressiveness of the model for some real-world problems that require loopy dependency between variables or more complex potential functions. Next, we present a framework that permits to efficiently perform approximate inference on loopy MRF models, where the domains of the variables can be any finite set of values and are not limited to regular grids anymore.

3.3.1.2 Fast Primal-dual MRF Optimization

In [Chekuri 2001], a framework has been presented to cast the problem of MRF optimization as an integer programming problem. A wide class of MRFs can be solved by this framework as it only requires the potential functions to satisfy $\psi(y_i, y_j) = 0 \iff y_i = y_j$ and $\psi(y_i, y_j) = \psi(y_j, y_i) \geqslant 0$.

The corresponding integer programming problem is written as follows:

$$
\begin{aligned}
\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}} \quad & \sum_i \Big(\sum_{y_i} e_i(y_i)\mathcal{I}_i(y_i) \Big) + \sum_{i,j} \Big(\sum_{y_i,y_j} e_{i,j}(y_i, y_j)\mathcal{I}_{i,j}(y_i, y_j) \Big) \\
\text{subject to} \quad & \sum_{y_i} \mathcal{I}_i(y_i) = 1 \qquad \forall Y_i \in \mathbf{Y} \\
& \sum_{y_i} \mathcal{I}_{i,j}(y_i, y_j) = \mathcal{I}_j(y_j) \qquad \forall y_j \in \mathcal{G}, (i,j) \in \mathbf{E} \\
& \sum_{y_j} \mathcal{I}_{i,j}(y_i, y_j) = \mathcal{I}_i(y_i) \qquad \forall y_i \in \mathcal{G}, (i,j) \in \mathbf{E} \\
& \mathcal{I}_i(.), \mathcal{I}_{i,j}(.,.) \in \{0, 1\}
\end{aligned}
\tag{3.15}
$$

In a MRF, $\mathbf{E}$ is the set of all edges in the graph, and the grid $\mathcal{G}$ specifies the state space of the variables, which is also called the *label set* in this formulation. The binary variables $\mathcal{I}_i(.)$ and $\mathcal{I}_{i,j}(.,.)$ are indicators of the label taken by the node. If all imposed linear constraints are satisfied, one can show that the above integer programming is equivalent to the original MRF energy minimization of Eq. (3.8) [Chekuri 2001, Komodakis 2007].

The primal-dual scheme is a very popular and powerful technique to solve integer programing problems [Komodakis 2007, Wang 2013]. In order to apply primal-dual scheme, [Komodakis 2007] proposed to relax the last integrality constraints (*i.e.* $\mathcal{I}_i(.), \mathcal{I}_{i,j}(.,.) \in \{0, 1\}$) to the constraints $\mathcal{I}_i(.) \geq 0, \mathcal{I}_{i,j}(.,.) \geq 0$. Depending on the properties of the MRF's potentials, different primal-dual optimization methods are proposed, which in practice, generate nearly optimal results. In [Komodakis 2008], an algorithm called *fast-PD* is proposed to solve the relaxed linear program. Fast-PD exploits information computed based on solutions for both the primal problem and also its corresponding dual problem to speedup the optimization. The evaluation results in [Komodakis 2008] have shown that the algorithm yields significant speedup over other primal-dual based techniques without making any compromises regarding the optimality of the results.

3.4 Chapter Summary

In this chapter, we present an overview of the probabilistic graphical models. We briefly describe two types of graphical models, namely Bayesian networks and Markov random fields. Inference in probabilistic graphical models is also discussed. An efficient version of belief propagation is presented for solving inference in tree-structured MRFs in linear time. We also describe the fast-PD algorithm for inference in loopy MRFs. In the following chapters, we use these models to develop our approaches for performing human

detection and pose estimation in real operating rooms.

For a more detailed explanation of modeling, learning and inference in probabilistic graphical models, we refer the reader to [Barber 2012]. A review on using MRFs in computer vision is presented in [Wang 2013].

4 Temporally Consistent 3D Pose Estimation using Markov Random Field Optimization Over a Complete Sequence

Chapter Summary

In this chapter, we address the problem of temporally consistent pose estimation in Operating Rooms (ORs). A solution to this problem is required in applications that rely on the locations of the body parts over time, such as radiation monitoring, where it is interesting to compute the accumulation of the dose received by each body part. We formulate the problem as a Markov Random Field (MRF) energy optimization defined over an entire set of frames. The proposed MRF energy formulation incorporates image

evidence along with body kinematic and temporal constraints in order to consistently track the body parts of medical staff in short RGB-D sequences. The proposed method is presented in Section 4.2. Evaluation results, presented 4.3, indicate that the proposed approach can consistently track the body parts of multiple persons over an entire sequence.

4.1 Introduction

One of the applications motivating our approach is radiation monitoring. The dramatic increase of the intra-operative usage of x-ray based imaging devices raises indeed the exposure of medical staff to radiation. It is well known that long-term exposure to x-ray can have negative effects on the body, which in the extreme can cause cancer [Vanhavere 2008]. As reported in [Carinou 2011, Ladikos 2010], a correct estimation of radiation exposure requires to compute the exposure at different body parts. Hence, the current practice of using a single dosimeter is not enough to provide an accurate estimation of the radiation exposure of the full body. However, it would be impractical to ask medical staff to wear a multitude of dosimeters on a regular basis, especially on their head and hands. Thus, there exists a need for a noninvasive radiation monitoring system that can be implemented by combining vision-based pose estimation with radiation simulation as in [Ladikos 2010] and [Loy Rodas 2015].

In order to accurately estimate the accumulation of radiation per body part over time, a pose estimation approach yielding temporally consistent results during the short bursts of emission from the x-ray device is necessary. Hence, in this chapter, we focus on consistent upper-body tracking of medical staff present in the operating room during such short sequences. Since the lower-body is generally less susceptible to movement and is occluded by the apron or by the patient table, we exclude it from the tracking approach. We use an RGB-D camera to capture the operating environment and propose to formulate the pose estimation problem as an optimization over the entire sequence using Markov random fields. We introduce a robust cost function that drives the body parts towards their optimal locations by relying on part detection confidences obtained using a body part detector. It also simultaneously enforces body kinematic and temporal constraints over the sequence.

The part detection confidences are computed using the random forest based approach presented in [Buys 2013]. The approach is inspired by the success of random forests in detecting body parts on foreground images [Shotton 2012]. [Buys 2013] relies on a pair of registered color and depth images (a sample pair is shown Figure 4.1(a,b)) and uses random forests in combination with a clustering algorithm to eliminate the need for background subtraction. The approach first assigns a unique label to each image pixel, which could be either background or one of the parts. Then, the predicted labels, color and depth are used to cluster the image pixels into a set of segments, called blobs, that are labeled based on majority voting. Finally, the detected parts, *i.e.* blobs with body part labels, are parsed using dynamic programming for body skeleton estimation. This approach focuses on single-frame pose estimation and the estimated skeletons are often

inconsistent between consecutive frames due to noise, motion and occlusions.

To enable smooth tracking, temporal cues and temporal dependency constraints between body parts in consecutive frames have been used in [Sapp 2011, Amin 2014, Ferrari 2008]. But, these works cannot guarantee long-term temporal consistency across the entire sequence. Approaches have been proposed to leverage information across the entire sequence for generating temporally consistent object trajectories [Butt 2013, Berclaz 2011]. These approaches do not incorporate object articulations, which is necessary for reliably tracking human poses. Few works address the body pose estimation problem as an optimization over the complete sequence. In [Baak 2009], the 3D pose estimation is refined on a complete multi-camera sequence by iteratively using action recognition for retrieving motion priors to restrict the space of possible poses. The evaluation results on single person scenarios show that the combination of pose estimation and motion prior improve both tracking and pose estimation. However, the approach relies on foreground segmentation to estimate body poses in each view, which is not easy to obtain in cluttered and dynamic environments such as operating rooms. In [Tokola 2013], an approach is proposed to generate a set of trajectory hypotheses for each body parts and estimate the body pose over the complete sequence by selecting a collection of body part trajectories. A greedy algorithm is used to build these trajectories, which does not deal with multi-person scenarios.

In this chapter, we propose an approach to consistently track upper-body parts of multiple persons over the entire sequence. We cast the problem as an MRF energy optimization over the complete sequence. The body part trajectories are constructed by using the body part detection responses in all frames in combination with the positions for the body parts in the first and the last frames, which are provided manually. Then, an efficient discrete optimization is used to solve the MRF optimization and drive the parts towards their optimal locations in 3D.

4.2 Method

We define the upper-body pose of a person using the positions of 17 keypoints as shown in figure 4.2. We follow the body kinematics to enforce dependencies between body parts, which are defined as a tree-structured graph over these parts and rooted at the left chest. This is the same skeleton as the one defined in [Buys 2013], but restricted to the upper-body and rooted at the left chest instead of the neck.

Given a set of consecutive RGB-D frames and upper-body poses for the persons in the first and last frames, our goal is to consistently track the poses of the persons over the whole set of frames. In order to consistently estimate and track upper-body pose over a sequence, we define the Markov random field graph $\mathbf{G}$ over the entire sequence. The graph $\mathbf{G}$ is constructed by connecting the upper-body skeleton tree of each person in consecutive frames, as shown in figure 4.3. $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is the set of nodes representing the upper-body parts and $\mathbf{E}$ is the set of edges defining the dependencies between the body parts of each person. Two types of edges are to be considered: kinematic edges ($\mathbf{E}^k$),

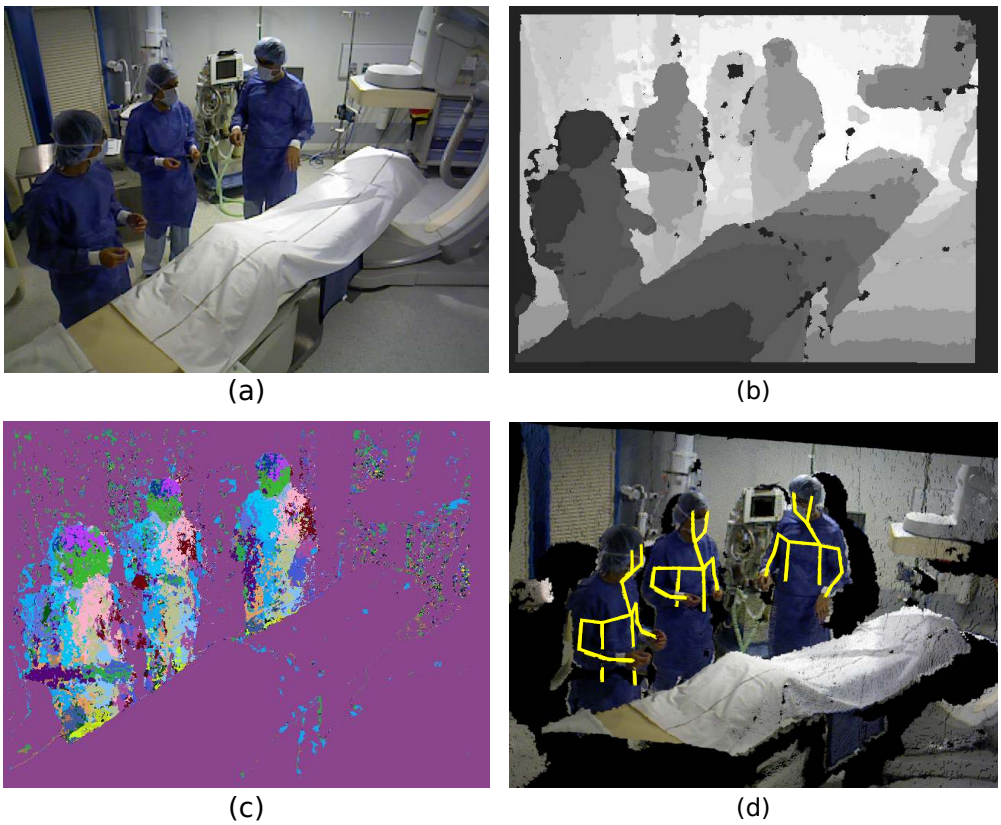


Figure 4.1: (a,b) Example of a pair of color and depth images captured using an consumer RGB-D camera. (c) A sample body part detector response obtained using [Buys 2013]. (d) Overlay of the estimated 3D upper-body skeletons on the reconstructed point cloud.

connecting body parts in each frame, and temporal edges ($\mathbf{E}^t$), connecting the same body parts in consecutive frames. As a result, each person has a connected graph over the sequence.

Pose estimation is then performed by iteratively optimizing an energy function defined over the Markov random field by using discrete optimization. In order to perform the optimization, we need to first initialize the part position (*i.e.* the nodes in the graph) which is obtained using part detection responses, described below. During the optimization, each node can take its value from a finite set of predefined values, often referred to as the label set. In our case, the label of a node specifies a 3D displacement with respect to the initial position of the part. Therefore, by solving this multi-label MRF optimization, the optimal label, *i.e.* relative 3D displacement, for each part can be found. The final 3D positions of the body parts are computed by using these 3D displacement labels to update the initial positions of the parts.

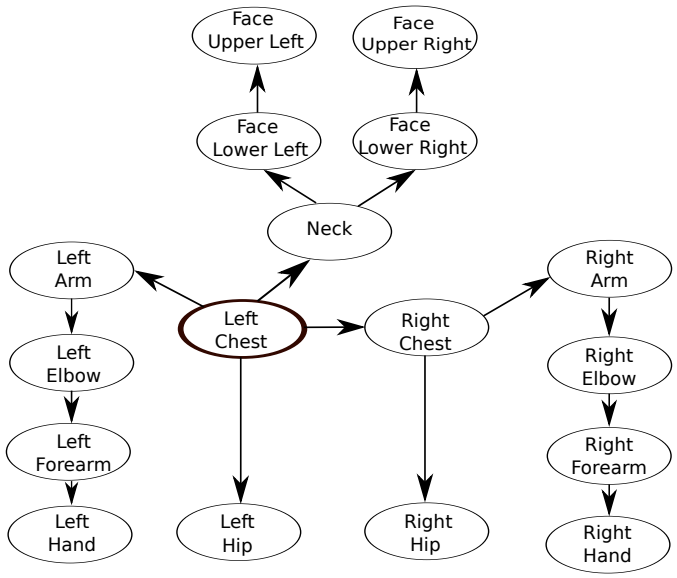


Figure 4.2: Upper-body kinematic tree consisting of 17 keypoints. The root node in this tree is the left chest.

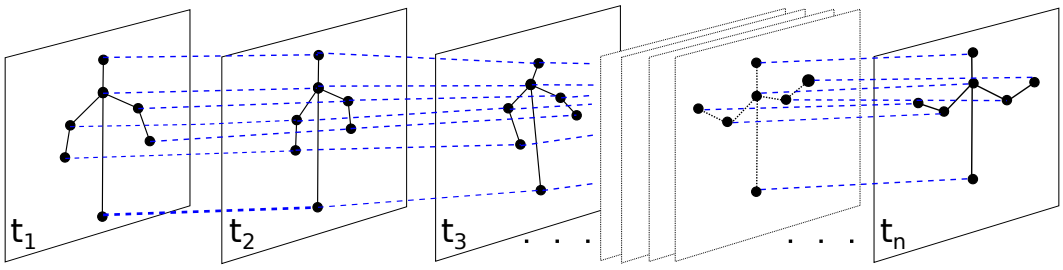


Figure 4.3: MRF graph with kinematic (black lines) and temporal (dashed blue lines) edges over body parts.

4.2.1 Body Part Detection and Person Trajectory Initialization

A depth body part detector is applied to compute the initial position and also to drive the MRF optimization. The random forest based detector of [Buys 2013] is used and hereafter referred to as BPD (Body Part Detector). The BPD uses a random forest model to exhaustively scan the depth image and determine the class of each pixel, which could be either one of the body parts or background. Figure 4.1(c) shows a body part detection response where different colors are used to indicate different classes. The resulting detections are then clustered into blobs based on their class labels, colors and 3D positions that are obtained by re-projecting the points into 3D by using their corresponding depth values. This results in a list of blobs per body part and frame. We associate a confidence value to each blob by counting the number of pixels voting for the part within the blob and by normalizing the values with the size of the largest blob in the frame corresponding to the same part. This gives low confidence to the small blobs that occur frequently in noisy data.

The blob list corresponding to left chest detections in all frames are used along with the left chest positions in the first frame to build initial 3D trajectories for all persons present in the video since the left chest is the root of the kinematic tree. To construct an initial trajectory for a person, the position of the left chest of the person in the first frame is required and provided manually or from the ground-truth data. For the rest of the frames, the left chest blob with highest confidence in a sphere of radius θ centered at the previous position is selected. If no left chest candidate is found in a frame, the previous position of the left chest is considered for this frame as well. The parameter θ is chosen to be the average 3D radius of the body trunk, so that the 3D trajectories of different persons do not get mixed.

4.2.2 Part Position Initialization

The person trajectories mentioned in the previous section only determine the positions of the left chests in the frames. They are used together with the list of detected body parts in each frame to initialize the positions of all parts. In case the part detector fails and does not provide any detection for some parts, we follow a default kinematic model to initialize these positions accordingly. The default upper-body kinematic model corresponds to a person standing in an upright position with the arms by the side of the body.

Given the detected body parts and the position of the root part, two different situations arise for each part: (1) *one or more blobs are available*: the blob with highest confidence value within a neighborhood around the parent position is used to set the position. As in the previous section, the average part size is used to define the neighborhood; (2) *no blob is available*: its position is predicted relative to its parent according to the default kinematic model. Following this procedure for each person, parts are associated in a greedy manner and initial body part trajectories are established. Then, a novel MRF energy function is proposed to derive the parts towards their correct positions, which is presented next.

4.2.3 MRF Model

The Markov random field framework provides a powerful formalism to elegantly model complex problems by jointly capturing both uncertainty and dependency constraints [Barber 2012]. The edges in a MRF graph represent dependencies between nodes. In this work, for obtaining smooth body part tracking over an entire sequence, we define the Markov random field graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ over a complete sequence of video frames to enforce longterm dependencies between body parts. The temporally consistent body part tracking over this graph is formulated as a minimization over the following energy function:

$$E(D) = \sum_{p \in \mathbf{V}} \phi_p(d_p) + \lambda^k \sum_{(p,q) \in \mathbf{E}^k} \psi^k_{p,q}(d_p, d_q) + \lambda^t \sum_{(p,q) \in \mathbf{E}^t} \psi^t_{p,q}(d_p, d_q), \tag{4.1}$$

where $D = \{d_p\}_{p \in \mathbf{V}}$ is a global labeling indicating the displacement for each node, d_p is the 3D displacement offset for node p encoded as a discrete label, $\phi_p(.)$ is the unary potentials representing the data term, $\psi^k_{pq}(.,.)$ and $\psi^t_{pq}(.,.)$ are the pairwise potentials defined respectively on kinematic and temporal edges and λ^k and λ^t are weighting coefficients. The superscripts k and t are used to indicate kinematic and temporal edges. The kinematic and temporal terms force the body parts to follow body physical constraints and to move smoothly along the frames. The data term incorporates the image evidence.

Due to the large search space in 3D, two different methods are compared to sample the search space and define the label set $\mathbf{L}$: dense sampling and sparse sampling. The set $\mathbf{L} = L(n, s)$ depends on two parameters: n the number of samples in each 3D direction and s the step size. In dense sampling, we sample the whole cube, while in sparse sampling we only sample along seven 3D directions, namely top-down, left-right, front-back and the four main cube diagonals [Padoy 2011].

4.2.3.1 Data Term

As mentioned above, the body part detector is used to classify pixels in the depth image, which are then clustered into body parts and background blobs. The list of these blobs is used to define the data term:

$$\phi_p(d_p) = \begin{cases} M(C(d_p)) & \text{if } \#\Big(blobs\big(frame(p), label(p)\big)\Big) > 0 \\ \beta & otherwise \end{cases}, \tag{4.2}$$

where $frame(p)$ returns the frame number of node p, $label(p)$ is the part label associated with the node, $blobs(f, l)$ returns the list of blobs in frame f labelled as part l, $\#(.)$ is the cardinality operator, $C(d_p)$ is the minimum cost defined below, β is a constant cost for parts without detection, and $M(.)$ is a robust error function (ROEF) chosen as

$$M(x) = \frac{x^2}{x^2 + \alpha^2}. \tag{4.3}$$

The function $C(d_p)$ computes the minimum cost of a 3D displacement d_p:

$$C(d_p) = \min_{b \in blobs\left(frame(p), label(p)\right)} \|P(d_p) - Centroid(b)\| * \big(\gamma - Conf(b)\big), \tag{4.4}$$

where $P(d_p)$ is the 3D position of node p moved by an offset d_p, $\|.\|$ is the ℓ_2-norm, and $Centroid(b)$ and $Conf(b)$ are respectively the centroid and the confidence value of blob b. The blob confidence $Conf(b)$ is always between 0 and 1. In practice, it is possible to have body part blobs with confidence of one. Therefore, the value for γ has to be strictly greater than one, *i.e.* $\gamma > 1$, in order to penalize larger distance to the blob centroid.

In Eq. (4.2), the cost of moving a node by a specified offset is computed according to the part detector's response. If no detection is available for the part, a constant cost is

used. As a result, undetected parts are only adjusted by the kinematic and temporal constraints.

4.2.3.2 Kinematic Term

The kinematic term is used to enforce body kinematic constraints between body parts. Following human body skeleton, the dependency between body parts are encoded using the tree-structure dependency graph shown in Figure 4.2. We define the pairwise kinematic potential as:

$$\psi^k_{p,q}(d_p, d_q) = |\|P(d_p) - P(d_q)\| - \mu_{pq}|, \tag{4.5}$$

where $(p, q) \in \mathbf{E}^k$, $|.|$ is the absolute value operator and μ_{pq} is the mean distance between the parts p and q in the kinematic model. In this term, kinematic dependencies between body parts are captured based on the relative displacement between the parts with respect to an average kinematic model, *i.e.* body part lengths. It is worth mentioning that since the part positions are expressed in 3D, the length of a body part does not vary much across different persons. Therefore, it is not required to learn person-specific kinematic models.

4.2.3.3 Temporal Term

The temporal dependency is established between body parts of a person in consecutive frames as shown by the dashed blue lines in Figure 4.3 to enforce temporal smoothness. Temporal consistency of the body parts is enforced by

$$\psi^t_{p,q}(d_p, d_q) = \|P(d_p) - P(d_q)\|, \tag{4.6}$$

where $(p, q) \in \mathbf{E}^t$. Here we assume that parts do not move very fast compared to the acquisition rate of the camera. It would however be possible to incorporate other types of dynamics if needed.

4.2.3.4 Optimization

In order to optimize the proposed energy function, we use the fast-PD algorithm [Komodakis 2008]. In fast-PD, the MRF minimization problem is cast as a linear programing problem, as described in Section 3.3.1.2. The fast-PD algorithm exploits information computed based on solutions for both primal and its corresponding dual problems to efficiently solve the optimization problem. At the end, the algorithm generates a set of labels for all nodes in the MRF graph, which in our case indicates a 3D displacement for each body part with respect to its initial position.

In practice, the initial position for some body parts might be too far from their optimal positions due to either detection failures or inaccurate 3D reprojections because of noisy depth values, which is common in depth images obtained from low-cost RGB-D cameras. Hence, we perform the optimization iteratively by starting with a coarse and

ID	#Frames	#Persons	Misdet.(%)	Room
S1	50	2	28	OR1
S2	100	2	29	OR2
S3	100	3	27	OR2
S4	110	3	32	OR2
S5	200	2	27	OR2
S6	200	2	47	OR1
S7	200	3	29	OR1

Table 4.1: Presentation of the *SV-RGBD-Seq* dataset (sequence IDs, number of frames, BPD misdetection rates and room IDs).

large search space for the labels. At the end of each iteration, the size of the search space is reduced. This is achieved by decreasing the step size s and keeping the number of samples n constant. As results, after each iteration, we use a finer discretization of the search space. After the first iteration, the part's initial positions are computed using the final displacement labels obtained in the previous iteration. During optimization, the nodes in the first and last frames are kept constant using the provided upper-body poses for the persons in the first and last frames.

4.3 Experimental Results

4.3.1 Experimental Setup

We have constructed a dataset, namely the *SV-RGBD-Seq* dataset, by recording seven simulated medical operations in two different operating rooms. The dataset has been recorded using an *Asus Xtion Pro* camera. The sequences have been recorded with a frame rate of 15fps and each sequence has a duration between 3 and 13 seconds. Two to three persons are present per recording. All sequences have been manually annotated to provide ground-truth positions for the skeleton body parts. Parts that are not visible due to occlusions have been annotated too, using positions predicted by the annotator. This is an easy task in case of self-occlusions that occur frequently for the arms. Annotation during inter-person occlusions was also possible in these datasets because they do not happen for long periods. The SV-RGBD-Seq dataset is summarized in table 4.1. The *BPD misdetection rate* is an indicator of the failure of the part detector. It indicates the ratio of parts in the ground-truth skeletons, for all persons and in the complete sequence, that cannot be associated with any part detection.

We perform three experiments to evaluate different aspect of the proposed approach. The first experiment compares the two 3D space sampling methods described in Section 4.2.3. The second experiment quantitatively evaluates the performance of our model for the task of 3D body part localization on the new manually-annotated dataset. The third experiment assesses the impact of noise during trajectory initialization by randomly drifting the parts away.

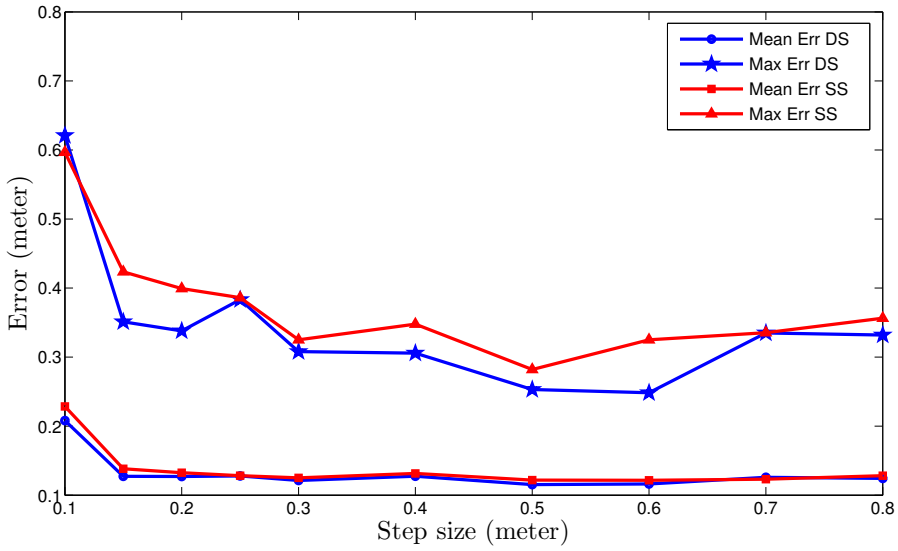


Figure 4.4: Mean and maximum body part localization error for two sampling methods, namely Dense Sampling (DS) and Sparse Sampling (SS), as a function of the initial step size s. The experimental results are obtained using sequence S1. Note that the number of samples n is selected so that the radius of the initial 3D space is around 0.8 meter, *e.g.* if $s = 0.15$, $n = 5$.

In all experiments, the discrete optimization step is iterated by shrinking the sampling step size by 20% at each iteration until the radius of the 3D search space covered by the labels becomes smaller than 5 centimeters. The initial radius of the 3D search space is chosen to be 60 centimeters. The parameters used in all experiments are $\theta = 0.4, \lambda^k = \lambda^t = 3, \beta = 5, \alpha = 0.1, \gamma = 1.01$. They have been determined using grid search over a complete sequence (S1). Errors and positions are expressed in meter. The accuracy is evaluated by computing the mean and standard deviation of the 3D Euclidean distances between the positions of the optimized body parts and their ground-truth potions in all frames.

4.3.2 Sampling Methods

To compare and evaluate the influence of different sampling methods on body part localization performance, we plot mean and maximum localization errors for sequence S1 as a function of initial step size. This localization errors are computed after iteratively optimizing the energy function in Eq. (4.1) as described above. The number of discrete labels per direction is chosen so that the initial 3D space covered has a radius of 0.8 meter. The mean localization errors for the two sampling methods are similar and after the initial step size of 0.2 reach a plateau (see Figure 4.4). However, the lowest value for maximum error is obtained at step size of 0.6, in case of the dense sampling method. Both the mean and maximum errors tend to increase after 0.6. Consequently, we choose

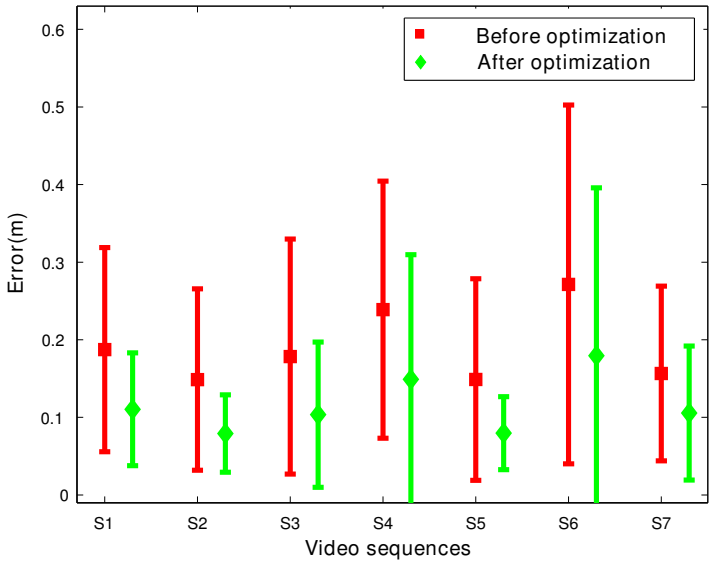


Figure 4.5: Average body part localization error per sequence. The average localization errors are shown before and after optimization.

to use the dense sampling $L(1, 0.6)$ for the experiments below. This corresponds to a set of 27 labels.

4.3.3 3D Body Part Localization

The approach is evaluated on all annotated sequences. Figure 4.5 summarizes the performance of our model for 3D body part localization per sequence. Mean and standard deviation of the part localization errors are shown for each sequence before and after optimization. The results are optimal for sequence S1 in the sense that the parameters have been selected using grid search for this particular sequence. It can be seen that the optimization performs equally well on the other sequences using the same parameters. In general, the mean error has decreased by over 30 percent and the standard deviation (STD) is lowered. Even though sequences S4 and S6 have the highest misdetection rate (see table 4.1), the optimization still reduces their mean error and std significantly. This implies that the optimization has correctly guided the detected and undetected parts toward their correct positions by using the image evidence and the kinematic and temporal constraints.

The detailed evaluation results for sequence S2 is presented in Figure 4.6. In this figure, we report the mean error for each part before and after optimization along with the part misdetection ratio. The results show that the optimization reduces the error considerably. Interestingly, by leveraging long term temporal smoothness and kinematic constraints, we are able to improve the body part localizations even for parts with high misdetection rates, for example see the right elbow (*Relbow*). The figure also compares the influence of the robust error function (ROEF) used in the data term ϕ_p. The ROEF

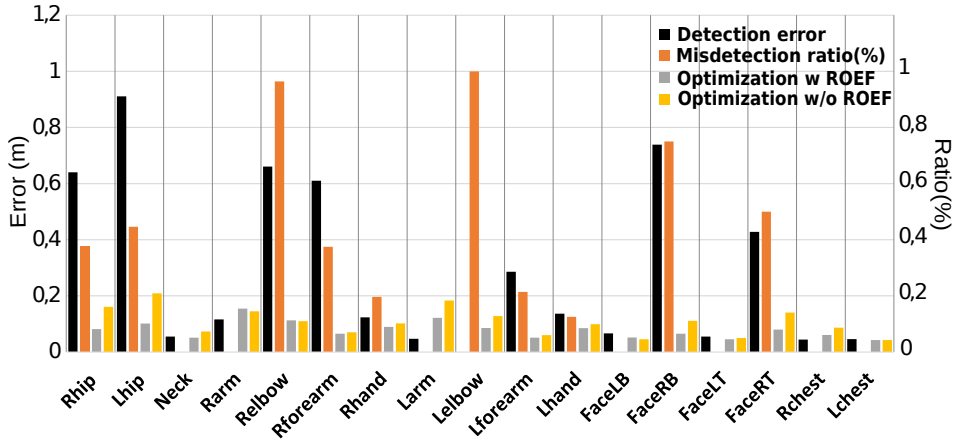


Figure 4.6: Body part localization error. Localization errors (at initialization and after optimization with and without robust error function) are reported for sequence S2 along with BPD misdetection rate.

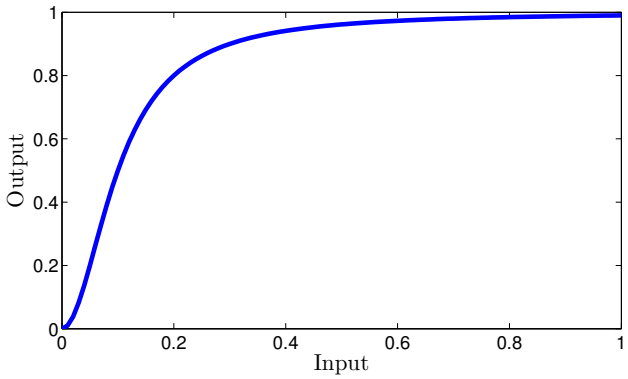


Figure 4.7: The robust error function in Eq. (4.3), where $\alpha = 0.1$.

largely reduces the error for parts with high misdetection rates. The ROEF is steep in the interval $[0, 2 * \alpha]$ and is almost flat in $[4 * \alpha, \infty]$, see Figure 4.7. Therefore, detections strongly attract close nodes but have a negligible impact on far nodes. This is crucial to avoid misleading the undetected parts, considering the high part misdetection rate in our multi-person scenarios.

Figure 4.8 shows estimated 3D skeletons for several frames from the SV-RGBD-Seq dataset. The estimated skeletons are overlaid on the reconstructed 3D point cloud. More qualitative results can be found in this supplementary video[1].

4.3.4 Noisy Initialization

The impact of noisy initialization is studied by adding random 3D displacements to the initial part positions in all frames. The random displacements are sampled from

[1]The supplementary video is available at https://www.youtube.com/watch?v=u0vBIn_h928

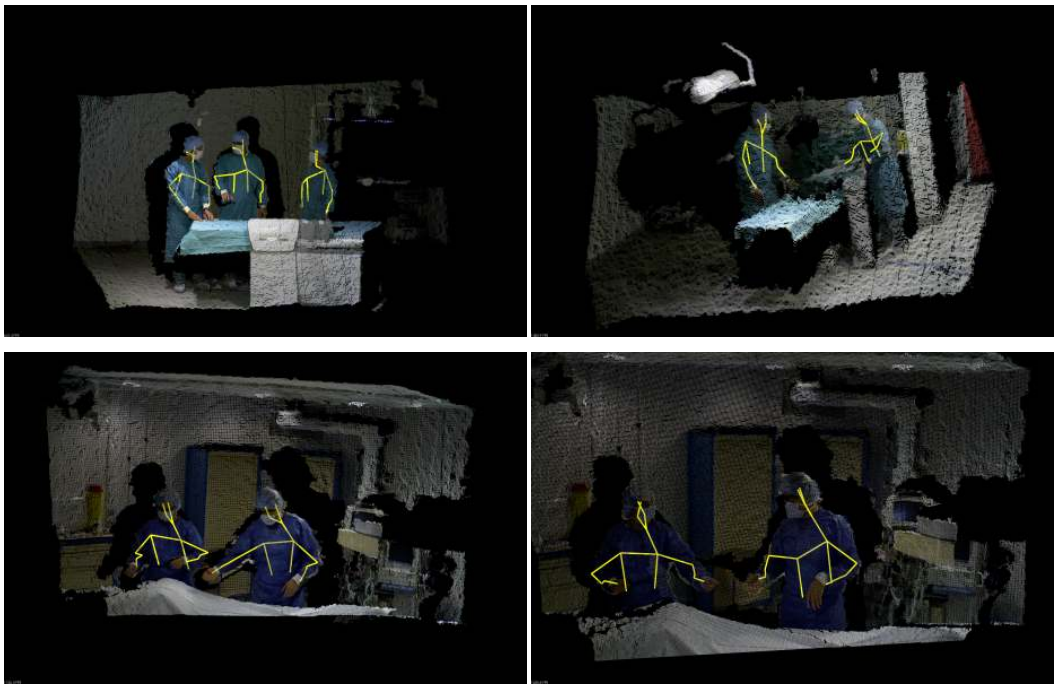


Figure 4.8: Examples of pose estimation results on frames from OR1 (top row) and OR2 (bottom row). The estimated 3D poses are overlaid over the reconstructed point cloud.

a uniform distribution with a magnitude of 50 centimeters. Two cases are considered: noise is added to a single part (the right hip) or to all parts at the same time. Table 4.2 reports the mean and std of the error before and after optimization. When noise is only added to the right hip, the results are reported for this single part. Results are reported on sequences S1, S2 and S3 that have little BPD noise to better identify the performance of our approach against initial noise. The results show that the approach can recover from a large amount of noise.

4.4 Conclusions

In this chapter, we propose an approach to track consistently the upper-body parts of persons present in an operating room over short RGB-D sequences. Due to the inherent visual challenges present in the operating room, the used BPD [Buys 2013] often fails to detect the body parts in individual frames. Consequently, we propose an approach based on optimization over the complete set of frames to recover from detection failures and improve tracking. Our approach uses discrete optimization in an MRF framework. We propose an energy function that incorporates both kinematic and temporal constraints in addition to the image evidence. We evaluated this approach quantitatively on seven manually-annotated RGB-D sequences captured in two different operating rooms. During the evaluations, we have observed that the part tracking error is reduced in average by half. The experiments also show robust results in the presence of multiple persons and

Setting	Sequence ID	Error after initialization	Error after optimization
Right hip	S1	1.02 ± 0.36	0.11 ± 0.05
All parts	S1	1.86 ± 1.17	0.32 ± 0.31
Right hip	S2	0.91 ± 0.36	0.16 ± 0.14
All parts	S2	1.81 ± 1.21	0.31 ± 0.33
Right hip	S3	0.94 ± 0.38	0.13 ± 0.13
All parts	S3	1.87 ± 1.25	0.36 ± 0.38

Table 4.2: Noisy initialization experiment. Mean error in meter with std before and after optimization for right hip and all parts.

occlusions, even when the number of part misdetections is high.

We notice that the reduction in error for the sequences with very high misdetection rates is relatively small. One should note that although the body part detector does not perform well due to the inherent visual challenges of the OR, it is the state-of-the-art approach for human pose estimation on traditional computer vision datasets and promising performances are indeed reported for common indoor scenes, such as working office or home. The results suggest that it is necessary to develop body part detector methods specifically targeting the part detection problem in the operating room and that it also also needed to train them on OR data in order to deal with the many visual challenges present in such a complex environment. In the next chapter, we present an approach to tackle this challenging problem of person detection and body pose estimation in operating rooms.

5 3D Pictorial Structures for People Detection and Pose Estimation

Chapter Summary

As discussed in Sections 1.3 and 4.3, state-of-the-art human pose estimation methods do not generalize well to operating room environments because of the inherent visual challenges present in such environments. Therefore, in this chapter, we propose an approach to detect persons and their body part configurations in operating rooms. The approach relies on a single RGB-D frame. We build our approach upon the pictorial structures framework and use depth data to extend this framework in three ways. First, we use both color and depth images jointly to construct more robust and discriminative part detectors. We also introduce a new descriptor on depth image, called *histogram of depth differences*. Second, we present 3D pairwise constraints to enforce interpart

dependencies directly in 3D instead of 2D since 2D constraints are often ambiguous due to perspective projection. Third, we propose an efficient algorithm for reducing the size of the 3D state space to make exact inference tractable. The method is introduced in Section 5.2. The evaluation results on data recorded during live surgeries are presented in Section 5.3. These results indicate that 3D pairwise constraints and RGB-D part detectors are important for reliably estimating poses of medical staff in visually challenging operating rooms and that the model generalizes well to other operating rooms.

5.1 Introduction

Despite the great success of human pose estimation methods on standard computer vision datasets [Andriluka 2014, Insafutdinov 2016, Yang 2013, Yang 2016], our experimental results on OR data, presented in Sections 4.3 and 5.3, show that there is still a large margin for improvement. The main reasons for such a drop in performance in the OR environment are occlusions, background clutter, the presence of multiple persons in close proximity of each other and the presence of many surfaces with similar colors. Therefore, in this chapter, we introduce a novel part-based approach that makes use of depth both for constructing part detectors that are robust to the visual similarities present in the OR and for assembling body part configurations in 3D. Assembling body configurations in 3D is essential to make more reliable constellations of body parts in cluttered environments.

Our approach is based on the Pictorial Structures (PSs) framework [Felzenszwalb 2005] that has been commonly used in the literature for two reasons: (1) its ability to generalize to unseen data with a relatively small training set and (2) its powerful formalism that allows to explicitly model part detection uncertainties and interpart dependencies. The body part detector in PSs, which is also called the appearance model, relies on color images. The PS approach encodes interpart dependencies using a tree-structured deformation model relying on 2D displacements between body parts in order to make exact inference tractable, as discussed in Section 3.3.1.1.

In order to deal with the inherent challenges of real ORs, our proposed approach extends pictorial structures in three ways: by constructing robust appearance models using both color and depth images, by enforcing pairwise dependencies in 3D, and by proposing an efficient algorithm for reducing the size of the state space for making exact inference tractable. We base our work on a modified version of PSs called Flexible Mixtures of Parts (FMP) [Yang 2013]. FMP uses multiple mixtures of body joints to capture the foreshortening of body parts and part appearance changes. However, the approach still relies on a color-based appearance model and on a 2D deformation model.

In a visually complex environment such as the OR, color images might not always carry enough descriptive information (see Section 1.1.4). Therefore, our approach relies on both color and depth images to build robust and discriminative appearance models, which is in contrast to FMP that only relies on color and to [Haque 2016, Munaro 2014] that are learning color- and depth-based appearance models separately. We have also introduced a new descriptor for depth images, named *Histogram of Depth Differences*

(HDD), that uses depth level changes to encode different object surfaces appearing in a depth image. The descriptor uses small convolution kernels for efficiency.

In general, in part-based methods, the image is scanned using the part detectors that give confidence scores at every image position for every part. The set of all possible positions for the parts is called the *state space*. Then, the deformation model is used to assign connectivity weights for every connections between pairs of parts in the state space. Finally, to recover the body configuration, an inference algorithm is used to propagate detections scores according to the connectivity map. A state space is shown in Figure 5.1(a). In this image, states, also called nodes, are represented by circles. The connectivity strength between two states is shown by the thickness of the connecting edge between these states, here estimated using 2D pixel distances. However, by relying on 2D pixel distances, a path that is connecting two nodes lying over the same person (α-β) can have a weight inferior to a path connecting two nodes lying over two different persons (α-γ). Therefore, performing inference relying on such a connectivity map can result in mixing part detections of persons who appear close to each other in a 2D image but are not necessarily nearby in 3D. Moreover, inference can be additionally confused by false detections on the background in case of weak part detections, which is common in cluttered scenes such as the OR.

In this chapter, to address this limitation we propose an approach based on a connectivity map that relies on the true 3D positions of the nodes instead of 2D ones [Andriluka 2012a, Yang 2013, Pishchulin 2016]. In order to recover the 3D positions, we use the depth map for back-projecting points into 3D. Figure 5.1(b) illustrates the connectivity map for the same state space as in Figure 5.1(a), but where 3D positions are considered. It can be noticed that the nodes lying on the same person are strongly connected (α-β), while connections crossing person boundaries are weak (α-γ). Hence, propagating messages across persons or across a person and the background is discouraged by properly weighting the connections in 3D.

In terms of inference, for the tree-structured Markov random field model used in pictorial structures, the Generalized Distance Transform (GDT) algorithm, described in Section 3.3.1.1 for a 2D state space, can be used in 3D. It has a complexity linear in the number of states [Felzenszwalb 2004, Felzenszwalb 2005]. However, it requires a fully connected regular 3D grid as a state space, which makes the dynamic programming step intractable in terms of memory [Burenius 2013]. With N_p parts and a 3D state space of size S_{3D}, the memory complexity of dynamic programming is $O(N_p * S_{3D})$. Consequently, to make the approach tractable, a very coarse discretization of the 3D state space would be required as discussed in [Burenius 2013], which would degrade the performance of the algorithm. Instead, we propose to use a smaller irregular 3D state space along with the standard dynamic programming approach, which has quadratic complexity. By using 3D information, the number of connections in the state space can be reduced significantly while retaining exact inference. This is achieved by excluding connections that are unreasonably far apart according to human body physical constraints. Although inference still has quadratic complexity, this reduction has a huge impact on run-time.

(a) (b) (c)

Figure 5.1: Two different connectivity maps for the same state space. Circles indicate nodes in the state space and edge thickness denotes the connectivity strength between two nodes: (a) connectivity map built using 2D pixel distances (b) the same connectivity map when real 3D positions of the nodes are taken into account (c) corresponding depth map used to back-project points into 3D. *Note: not all edges are represented in this picture.*

Furthermore, the reduced map contains as many 3D nodes as 2D locations where the detector is evaluated. As a result, no appearance information is lost.

To evaluate our approach, we have generated an annotated dataset, namely the *SV-RGBD-CT* dataset, from seven half-day recordings using an RGB-D camera in an operating room equipped with an intra-operative CT scanner. We have manually generated two types of annotations: upper-body bounding boxes of all clinical staff present in the scene to evaluate human detections and upper-body poses of clinical staff that have at least half of their upper-body parts visible to assess human pose estimation. In Figure 5.2 (top row), we show sample images from the SV-RGBD-CT dataset. This dataset has been divided into disjoint subsets used for either learning the model parameters or performing quantitative evaluation. We have also constructed another dataset, called *MV-RGBD-CArm*, by recording all activities in an OR equipped with a C-arm device by using a two-view RGB-D camera system. Sample images from this dataset are shown in Figure 5.2 (bottom row). Since there is no annotation available for the MV-RGBD-CArm dataset, we have only used this dataset for qualitative evaluation.

5.2 Method

In this section, we briefly present the flexible mixtures of parts model used as the basis to develop our approach for clinician detection and pose estimation. We then introduce our novel 3D pictorial structures approach on RGB-D data followed by a description of the proposed feature descriptor and 3D pairwise constraints. Finally, an algorithm

Figure 5.2: Sample images from two different datasets recorded in different operating rooms during live surgeries. In each row, we have shown sample images from one dataset. In the top row, sample images of the *SV-RGBD-CT* dataset recorded from three different view points are shown. The bottom row shows frames from the *MV-RGBD-CArm* dataset recorded using a two-view RGB-D camera system (the first two images are captured from the same viewpoint and the right-most image is captured from the other one).

is presented to make exact inference tractable in pictorial structures with 3D pairwise constraints.

5.2.1 Flexible Mixtures of Parts (Recap)

The FMP approach represents human body poses by a flexible configuration of body joints [Yang 2013]. The state of a joint i is given by $l_i = (x_i, y_i)$, where (x_i, y_i) represent the joint position in image coordinates and by its joint type $t_i \in \{1...T\}$, where T is the number of joint types. These joint types are defined based on training data annotations, such as joint locations or semantic annotations (a closed versus opened hand) to capture appearance changes. The body pose estimation is broken down into a set of 2D joint detectors combined with pairwise constraints between body joints. The model is formally represented as an energy function defined over a tree-structured Markov random field. Let $G = (V, E)$ be the MRF graph whose nodes are the body joints and whose edges are indicating dependency constraints between body joints. Given the image evidence I, the energy function $S(.,.,.)$ is defined as:

$$S(I, l, t) = \sum_{i \in V} w_i^{t_i} . \rho(I, l_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} . \psi_{2D}(l_i - l_j) + \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i, t_j}, \qquad (5.1)$$

where in a model with N_p body joints, $l = (l_1, ..., l_{N_p})$ denotes the body pose of the person and $A.B = \sum_{i=1}^{n_e} A_i B_i$ is the dot product of two vectors of n_e dimension. The

first term is the body joint appearance model and the second term is the deformation model that enforces spatial pairwise dependency constraints between body joints. The two last terms are capturing joint compatibility, where $b_i^{t_i}$ is the score of choosing a particular mixture type for joint i and $b_{ij}^{t_i,t_j}$ encodes the co-occurrence probability of body joint types.

Appearance model. The appearance model, also referred to as part detector, assigns the score of having a body joint at image location l_i using a part template w_i that is learned during supervised model training and a feature vector $\rho(I, l_i)$. The feature vector $\rho(I, l_i)$ is extracted at image location l_i. FMP uses the Histogram of Oriented Gradients (HOG) descriptor on color image [Yang 2013].

Deformation model. The spatial pairwise constraints between body joints are enforced by w_{ij} and $\psi_{2D}(l_i - l_j)$. The weights w_{ij} encode the deformations between pairs of joints and are learned during supervised model training. $\psi_{2D}(l_i - l_j) = [\overline{dc}, \overline{dc}^2, \overline{dr}, \overline{dr}^2]^T$ captures the relative displacement of joint i w.r.t. joint j, where $[\overline{dc}, \overline{dr}] = [dc, dr] - [ac_{ij}, ar_{ij}]$, dc and dr are the displacements along the columns and rows of the image, and ac_{ij} and ar_{ij} are the average kinematic distances estimated during training between these two body joints. Note that this notation for ψ_{2D} is slightly different from the one in [Yang 2013] by including ac_{ij} and ar_{ij} to allow for better comparison with the generalization to 3D given below.

Learning and inference. In a supervised learning paradigm, the model parameters, namely part templates, deformation parameters and co-occurrence relations, can be learned using a structured prediction objective function. More details can be found in [Yang 2013]. Given an input color image and the learned model parameters, inference corresponds to finding $(l^*, t^*) = \text{argmax}_{l,t}\, S(I, l, t)$. For tree-structured and pixel-based pairwise dependencies, this optimization can be solved efficiently and exactly using the generalized distance transform (GDT) algorithm and dynamic programming.

5.2.2 3D Pictorial Structures on RGB-D Data

Given the availability of synchronous and aligned color and depth images, it is natural to think of models that could benefit from such complementary information. This can be achieved in the pictorial structures framework by using this information to extend both the appearance and deformation models:

$$S(I, D, l, t) = \sum_{i \in V} w_i^{t_i}.\phi(I, D, l_i) + \sum_{ij \in E} w_{ij}^{t_i,t_j}.\psi_{3D}(D, l_i, l_j) + \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i,t_j}, \quad (5.2)$$

where D is the synchronized and aligned depth image of the color image I and ψ_{3D} models the 3D pairwise constraints. The feature vector $\phi(I, D, l_i)$ is computed by concatenating the features extracted from the color and depth images. Following common practice in the literature, we use the HOG descriptor on intensity images (I-HOG). We compare three descriptors on depth images, namely D-HOG (HOG applied on the depth image),

0	0	0
-1	0	1
0	0	0

-1	0	0
0	0	0
0	0	1

0	-1	0
0	0	0
0	1	0

0	0	-1
0	0	0
1	0	0

Figure 5.3: Four different kernels that capture local level changes in depth images.

Histogram of Oriented Normal Vectors (HONV) [Tang 2013] and Histogram of Depth Differences (HDD), defined below.

5.2.3 Histogram of Depth Differences (HDD)

A depth image encodes object surface distances with respect to the depth camera. We therefore propose a novel descriptor on depth images to represent surfaces in the scene based on relative surface distance changes. A related idea has been investigated earlier in [Shotton 2012] that uses an ensemble of deep trees as body part detectors. For each non-leaf node in the trees, a decision function is learned based on the relative surface distances. However, training such a deep forest requires a very large training set and the forest also needs to be retrained for each application. Instead, we present a simple yet efficient new descriptor that captures local surface level changes. The descriptor uses four kernels that are shown in Figure 5.3. Let K_k be one of the HDD kernels with $k \in \{1, ..., 4\}$. At the image position (x, y), the normalized convolution response is defined as:

$$C_{ku}(x, y) = (K_k \ * \ P_u(x, y))/D_u(x, y), \tag{5.3}$$

where $u \in \{1, ..., N_u\}$ is the scale of the depth image. $P_u(x, y)$ and $D_u(x, y)$ are respectively the depth image patch and the depth value at location (x, y) for scale u. The normalization of the response by the inverse of the depth value at the patch center ensures that the feature is depth invariant. We have also applied the convolution over a scale space to encode the changes in different spatial neighborhoods. In order to compute the descriptor, we quantize the convolution responses and also divide the image into non-overlapping windows, called cells. Then, the descriptor is built per cell by computing a 3D histogram of kernel, scale and quantization levels.

5.2.4 3D Pairwise Constraints

In this section, we introduce novel pairwise constraints to enforce body kinematic constraints in 3D. We define

$$\psi^1_{3D}(D, l_i, l_j) = [\overline{dx}, \overline{dx}^2, \overline{dy}, \overline{dy}^2, \overline{dz}, \overline{dz}^2]^T, \tag{5.4}$$

where $[\overline{dx}, \overline{dy}, \overline{dz}] = [dx, dy, dz] - [ax_{ij}, ay_{ij}, az_{ij}]$, $(ax_{ij}, ay_{ij}, az_{ij})$ are the average kinematic displacements in each direction between joints i and j estimated during training and (dx, dy, dz) are the relative displacements between the two body joints in x, y and

z directions. The pairwise constraints ψ^1_{3D} is a natural generalization of ψ_{2D} to 3D. Since body part lengths, *i.e.* the distances between body joints, are relatively constant in 3D, we enforce these constraints explicitly by using the absolute 3D Euclidean distance between body joints. Let ψ^2_{3D} and ψ^3_{3D} be defined as

$$\psi^2_{3D}(D, l_i, l_j) = [|\overline{d_{3D}}|, \overline{dx}, \overline{dx}^2, \overline{dy}, \overline{dy}^2, \overline{dz}, \overline{dz}^2]^T, \tag{5.5}$$

$$\psi^3_{3D}(D, l_i, l_j) = [|\overline{d_{3D}}|, \overline{dx}, \overline{dy}, \overline{dz}]^T, \tag{5.6}$$

where $|.|$ is the absolute value operator, $\overline{d_{3D}} = \|[dx, dy, dz]\| - a_{ij}$ and a_{ij} is the average 3D Euclidean distance between body joints i and j estimated during training. ψ^2_{3D} enforces body kinematic constraints by using not only relative displacement directions and magnitudes along the 3D axes, but also absolute 3D Euclidean distances between the body joints. In ψ^3_{3D}, the square terms are dropped in order to only rely on absolute 3D Euclidean distances $\left(\overline{d_{3D}}\right)$ for enforcing part lengths.

In practice, since 3D positions are computed by back-projecting 2D points into 3D using a depth image, the precision of the 3D positions are driven by the quality of the computed depth image. In the case of low-cost RGB-D cameras, the quality of the depth image is often degraded due to noise and the low resolution of the depth sensor, which decreases quadratically with increasing distance to the sensor [Khoshelham 2012]. In such cases, the 2D annotations are therefore incorrectly back-projected to 3D. During training, these incorrect 3D points can introduce a large error to the part lengths. Thus, we propose a pairwise model ψ^4_{3D} that combines 2D and 3D constraints by relying both on absolute 3D Euclidean distance and on pixel displacement consistency. As will be shown in the experiments, the incorporation of 2D distances into the pairwise constraints is highly beneficial to prevent the learning algorithm from being misled by incorrect 3D positions. The combined pairwise constraints ψ^4_{3D} is defined as:

$$\psi^4_{3D}(D, l_i, l_j) = [|d_{3D}|, \overline{dc}, \overline{dc}^2, \overline{dr}, \overline{dr}^2]^T. \tag{5.7}$$

5.2.5 Learning and Inference

All model parameters are automatically learned using the same approach as in [Yang 2013]. To present the inference, as in [Yang 2013], we denote $z_i = (l_i, t_i)$ for the sake of notation clearness and re-write the optimization as

$$z^* = \underset{z_i}{\operatorname{argmax}} \sum_{i \in V} f_i(I, D, z_i) + \sum_{ij \in E} d_{ij}(D, z_i, z_j), \tag{5.8}$$

where $z^* = (l^*, t^*)$, $f_i(I, D, z_i) = w_i^{t_i}.\phi(I, D, l_i) + b_i^{t_i}$ and $d_{ij}(D, z_i, z_j) = w_{ij}^{t_i,t_j}.\psi_{3D}(D, l_i, l_j) + b_{ij}^{t_i,t_j}$. For a model with a tree-structured pairwise dependency graph, we can write the

Algorithm 5.1 Construction of the state space's neighborhood map

```
1: maxDist_3D ← 0.9                                   ▷ Distances are in meter
2: neighbors = ∅                                      ▷ 2D array to store the neighbourhood map
3: for i = 1 to L do                                  ▷ L: total number of the states
4:     neighbors[i] = ∅                               ▷ Sorted array based on the distances
5:     C = getCandidates(x_i, depth[i])               ▷ depth: array containing the depth values
6:     for each x_n ∈ C do
7:         nodeDist = distance3D(x_i, x_n, depth)     ▷ Euclidean distance between x_i and x_n
8:         if nodeDist < maxDist_3D then
9:             insert(neighbors[i], (nodeDist, x_n) )
10:        end if
11:    end for
12: end for
```

inference as

$$\begin{cases} \underset{z_r}{\operatorname{argmax}} \left(f_r\left(I, D, z_r\right) + \sum_{q \in child(r)} \mu_{z_q \to z_r} \right) \\ \mu_{z_q \to z_p} = \underset{z_q}{\operatorname{argmax}} \left(f_q(I, D, z_q) + d_{qp}(D, z_q, z_p) + \sum_{ch \in child(q)} \mu_{z_{ch} \to z_q} \right) \end{cases}, \quad (5.9)$$

where r stands for the root node in the tree.

We start propagating scores from the leaf nodes in the tree upward to the root node by using dynamic programming. Once all the scores have traversed the tree and reached the root node, pose confidence is available in the root and the corresponding body joint configuration can be recovered by back-tracing the scores. Standard child-parent score propagation is quadratic in the size of the state space since every combination of child-parent nodes needs to be evaluated. Inference corresponds to the same optimization problem as in [Yang 2013]. However, as mentioned in the introduction and in Section 3.3.1.1, to use linear time generalized distance transform, it is required to use a 3D regular grid that makes dynamic programming intractable in memory. In our approach, we therefore use an irregular state space and keep it as small as possible by only back-projecting 2D states into 3D using the depth information. We rely on 3D information to significantly reduce the number of connections in the neighborhood map of the state space. This is achieved by removing the connections between nodes that are too far apart in 3D according to the body physical constraints. Even though this will only reduce the quadratic complexity of the inference by a constant value, it has a dramatic effect during the learning stage that uses the inference intensively. In Algorithm 5.1, we present an algorithm to efficiently construct the neighborhood map by removing the need for comparing the distances between all states. We define $maxDist_{3D}$ to be the maximum allowed part length in 3D. For a given $maxDist_{3D}$ and node x_i at location l_i, only the

2D nodes at a 2D distance $maxDist_{2D}$ of x_i need to be inspected, where

$$maxDist_{2D} = maxDist_{3D} / (res \times depth(l_i)) \tag{5.10}$$

and res is the resolution of a camera pixel obtained from the intrinsic parameters. Since we can store the nodes in a 2D array, the candidate nodes can be accessed in constant time. As the 2D criteria cannot guaranty that the corresponding 3D nodes are within distance $maxDist_{3D}$, this condition is further checked among all potential candidates. A large distance of 0.9 meter is chosen for $maxDist_{3D}$, to make sure that the global optimum is not missed. Once the neighborhood map is constructed, it is used in Eq. (5.9) to propagate the scores between all body parts.

5.3 Experimental results

In this section, we evaluate the proposed approach on two different datasets recorded in different operating rooms. We demonstrate that our new approach achieves significant improvement compared to the original flexible mixtures of parts method on the two tasks of clinician detection and pose estimation.

5.3.1 Datasets

For evaluation, we use two RGB-D datasets, namely the SV-RGBD-CT and the MV-RGBD-CArm datasets. Both datasets have been recorded using *Asus Xtion Pro* cameras.

The SV-RGBD-CT dataset has been recorded using a single RGB-D camera. In order to capture the room from different viewpoints, the camera position is changed among three possible locations. Three sample images from this dataset are shown in Figure 5.2 (top row), where each image shows the room from one of the three different views.

The SV-RGBD-CT dataset includes 1451 annotated frames that are evenly selected across seven half-days of recordings, and 173 negative frames that do not contain any human for training. This dataset contains 3023 bounding boxes annotating all surgical team members who appear in the images. If the head or more than 50% of the upper-body of a person is occluded, it is labeled with a *difficult* flag. There exist 476 persons with difficult flag in the dataset. We have also annotated the clinical staff with ground-truth positions for nine upper-body joints, namely neck, left and right shoulders and hips, as well as left and right elbows and wrists, and the head that is indicated by a point at the center of the head at the eye level. The pose annotation is only provided for staff who have the head and more than five body joints visible. We therefore obtain 1991 persons with pose annotations.

The MV-RGBD-CArm dataset has been also recorded during live surgeries. This dataset has been recorded using a two-view RGB-D camera system. In Figure 5.2 (bottom row), we show sample images from this dataset. There is no annotation for this dataset.

Both datasets cover many visual challenges, such as severe part foreshortening, clutter, occlusion and multi-person scenarios. Since we only have ground-truth annotations for

the SV-RGBD-CT dataset, we use this dataset both for quantitative evaluations and for training the model parameters. In oder to construct disjoint train and test sets, we divide the SV-RGBD-CT dataset into seven disjoint sets where each set only contains frames that belong to the same half-day recording. A leave-one-out scheme is used during our experiments, so that one set is used as test set and the rest as training set. We report the average results of the seven-fold cross validation during the evaluation.

5.3.2 Experimental Setup

We compute all the descriptors using the same parameters as in [Yang 2013]. In other words, we use the cell size of 6×6 pixels and six mixtures for each body parts. Also, we similarly normalize the descriptor responses using the L2-Hys normalization scheme, defined as a L2-norm where the maximum value is limited to 0.2. The HDD descriptor is computed in three scales, and the convolution responses are coarsely quantized into ten levels to be robust to noise and spatial distortions.

During training, we build the mixtures by clustering training data based on ground-truth labels. This step can be performed either in 2D or in 3D. During the experiments, we noticed that clustering in 3D reduces the performance (by $\sim 5\%$ for $\psi_{3D}^{\{1-3\}}$). We believe that this is due to two reasons: *(1) noisy depth:* when the depth value for a ground-truth point is noisy, the 3D back-projection will be inaccurate. Therefore, noisy clusters are generated that lead to an inaccurate division of the part samples. (2) *insufficient number of samples for 3D clustering:* more clusters are required to avoid coarse clustering of the larger 3D space. However, increasing the number of clusters results in a smaller number of samples per mixture, leading in turn to weaker part detectors. Hereafter, we therefore report the results when the part types are generated using 2D clustering.

At test time, in order to be robust to body part scale changes, an image pyramid is constructed by repeatedly smoothing and subsampling the image. We then evaluate our model on the image pyramid and perform non-maximum suppression to detect body parts of different sizes. We have performed our experiments on an Intel i7 machine, with six cores running at 3.20 GHz and 64 GB RAM. Our approach is implemented in MATLAB. In average, it takes five seconds and twelve seconds to test one image using our approach with a 2D deformation model and a 3D deformation model, respectively. But, if we do not use Algorithm 5.1, the computation with a 3D deformation model is 15 times slower. Even though the time complexity remains quadratic in the size of the state space, these results show that the proposed inference approach significantly speeds up the running time.

5.3.3 Clinician Pose Estimation

Evaluation metric. Following common practice in the literature, we use the Percentage of Correct Keypoint (PCK) metric to evaluate human pose estimation accuracy [Yang 2013]. PCK measures the accuracy of localizing body joints. To compute

ψ_{2D} (I-HOG) ψ_{2D} (I-HOG+HDD) ψ^4_{3D} (I-HOG) ψ^4_{3D} (I-HOG+HDD)

Figure 5.4: Examples of pose estimation results for two different appearance models combined with two different pairwise constraints. (Picture best seen in color)

PCK, we calculate a matching distance $\tau = \alpha.max(bb_w, bb_h)$ for each person, where bb_w and bb_h are the width and height of the tight bounding box for that person, and α is 0.2 as suggested by [Yang 2013]. A body joint prediction is correct if its distance from the corresponding ground-truth is less than the matching distance τ.

The average performance results of the seven-fold cross validation are reported in Table 5.1. In column '2D', the result for the I-HOG combined with ψ_{2D} corresponds to the FMP model trained on the SV-RGBD-CT dataset. As discussed in Section 1.3, we have also trained an FMP model on the Buffy dataset, that is widely used in the computer vision community for training and evaluating human pose estimation methods. On the same evaluation setup, the performance of the model trained on Buffy is 32.6% PCK, a drop of 30.7% PCK compared to the model with the same parameters trained on the SV-RGBD-CT dataset. This drastic difference in the performance of these two models suggests that the appearance of the people in operating room environments are very different from what we often find in public computer vision dataset. Therefore, training on OR data is crucial to obtain reliable models for visually challenging OR environments. Hereafter, we consider the FMP model trained on the SV-RGBD-CT dataset as the baseline model.

In Table 5.1 column '2D', we present the results of different appearance models in comparison with the baseline FMP model on the same evaluation setup. The results show that the representation based on D-HOG significantly improves the performance over FMP. Since both I-HOG and D-HOG are using the same descriptor, these results demonstrate that the depth gradient is more reliable than the color-based one in environments with

Descriptor		2D	3D			
Color	Depth	ψ_{2D}	ψ^1_{3D}	ψ^2_{3D}	ψ^3_{3D}	ψ^4_{3D}
I-HOG	–	63.3*	64.8	38.0	56.0	**69.3**
–	D-HOG	72.5	66.6	37.0	61.7	**76.5**
I-HOG	D-HOG	75.3	73.6	58.4	66.6	**79.5**
–	HONV	65.6	67.3	39.5	54.4	**71.0**
I-HOG	HONV	75.4	72.9	55.0	68.8	**80.1**
–	HDD	74.7	73.0	46.7	67.7	**79.1**
I-HOG	HDD	76.6	76.6	70.3	72.3	**81.5**

Table 5.1: PCK results. Comparison of five deformation models in combination with seven different appearance models. Each row shows the evaluation results for an appearance model in combination with the 2D pairwise constraint ψ_{2D} or one of the proposed 3D pairwise constraints ψ^{1-4}_{3D} as deformation model. Note(*): ψ_{2D} with I-HOG is the FMP model [Yang 2013], that is trained on the SV-RGBD-CT dataset.

high color similarities and illumination changes. In general, the depth-based appearance models always outperform the color-based one and the best performance is obtained by HDD. The HDD-based appearance model improves the performance over the baseline by $\sim$ 11%. This highlights the benefit of the proposed coarse and depth invariant representation in describing surface level changes. One can also notice that combining a depth-based descriptor with I-HOG always leads to further improvements by building stronger appearance models that make use of complementary information coming from both color and depth images. For the sake of comparison, we have also built a depth appearance model by combining all depth descriptors, namely D-HOG, HONV and HDD. However, this model does not yield any significant improvement.

The evaluation results for the proposed 3D pairwise constraints in combination with different appearance models are also reported in Table 5.1. In general, the performance dose not improve when $\psi^{\{1-3\}}_{3D}$ are used. We believe that this is due to the noisy depth measurements coming from the low-cost RGB-D camera. The noisy depth leads to inaccurate ground-truth back-projection into 3D, which in turn results in incorrect estimations of part lengths and relative displacements $\left(\overline{dx}, \overline{dy}, \overline{dz}\right)$. Moreover, this noisy 3D data will affect not only the deformation model, but also the part detectors since all parameters are learned in a unified framework. These inaccurate part lengths have more impact on ψ^2_{3D} that uses both the absolute 3D Euclidean distance and the magnitudes of relative 3D displacements between body joints along the axes.

The best performance is always achieved by the model using ψ^4_{3D}, which significantly improves the performance over the model with 2D pairwise constraints using the same descriptor. We observe a significant performance gain (+6%) for the appearance model that only relies on color and does not use any depth information. In the case of the color-based appearance model, the part detector provides noisy detections due to the high color similarity in the images. The 2D deformation model is also not able to resolve the

Figure 5.5: Examples of pose estimation results obtained with the proposed 3D pictorial structures approach using ψ^4_{3D} with I-HOG+HDD. (Picture best seen in color)

uncertainty caused by these weak detections, contrary to the proposed 3D deformation model. These results indicate that by using more reliable pairwise dependencies, the PS model can better resolve the uncertainty of the part detector. These improvements achieved by ψ^4_{3D} also demonstrate that this pairwise term provides an elegant way to benefit from the 3D distances to learn a more reliable deformation model and to use the 2D positions to be more robust to the noise present in the back-projected 3D positions.

Body parts	I-HOG		I-HOG+HDD	
	ψ_{2D}	ψ^4_{3D}	ψ_{2D}	ψ^4_{3D}
Head	84.1	**92.3**	92.8	**96.4**
Shoulder	72.7	**80.5**	84.1	**87.7**
Elbow	57.0	**59.7**	71.1	**76.6**
Wrist	56.5	**64.4**	71.6	**76.8**
Hip	45.9	**52.6**	63.6	**69.9**
Average	63.3	**69.3**	76.6	**81.5**

Table 5.2: PCK evaluation results per body part. Part detection for three variants of our approach compared with baseline FMP (I-HOG+ψ_{2D}) [Yang 2013] on the same experimental setup.

Figure 5.4 shows the estimated poses using ψ_{2D} with I-HOG (*i.e.* the baseline FMP model), ψ_{2D} with I-HOG+HDD, ψ^4_{3D} with I-HOG and ψ^4_{3D} with I-HOG+HDD. We observe that 2D PS is often confused by false detections on the background and also mixes up detections between persons. The last row shows cases where the 3D PS approach does not localize the arms correctly, although the heads and shoulders are correctly estimated. It is either due to weak part detection responses, occlusions or side view poses. Figure 5.5 shows more qualitative results obtained by the proposed 3D pictorial structures that relies on I-HOG+HDD and ψ^4_{3D}. In summary, the use of 3D information always improves the performance, when used in the appearance model alone, in the deformation model alone or in both. The best results are obtained when 3D information is used in both models.

In Table 5.2, we report the detailed performance results per body part. The results are presented for ψ_{2D} and ψ^4_{3D} as well as for the I-HOG and I-HOG+HDD appearance models. It is important to point out that 2D pictorial structures uses exact inference and that these results are therefore the best possible using these appearance models. These results are remarkably improved when the proposed 3D deformation model is used. We therefore show that a reliable 3D deformation model permits to efficiently deploy PS on RGB-D data, while the experiments suggest that 2D-based deformation models are limited by their unreliable pixel-based distance metric.

5.3.4 Clinician Detection

We detect clinician and clinical staff in the operating room using our clinician pose estimation approach. Given a pair of RGB-D images, our clinician pose estimation returns a set of body part configurations. To estimate detection windows from a set of estimated poses, we fit a tight bounding box around each estimated body pose. We compare our approach with deformable part models (DPM) [Felzenszwalb 2010], which has achieved competitive results on challenging datasets for human and object detection. Similarly to FMP, the DPM model is a part-based approach, which also uses

Appearance model	DPM		PS(ψ_{2D})		PS(ψ^4_{3D})	
	N	N+D	N	N+D	N	N+D
I-HOG	75.5	70.0	68.8	64.0	77.3	72.0
I-HOG+HDD	80.3	75.1	86.8	79.1	**89.7**	**80.8**

Table 5.3: Person detection results using AP score. Two variants of our approach are compared with DPM on the same appearance models. N indicates a set of annotated staff who have at least half of their upper-body visible in the view. N+D contains all annotated staff appearing in the view.

multiple mixtures. But, in DPM, parts are automatically discovered using a discriminative approach given a bounding box annotation for each person. Person detection is performed using an energy function similar to FMP, which also consists of three terms: appearance model, deformation model and co-occurrence compatibility score. To build a stronger baseline, we have also extended the appearance model in DPM to use both color and depth images. However, since the parts in DPM are specified with bounding boxes that can include both foreground and background, it is not straightforward to extend the deformation model to 3D.

Evaluation metric. To evaluate clinician detection, we use the Average Precision (AP) score that is commonly used for object and human detection in the literature as well as in [Felzenszwalb 2010]. A detection box is considered as true positive if the overlap between this box and a ground-truth bounding box is more than 50%. Multiple detections are penalized, *i.e.* if more than one detection for a ground-truth occur, one detection will be accepted as a true positive and the rest are false positives. This criteria is used to compute a precision-recall curve and AP is the area under the curve.

Table 5.3 shows clinician detection results. In this table, the results are reported for two appearance models: I-HOG that is used in both FMP [Yang 2013] and DPM [Felzenszwalb 2010], and I-HOG+HDD that was the best appearance model according to our experiments for clinician pose estimation. Following the same reasoning, we also use the two pairwise constraints ψ_{2D} and ψ^4_{3D}. For clinician detection, we consider two cases: *(1) Normal staff:* we compute the true/false positives and negatives only for staff that are labeled as normal, indicated by N in the table. The first detection for a difficult staff is not considered as false positive. If a staff with difficult flag does not have a detection, it is not considered as a false negative. *(2) Normal and Difficult staff:* Missing detection of any staff is considered as false negative, indicated by N+D in the table.

We observe that using the I-HOG+HDD representation improves the performance of the original DPM (I-HOG) by $\sim 5\%$. These consistent improvements indicate that the jointly learned appearance model is highly beneficial in visually challenging environments for both task of human pose estimation and detection. Figure 5.6 shows clinician detection results on normal staff using precision-recall curves computed on the first fold. The increase in the precision and recall of the approaches based on the proposed I-HOG+HDD representation indicates the benefits of this representation in building more reliable and

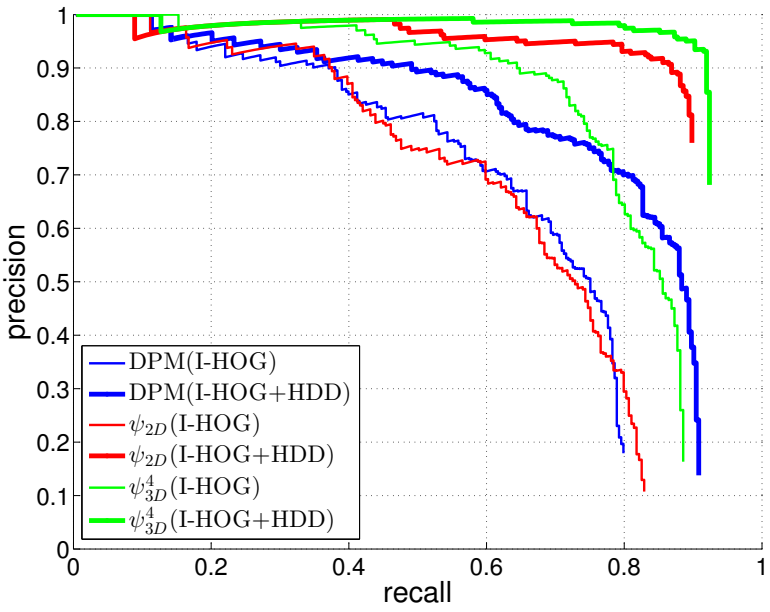


Figure 5.6: Precision-recall curves computed for the detection of normal staff in the first fold of the cross validation. Results for DPM, ψ_{2D} and ψ^4_{3D} in combination with the I-HOG and I-HOG+HDD representations.

discriminative models. Furthermore, this representation enables all models, namely DPM, ψ_{2D} and ψ^4_{3D} to obtain a similar maximum recall value. The high precision for ψ^4_{3D} also highlights the advantages of the 3D deformation model in pruning false positives.

The original DPM (I-HOG) outperforms the baseline FMP model on clinician detection for two reasons. First, since the number of bounding box annotations is much higher than the number of pose annotations, a larger training set is used to learn DPM models. Second, DPM clusters the training data based on box sizes to learn different mixtures. Since the box sizes are normally changing according to the distance of the person to the camera, this enables DPM to learn different mixtures for people at different distances. However, it can be seen that by using the I-HOG+HDD representation, which can encode 3D information, our clinician detection approach consistently outperforms DPM. The best clinician detection result is achieved by using our clinician detection approach with deformation model ψ^4_{3D}. These results further highlight the advantages of 3D pairwise constraints for human detection in cluttered and crowded scenes.

5.3.5 Qualitative Evaluation on the MV-RGBD-CArm Dataset

In order to evaluate the generalization of our model to other operating rooms, a 3D pictorial structures model that uses I-HOG+HDD and ψ^4_{3D}, is trained on the SV-RGBD-CT dataset and evaluated on the MV-RGBD-CArm dataset. Figure 5.7 shows qualitative results obtained by the pre-trained model for several frames in the MV-RGBD-CArm

Figure 5.7: Examples of pose estimation results of a model trained on the SV-RGBD-CT dataset and tested on the MV-RGBD-CArm dataset.

dataset. Note that different viewpoints are used to capture the MV-RGBD-CArm dataset. In general, our model often detects all clinical team members and generates very few false positive detections. For detected people, wrist localization is not very precise. This could be due to occlusions and severe foreshortening of the forearms. But, torso and the arms are correctly localized for most of the detected persons. These results shows that our model generalize well to unseen OR data.

5.4 Conclusions

In this chapter, we propose a novel approach based on pictorial structures for human pose estimation and detection in operating rooms. We extend pictorial structures to 3D on RGB-D data by designing appearance models based on both color and depth images as well as deformation models based on 3D pairwise constraints. We also introduce a new feature descriptor for depth images, the histogram of depth differences, which encodes surface level changes in a coarse, multi-scale and depth invariant representation. Finally, we quantitatively evaluate the approach on a novel and challenging dataset generated from several days of recordings during live surgeries. Different combinations of the proposed appearance and deformation models are compared to state-of-the-art methods for human pose estimation [Yang 2013] and human detection [Felzenszwalb 2010].

Experimental results demonstrate the strength of the proposed appearance model, where the best performance is always obtained by I-HOG+HDD, supporting our hypothesis that color and depth images are complementary. Furthermore, the results demonstrate how the 3D pairwise constraints significantly improve the performances for both clinician detection and pose estimation in a cluttered and busy environment like the OR. Key to this improvement is the use of 3D information to (1) construct 3D nodes; (2) reduce the number of edges in the connectivity map of the state space; and (3) propagate information in the state space by considering 3D distances between the nodes, while retaining an exact solution. To the best of our knowledge, this is the first

time that an approach for articulated clinician detection is proposed and evaluated on a large dataset recorded during real surgeries.

In multi-person and cluttered environments such as ORs, the likelihood of having occluded body parts is quite high due to both inter-person and object occlusions. Capturing the environment using a multi-view system can significantly reduce the likelihood of such occlusions. In the next chapter, we therefore extend our approach to multiple views in order to further improve the results in such environments.

6 A Multi-view RGB-D Approach for Human Pose Estimation

Chapter Summary

In cluttered and multi-person environments like operating rooms, the risk of body part or even person occlusions is very high. These occlusions can dramatically degrade the reliability of the person detection and pose estimation methods. In order to reduce the risk of occlusion, we propose to capture the environment using a multi-view camera system. We introduce an approach for multi-view human pose estimation from RGB-D images and demonstrate the benefits of using the additional depth channel for pose refinement, beyond its mere use for the generation of improved features. The proposed method permits the joint detection and estimation of the poses without knowing a priori the number of persons present in the scene. We evaluate this approach on a novel multi-view RGB-D dataset acquired during live surgeries and annotated with ground-truth 3D

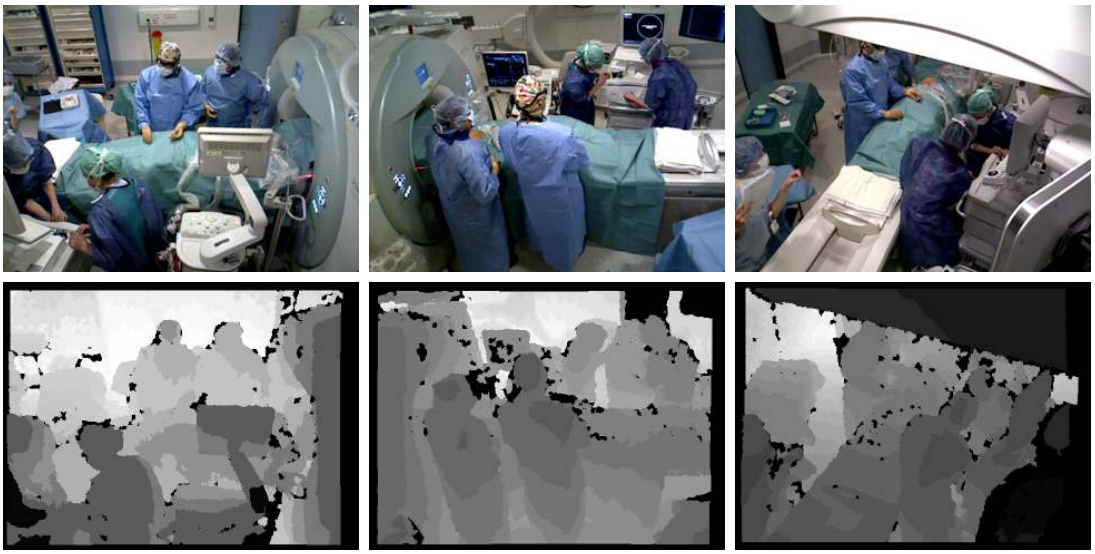


Figure 6.1: Synchronized pairs of color and depth images from a novel multi-view dataset, called the *MV-RGBD-CT* dataset. The images are recorded during live surgeries using a three-view RGB-D camera system.

poses.

6.1 Introduction

In Chapter 5, we demonstrate that the combination of color and depth information along with the use of depth information to model 3D constraints between neighboring body parts greatly improves the pose estimation results. However, occlusions that occur often in such a crowded and cluttered environment, can degrade the performance of the method. We argue that in the operating room, where the working volume is known a priori, multi-view systems can be used to capture the environment from different viewpoints for reducing the risk of occlusions. Following our findings in Chapter 5, we propose in this chapter to use a multi-view RGB-D system to capture the environment from three complementary views. We introduce a novel approach to leverage a priori information about the room for making more reliable predictions and to incorporate evidence across all views for localizing body parts. We show that the advantages of using depth maps in such a multi-view approach go beyond the generation of improved appearance features.

In a multi-view RGB system, correspondences across views are traditionally established by relying on appearance similarity and triangulation [Belagiannis 2014b, Gall 2010, Amin 2014, Belagiannis 2016], which is unreliable in OR environments containing many surfaces that are visually similar (see Figure 6.1). Instead, the depth data enables us to back-project points to 3D. It is also not affected by the visual appearance of the surfaces in the scene. It further enables us to back-project points that are only visible in one view, while in multi-view RGB systems, points should be visible in at least two views.

Current multi-view human pose estimation approaches have been proposed either for single-person scenarios [Gall 2010, Burenius 2013, Hofmann 2011, Amin 2014] or for multi-person scenarios in which the number of persons is known in advance [Luo 2010, Belagiannis 2014b, Belagiannis 2016]. The approach proposed in this chapter makes no assumption about the number of persons in the scene. To this end, our approach first processes each view separately to detect putative skeletons. Next, a priori information about the environment, modeled using random forests, is applied to filter spurious skeletons. The resulting skeletons are then merged across views[1]. Finally, a novel energy function is optimized to incorporate evidence across views and update initial part positions directly in 3D.

Our single-view RGB-D pose estimation approach extends 3D Pictorial Structures (3DPS) presented in Chapter 5 by incorporating Convolutional Neural Networks (ConvNets) for the part detection [Insafutdinov 2016]. ConvNets have recently enjoyed a great success in solving many vision-based tasks including human pose estimation [Toshev 2014, Tompson 2015, Schmidhuber 2015, Insafutdinov 2016, Pishchulin 2016]. They are capable of learning strong detectors that can incorporate a wide image context through deep network architectures with large receptive fields [Wei 2016]. Having access to a wide image context makes the ConvNet-based detectors less subject to false detections, as can be seen in Figure 6.4. This is important for the subsequent multi-view skeleton estimation algorithm to avoid being misled by false detections. However, mutual spatial constraints among body parts are not explicitly modeled, even though they are essential to guarantee joint consistency in the predicted body configuration, especially in multi-person and cluttered environments such as ORs. Therefore, we use a deep ConvNet-based part detector constructed for RGB-D data in conjunction with a 3D pairwise dependency model to enforce body kinematic constraints directly in 3D. This is in contrast to current methods that rely on 2D displacements [Yang 2016, Jain 2014, Tompson 2014] or visual similarities [Insafutdinov 2016] among body joints. As shown in Chapter 5, enforcing body kinematic constraints in 3D is crucial to reliably estimate body part configurations of different individuals who are close to each other in the projected 2D image and are visually similar.

Incorrect detections and occlusions can however result in spurious skeleton candidates in each view that can mislead the multi-view merging algorithm. We argue that in a specific environment like the operating room, *a priori* information about the room should be leveraged to identify spurious candidates. Therefore, we also propose a method to learn a prior on the 3D body kinematic and room layout constraints. This prior, based on random forests, is used to recognize and remove skeletons with unlikely 3D shapes or positions. Relying directly on high level 3D skeleton information enables the model to better explore the a priori information of the OR and to build a stronger prior compared to traditional pose priors that are based on the displacement or visual similarity among parts [Yang 2016, Insafutdinov 2016, Kadkhodamohammadi 2017a].

Single-view skeleton candidates are then merged using a multi-view fusion algorithm

[1] We assume that the extrinsic parameters of the cameras are known.

to generate a set of initial multi-view skeletons. We use our novel multi-view energy function to drive body parts towards their optimal locations by leveraging depth data and reasoning across all views. We use the depth data not only to establish correspondences but also to a compute reprojection cost that is otherwise often computed based on appearance similarity and triangulation [Amin 2014, Belagiannis 2016, Burenius 2013].

We have generated a multi-view dataset, called *MV-RGBD-CT*, from several days of recordings. We have manually annotated the dataset for both 2D and 3D positions of upper-body skeletons. The dataset has been used for evaluating our approach and also for performing comparisons with several state-of-the-art methods.

6.2 Method

We start this section by recapitulating the 3D Pictorial Structures (3DPS) of Chapter 5 and then present the different components that lead to our multi-view RGB-D approach.

6.2.1 Single-view Body Pose Estimator

The 3DPS model represents the body as a set of N_p joints and learns multiple mixtures of parts to capture appearance changes. This model uses ten body joints to indicate upper-body poses, since lower body parts are often occluded in operating rooms. A body configuration is specified by a pair (l, t), where $l = \{l_1 ... l_{N_p}\}$ indicates the 2D positions of the body joints and t_i belongs to a set of T possible mixture types $t = \{t_1 ... t_{N_p}\}$ for each body joint. The pose estimation is defined as an energy minimization over a tree-structured graph $G = (V, E)$, whose nodes are the body joints and whose edges indicate dependencies between joints. The body joint dependencies are defined following the human body skeleton. Given a pair of aligned color and depth images denoted by I and D, respectively, the score associated with a body configuration (l, t) is defined as:

$$S(I, D, l, t) = \sum_{i \in V} \phi(I, D, l_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} . \psi(D, l_i, l_j) + \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i, t_j}, \qquad (6.1)$$

where, similarly to 3DPS in (6.1), the first term is the appearance model (or the part detector), and the second term is the deformation model that enforces pairwise dependencies between body joints and the last two terms are part type compatibility score functions.

The part detector assigns a confidence score for placing the body joint i at image location l_i. 3DPS relies on handcrafted features, namely Histogram of Oriented Gradients (HOG) and Histogram of Depth Differences (HDD). Here, we compute the part detection scores using the deep ConvNet model presented in Section 6.2.2.

We follow the 3DPS formulation presented in Section 4.2 to construct the last three terms. The part detector is learned using the method presented next, but other model parameters, namely compatibility score functions and $w_{ij}^{t_i, t_j}$, are learned using a structured support vector machine formulation similar to 3DPS. The inference is also performed using the algorithm presented in Section 5.2.5.

6.2.2 ConvNet-based RGB-D body part detector

Motivated by the great success of deep convolutional neural networks in recent years [Schmidhuber 2015, He 2015, Tompson 2015, Insafutdinov 2016, Newell 2016], we propose to use RGB-D body part detectors based on deep ConvNets in order to automatically learn feature representations instead of relying on engineered feature representations such as HOG or HDD. To this end, we build on the very deep residual network [He 2015], which has recently been used for part detection and shown promising results [Insafutdinov 2016]. The body part detection is formulated as a multi-label classification problem, where a set of n scores is generated at each image location to denote the probability of part presence. The scores are obtained by using sigmoid activation functions on the output neurons. We adapt the network to learn body part detectors for pairs of color and depth images. We change the input layer to accept four dimensional data (*i.e.* three color channels and depth channel). We also change the `res3d_pose` layer to generate part score maps for ten upper-body parts instead of the fourteen full body parts. During pose estimation, we use the ConvNet-based body part detector to predict confidence scores for all parts at every image locations. Hereafter, we refer to this human pose estimation model as *Deep3DPS*.

Fine-tuning. We initialize the network from the pre-trained model of [Insafutdinov 2016], which is trained on the *MPII Human Pose* dataset. We fine-tune the network on the *SV-RGBD-CT* dataset from Chapter 5 using the Caffe framework [Jia 2014]. We scale the images down to 85% and use a batch size of two. Similarly to [Insafutdinov 2016], we generate target training score maps for all body joints by assigning the positive label 1 for all image locations within 15 pixels to the ground-truth locations and negative label 0 otherwise. During training, we use all positive samples and keep at most three times more negative samples. The network is trained with cross entropy loss and stochastic gradient descent for 50k iterations. The initial learning rate is set to 5×10^{-5} for the adapted layers and 5×10^{-6} for the rest. This yields the best results in our experiments. In [Insafutdinov 2016], the network is trained for three tasks: body part detection, location refinement, which is the relative row and column displacement from a scoremap location to the ground-truth, and regression to other parts. However, training for the last two tasks did not yield any performance improvement during our experiments. We therefore only train for the body part detection task.

6.2.3 Random Forests Based Prior

To design a robust method, we believe that it is essential to include priors specific to the environment. Even though a general body kinematic prior is included in the pose estimation model through pairwise constraints, it cannot be guaranteed that these constraints are always properly enforced due to the high complexity of the pose estimation model that predicts human poses directly from image pixel values. In addition, this prior only captures body kinematic constraints and does not incorporate a priori information about the environment. In an environment like the OR, constraints such as possible

human poses and possible locations can also be used to improve the reliability of the method. Such constraints cannot be easily handcrafted. Furthermore, including them in the pose estimation model would need higher-order dependency terms. Adding such terms would increase the number of model parameters and, more importantly, dramatically increase the complexity of the inference algorithm. We therefore propose to automatically learn the prior, which we formulate as a binary classification problem that takes a skeleton estimated by the single-view detector as input and outputs whether this skeleton corresponds to a spurious detection or not.

We base our approach on Random Forests (RF), which are an ensemble of decision trees consisting of two types of nodes: split and leaf nodes. In each split node, a decision function is implemented to forward samples to one of the branches until they finally reach a leaf node containing a prediction function. In our case, we use RF with binary trees and the mean over all predictions to aggregate the votes across all trees. The trees are learned automatically given a labeled training set, which we construct using the skeletons estimated by our single-view pose estimator on a set of images for which ground-truth is available. The detected skeletons are compared to the ground-truth using the probability of correct keypoints (PCK) metric, which is commonly used for evaluation in multiple-person pose estimation [Yang 2013, Pishchulin 2016]. We label a detected skeleton as positive if the head, neck, and left and right shoulders are correctly localized according to PCK.

For RF training, we propose to combine various features computed from the 3D skeletons, which are all expressed in the common room reference frame. The reference coordinate system is chosen w.r.t. the operating table in default position, which makes the prior generalizable to other ORs. This enables our prior to encode two types of information: room layout and possible clinician poses. Certain parts of the room, such as the floor or the ceiling, are for instance not expected to have clinicians or certain body parts. Thus, as first set of features, we use the positions of the 3D body parts to enable the RF to build an internal representation of their spatial occupancy probability. To capture the set of possible human poses in the OR, we include a second set of features, namely the relative 3D displacements between all pairs of body joints. The prior also serves to verify 3D part lengths and exclude incorrect skeletons that may occur due to weak detections and foreground/background confusions. As third feature, we include the detection score of the individual skeleton to incorporate detection confidence. To enable our prior to better encode high-level information, we use the RF method in a multi-layer scheme, referred to as auto-context in the machine learning literature. A multi-layer model is learned, where the first RF layer is constructed using only the three aforementioned types of features, while the other layers use another extra feature that is the classification confidence generated by the previous RF layer.

6.2.4 Multi-view Human Pose Estimation

6.2.4.1 Multi-view fusion

The objective of the multi-view fusion is to combine the 3D skeletons across all views. For a given *frame*, defined as a set of RGB-D images recorded from all cameras at the same time step, detections from all views are first put in a set. The two closest skeletons that do not originate from the same view are then merged. This procedure is iterated until no pair of merging candidates is left in the set, where the condition for merging two skeletons is that the distance between their heads and the distance between their necks are both smaller than a constant T_s. Since the left/right side labels of the individual detections are not always reliable, to ensure a consistent merging of the 3D joints we use the 3D positions of the shoulders to find the correct association between the two skeletons. Finally, for all skeletons resulting from a merging step, the left and right side labels are set based on a majority vote among the supporting skeletons. If a merged skeleton originates from only two supporting skeletons, which do not agree on the side label, we set the side according to the skeleton with highest confidence.

As a result, we obtain a set of initial 3D skeletons generated from skeletons coming from one or more views. Then, a new multi-view energy function, presented next, is used to drive the body parts towards their optimal 3D locations by jointly optimizing over all views.

6.2.4.2 Multi-view RGB-D Optimization

We formulate our multi-view RGB-D approach as an energy minimization over the same graph G as in Section 6.2.1 and define the energy function $E(\Delta)$ over the graph as:

$$E(\Delta) = \sum_{i \in V} \Big(\lambda_1 . \Phi^{conf}(\delta_i) + \lambda_2 . \Phi^{depth}(\delta_i) \Big) + \sum_{(i,j) \in E} \Psi_{i,j}(\delta_i, \delta_j), \tag{6.2}$$

where λ_1 and λ_2 are weighting coefficients, $\Delta = \{\delta_1 ... \delta_n\}$ is a set of displacement labels for all body parts, $\delta_i \in \mathbb{R}^3$ is a 3D displacement offset for part i, $\Phi(.)$ are the unary potentials and $\Psi_{i,j}(\delta_i, \delta_j)$ is a pairwise dependency term enforcing body physical constraints.

The first term in (6.2) incorporates part detection confidence scores computed by the ConvNet part detector. Given the list of all views *views*, we define:

$$\Phi^{conf}(\delta_i) = \sum_{\mathtt{v} \in views} conf\Big(proj\big(P(\delta_i), \mathtt{v}\big) \Big), \tag{6.3}$$

where $P(\delta_i)$ is the 3D position of part i displaced by an offset δ_i and $proj(p_{3D}, \mathtt{v})$ projects the 3D point p_{3D}. In order to provide a smooth cost function, we compute the distance transforms of the deep ConvNet score maps using the generalized distance transform algorithm [Felzenszwalb 2005]. We find that this transformation is necessary to avoid local minima. $conf(p_{2D}) \in [0..1]$ is the value of the distance transform of the score map

of part i at location p_{2D}. The second term is defined as:

$$\Phi^{depth}(\delta_i) = \sum_{\mathtt{v} \in views} \Big| D\Big(proj(P(\delta_i), \mathtt{v})\Big) - Z(P(\delta_i), \mathtt{v}) \Big|, \tag{6.4}$$

where $D(p_{2D})$ is the depth value at image location p_{2D}, $Z(p_{3D}, \mathtt{v})$ is the z value of the 3D point p_{3D} in the coordinate system of the view and $|.|$ is the absolute value operator. To reduce the effect of the noise present in the depth image, we smooth the depth image with a median filter of size 7×7px. This term quantifies the distance between the displaced 3D joint and the surfaces captured by the depth cameras. Therefore, it can help to avoid placing parts in ghost 3D locations that do not correspond to any surface in the scene. These two unary terms incorporate multi-view cues, where the RGB-D ConvNet is used to include image evidence and depth is used to integrate a reprojection cost across all views.

The pairwise term is used to enforce kinematic constraints, namely body part lengths between pairs of joints. Let $\Psi_{i,j}$ be defined as:

$$\Psi_{i,j}(\delta_i, \delta_j) = |\|P(\delta_i) - P(\delta_j)\| - \mu_{i,j}|, \tag{6.5}$$

where $\|.\|$ is $\mathcal{L}_2$-norm and $\mu_{i,j}$ is the average distance between joints i and j, *i.e.* average part length. The average part lengths are computed over the entire training dataset. Note that since the body part lengths are relatively constant in 3D, it is here not needed to learn person-specific average part lengths.

Inference. In order to recover 3D body part configurations, we need to perform inference in 3D. This problem corresponds to optimizing the energy function in Eq. (6.2). Note that using the optimization algorithm of 3DPS would require to construct a 3D state space that includes all 2D positions back-projected to 3D (amounting to the number of views multiplied by the size of the images) augmented with extra neighboring nodes for each back-projected node to account for occlusions. Such a large state space would degenerate the performance and slow down the inference. Similarly, the inference approach from [Burenius 2013] would limit us to use simple binary pairwise terms. Instead, we perform discrete optimization using the *fast-PD* algorithm [Komodakis 2008], which casts the optimization problem in an integer programming framework and exploits solutions from both primal and dual problems for efficiency. To perform the optimization, we define a set of discrete displacement labels $\mathcal{L}$ for each body joint by sampling densely from a cube centered at the initial joint position. The sampling function is parametrized by (n, s), where n is the number of samples along each 3D direction and s is the step size between the samples. We perform the optimization iteratively by starting with a coarse label set with a large step size to cover a large 3D space. At the end of each iteration, we update the part positions based on the displacement labels and then generate a finer label set for the next iteration.

Figure 6.2: Annotation tool. Three views and the 3D point cloud (bottom right) are shown in the window. Right side body parts are indicated in green color. Occluded body parts are denoted by crosses. The annotator can move points in either 2D or 3D. The correctness of an annotated skeleton can be verified using both the 3D point cloud and its reprojection to the views.

6.3 Experimental Results

Datasets. We have generated a multi-view RGB-D dataset, illustrated in Figure 6.1, by recording all activities in an operating room for four days. This dataset is called *MV-RGBD-CT*. For quantitative analysis, the 3D upper-body poses of 1378 clinicians have been manually annotated in 741 multi-view frames that are evenly distributed across the dataset. All clinicians who have more than 50% of their upper-body parts visible in at least one view have been annotated in these frames. The annotations are performed using a tool that displays a 3D point cloud reconstructed from all three views as well as the corresponding individual 2D images. This tool allows the user to move the body joints either in the 2D views or in the 3D point cloud. The annotator often starts by annotating 2D positions for all body parts in all views. For each person, an average 3D skeleton is computed by back-projecting 2D annotations into 3D using the depth map. Whenever a joint is moved in 2D, the 3D position of the corresponding joint in the average 3D skeleton is updated. However, noisy depth and side-view of a person can lead to inaccurate average 3D skeletons. Therefore, to ensure the correctness of the 3D skeletons, the annotator verifies the reprojection of the average 3D skeletons to all views and move joints directly in 3D. Figure 6.2 shows a snapshot from the annotation tool developed by our group. All 2D images and the reconstructed 3D point cloud (bottom right window) are shown simultaneously. The annotator can check the correctness of the

Figure 6.3: Multi-view examples illustrating the results of the RF-based prior. Accepted skeletons are shown in orange and rejected skeletons in purple.

Figure 6.4: Part detection score maps. These score maps are generated using *Deep3DPS (RGB-D)* and overlaid over the corresponding color images

reprojections and move joints in 2D or 3D.

In order to have enough data for ConvNet training and disjoint test set, we use the SV-RGBD-CT dataset presented in Chapter 5 to train all 3DPS single-view pose estimation models and for fine-tuning the network. For all models, evaluation is performed on the MV-RGBD-CT dataset. As the MV-RGBD-CT dataset is used for random forest training, to evaluate the model, a 4-fold leave-one-out cross-validation is performed, where three folds are used for training and the rest for testing. The evaluation reports the average results of the cross-validation.

6.3.1 Single-view Pose Estimation

Table 6.1 reports the performance of different models on the MV-RGBD-CT dataset using the *PCK* metric [Yang 2013, Insafutdinov 2016]. All 3DPS models have been trained on the SV-RGBD-CT dataset used in Chapter 5. The 3DPS method using

Setting	Head	Shld	Elbow	Wrist	Hip	Avg
Deep3DPS (DeeperNet)	89.6	56.5	50.6	54.3	42.9	58.8
Deep3DPS (RGB)	**93.7**	74.9	69.6	71.8	66.6	75.3
Deep3DPS (Depth)	91.0	75.0	69.1	68.0	63.2	73.2
Deep3DPS (RGBD)	93.4	**77.0**	**71.5**	**73.7**	**69.1**	**76.9**
+Auxiliary tasks	91.4	72.1	64.9	68.4	63.5	72.1
3DPS (IHOG+HDD)	90.8	74.2	62.2	63.4	57.5	69.6
Insafutdinov et al. [Insafutdinov 2016][2]	91.1	53.7	47.5	50.1	38.4	56.2
Yang and Ramanan [Yang 2013][3]	30.4	35.2	19.6	24.3	16.7	25.2

Table 6.1: Pose estimation results of several single-view approaches using PCK metric.

the pre-trained network of [Insafutdinov 2016] as body part detector, referred to as *DeeperNet*, achieves a better performance compared to the full *DeeperCut* approach from [Insafutdinov 2016] that estimates the body poses via a joint optimization across all people. These results indicate that in an environment with many visually similar surfaces, a 3D deformation model, even with tree-structured graph, is more reliable than a fully connected deformation model which relies on appearance and 2D displacement constraints. Fine-tuning the network on the SV-RGBD-CT dataset significantly improves the results (*Deep3DPS (RGB)*: 75.3% vs. 58.8% PCK), as it allows the network to adapt its representation for learning a better encoder for such an environment. We have also trained the network to detect body parts using only depth data, *Deep3DPS (Depth)*, which achieves competitive results. The best performance is obtained when the network relies on both color and depth images: the resulting model, called *Deep3DPS (RGB-D)*, is therefore used as single-view pose estimator during the rest of the experiments. Exemplary score maps generated by *Deep3DPS (RGB-D)* are shown in Figure 6.4. It can be seen that the detector generates very few false detections, which is crucial for the multi-view merging method that relies on these detections (Φ^{conf}). But, we observe that on this data, training the network for the auxiliary tasks suggested in [Insafutdinov 2016], namely location refinement and regressing to other parts, degrades the performance. We believe that this is due to both a much smaller training set and to the strong foreshortening of the body parts because of the top views of the cameras.

As baseline, we report the results of the 3DPS model from Chapter 5, which relies on a 3D pairwise deformation model similar to our approach, but with handcrafted color and depth features. Our best model improves the performance over this baseline by ∼7% on the same experimental setup. This highlights the benefits of deep ConvNets in constructing more discriminative body part detectors by automatically learning feature representations and also incorporating a wider context. Evaluation of state-of-the-art RGB models [Insafutdinov 2016, Yang 2013] trained on common computer vision datasets shows that they do not generalize to the OR environment due to both loose clinical clothes and the presence of many visually similar surfaces[4].

[4]Note that the Kinect tracker [Shotton 2012] cannot be evaluated quantitatively, as it is not available off-line and cannot be run on individual frames. Qualitative results for the Kinect skeleton tracker, and

Figure 6.5: Examples of multi-view pose estimation results. Each row shows a multi-view frame. The 3D skeletons obtained after multi-view energy optimization are projected to the views.

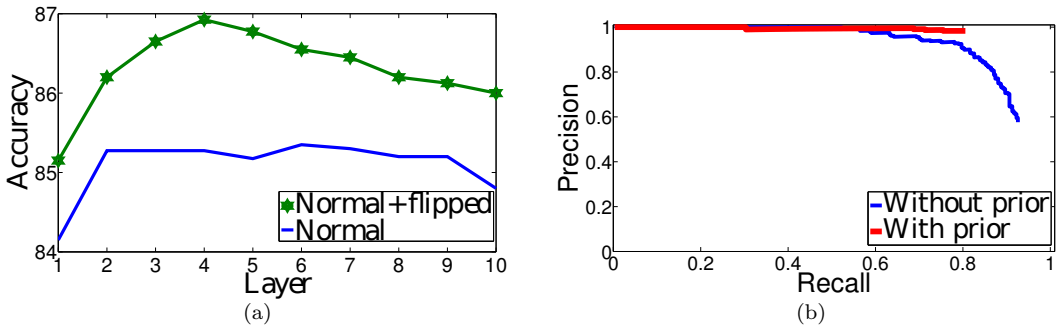
6.3.2 Random Forest Based Prior

The Deep3DPS (RGB-D) model is applied to detect skeletons in each view of the MV-RGBD-CT dataset separately. The skeletons are back-projected into 3D and transformed into a common reference frame. We use these 3D skeletons to train our prior, as explained in section 6.2.3. We also augment the data by flipping the skeletons to exchange the left-right body parts. Due to the small size of the training set, we learn 100 shallow trees with a maximum depth of 10. Figure 6.6(a) shows the detection accuracy of

the methods [Yang 2013] and [Insafutdinov 2016] are presented in Chapter 1.

Part name	One view			Two views			Three views		
	initial	after opt.	opt.-Φ^{depth}	initial	after opt.	opt.-Φ^{depth}	initial	after opt.	opt.-Φ^{depth}
Head	7 ± 4	7 ± 4	7 ± 4	6 ± 3	6 ± 3	6 ± 3	5 ± 2	5 ± 2	5 ± 2
Neck	7 ± 4	7 ± 4	7 ± 4	5 ± 3	5 ± 3	5 ± 3	4 ± 2	4 ± 2	4 ± 2
Shld	25 ± 25	19 ± 16	21 ± 18	22 ± 16	15 ± 10	19 ± 13	14 ± 14	10 ± 7	12 ± 9
Hip	28 ± 22	27 ± 19	28 ± 20	24 ± 13	23 ± 13	24 ± 14	18 ± 10	17 ± 9	18 ± 10
Elbow	31 ± 22	27 ± 19	30 ± 21	30 ± 18	23 ± 15	27 ± 18	19 ± 14	16 ± 11	18 ± 14
Wrist	42 ± 34	32 ± 21	35 ± 24	34 ± 22	25 ± 16	28 ± 18	24 ± 18	18 ± 13	20 ± 15
avg†	32 ± 26	$\mathbf{26 \pm 19}$	29 ± 21	28 ± 17	$\mathbf{22 \pm 14}$	25 ± 16	19 ± 14	$\mathbf{15 \pm 10}$	17 ± 12

Table 6.2: Mean and standard deviation of 3D part localization error in centimeter. The results are presented as a function of the number of supporting views used to generate the initial 3D skeletons (distribution: 1 view: 30%; 2 views: 43%; 3 views: 27%). † The average is computed for all parts except the head and neck since they are not included in the optimization. See Section 6.3.3 for details.



(a) (b)

Figure 6.6: (a) Accuracy of the RF-based prior in detecting spurious skeletons. (b) Precision-recall curves for 3D clinician detections.

the RF-based method in distinguishing spurious detections. The results show that the method always detects valid skeletons with an accuracy superior to 84%. One observes that augmenting the training set by flipping the skeletons consistently improves the results by enabling the forest to learn a richer prior model that is not confused by the noisy side detections. The results show that auto-context enhances the performance up to the fourth iteration and then tends to overfit. We therefore use the output of the RF trained on the augmented training set at the fourth layer to identify spurious skeletons during the remaining evaluations. Figure 6.3 illustrates the results of the approach on sample frames from the MV-RGBD-CT dataset and also on a few frames from the MV-RGBD-CArm dataset, which has been recorded in a different room from totally different viewpoints. It can be seen that the method has correctly identified spurious skeletons in both datasets and generalizes well. The proper generalization is due to the fact that the reference frame is defined on the floor at the center of the operating table, the main element in any operating room.

6.3.3 Multi-view 3D Person Detection and Pose Estimation

We set T_s to 30 cm to avoid merging skeletons across persons who are close to each other. We evaluate 3D clinician detection using the precision-recall curves. We accept a detection as a true positive if the distance between the ground-truth and the detection is below 30 cm for both the head and neck. We use the fusion algorithm described in Section 6.2.4.1 to generate a set of 3D skeleton candidate per frame. Figure 6.6(b) shows the 3D clinician detection results after multi-view fusion with and without the RF-based prior. The high precision obtained by our method when the OR prior is used indicates the high quality of the retrieved skeletons. To optimize part positions based on multi-view cues, we generate four label sets $\{(n, s) : (3, 50),\ (5, 10),\ (7, 2),\ (7, 1)\}$, where the step sizes are in centimeter. We solve the optimization in four iterations by going from a large and coarse search space towards a small and fine search space, which allows us to more efficiently explore the 3D space. The parameters used in all experiments are $\lambda_1 = 2$ and $\lambda_2 = 0.5$, that are selected using grid search over a set of 50 frames from the MV-RGBD-CT dataset. The mean and standard deviation (STD) of the 3D Euclidean distances between the predicted body part positions and the ground-truth positions are used to evaluate 3D body part localizations.

In Table 6.2, we present the evaluation results for multi-view body part localization as a function of the number of supporting views. Please note that since the head and neck localization errors are close to the expected error in low-cost RGB-D cameras, we do not update these two joints during our optimization. This table presents localization errors for the initial 3D skeletons obtained by the fusion algorithm and the error after performing the multi-view optimization. One can notice that the proposed multi-view fusion method correctly associates skeletons across views by consistently reducing the localization errors as the number of supporting views increases. However, we observe that if we ignore the left and right labels of the detections and assign the label based on shoulder distances with ground-truth, the localization errors decrease by ∼10 cm for skeletons with one or two supporting views and ∼3 cm for skeletons with three supporting views. These results indicate that the side detection in individual views is not very reliable. But, if a person is detected in all views, the proposed voting algorithm can make a more reliable prediction. The multi-view optimization significantly reduces the localization error for skeletons with any number of supporting views. Interestingly, the optimization improves the results even for skeletons with one supporting view by properly incorporating the depth-based reprojection costs and detection confidences. The deep RGB-D part detector is the main driver of the optimization. To evaluate the effect of the depth-based reprojection cost, we also report the results without this term in column 'opt.-Φ^{depth}'. The drop in performance highlights its importance. The 2D projections of 3D poses obtained using the proposed multi-view optimization are shown for a few frames in Figure 6.5.

6.4 Conclusions

In this chapter, we propose a multi-view RGB-D approach for detecting and estimating the body part positions of medical staff in 3D. A ConvNet-based body part detector combined with a 3D pairwise deformation model is used to recover body poses in each view. A method based on multi-layer random forests is then proposed to automatically learn a priori information about the OR and remove spurious detections per view, which allows us to reliably detect the body poses of persons in the scene. Then, these detections are back-projected to 3D and merged across views. Finally, a novel optimization function is introduced to update the part positions by relying jointly on the body part confidence maps, depth data and multi-view cues. The method has been quantitatively evaluated on a new multi-view dataset acquired during live surgeries. Experimental results show significant improvements over state-of-the-art methods for the task of single-view pose estimation in multi-person scenarios, indicating the benefit of combining deep part detectors and 3D pairwise constraints in building robust models. The multi-view formulation also achieves very promising results showing the benefit of the deep ConvNet detector and of depth data for correctly driving parts towards their optimal locations.

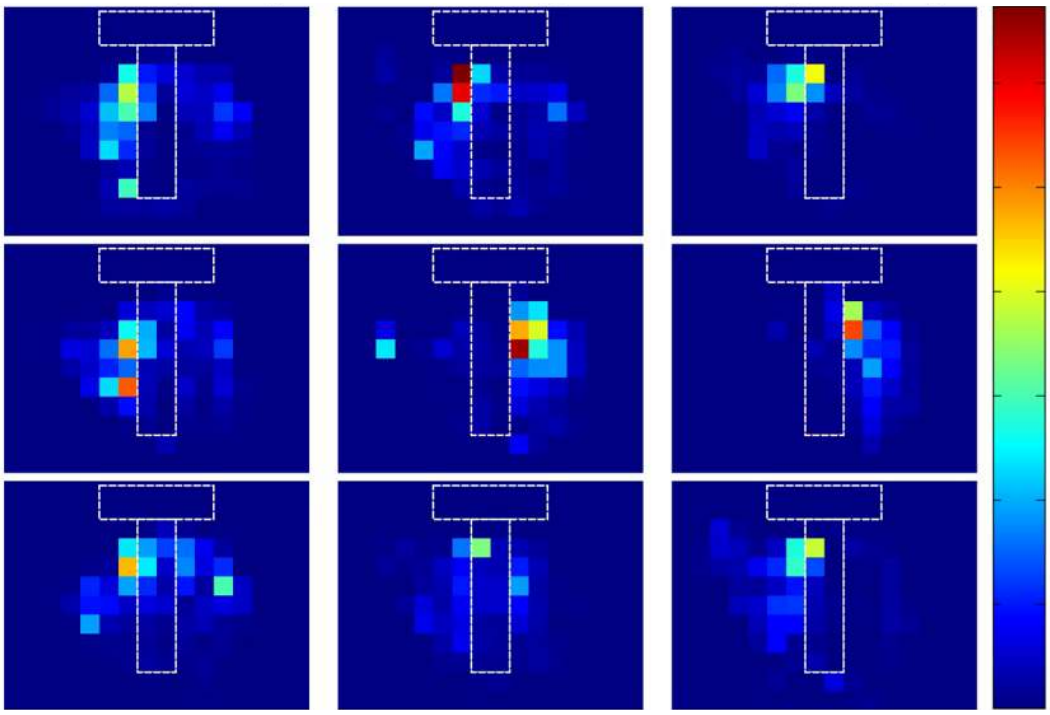
7 Potential Applications of Clinician Detection and Pose Estimation for the Operating Room

Chapter Summary

In this chapter, we demonstrate how human detection and pose estimation can be used to address several applications for the operating room. To detect and estimate the poses of the clinical staff, we use our multi-view model presented in Chapter 6. Then, the estimated 3D body poses are applied to develop solutions for several applications, namely room occupancy analysis, smart video browsing and radiation exposure estimation. For all these applications, we use a fully calibrated multi-view RGB-D camera system to capture the operating room from complementary views during real surgeries.

7.1 Room Occupancy Analysis

We apply the multi-view model of Chapter 6 to extract room occupancy maps per surgery, hypothesizing that the spatial room usage patterns vary among different types of procedures and are similar for different instances of the same type. To this end, with the help of our clinical partner, we have selected three types of fluoroscopy-guided surgical interventions, namely vertebroplasty, drainage and lung biopsy. These interventions are chosen because they are frequently performed in the interventional radiology department of the University Hospital of Strasbourg and their time durations, which are usually about one hour, are similar. We have recorded three instances of each kind in an operating room that is equipped with an inter-operative CT scanner device and a mobile C-arm device.



Vertebroplasty Drainage Lung biopsy

Figure 7.1: Room occupancy heat maps for three instances of three types of surgeries: Vertebroplasties, Drainages and Lung biopsies. The horizontal dashed boxes indicate the CT scanner device and the vertical ones indicate the operating table.

We apply our multi-view model to detect clinical staff at 1 fps. The 3D positions of the heads in the detected skeletons are used to indicate the staff's locations in the room. The floor is then discretized into squares of 30×30 cm. We accumulate the number of persons detected in each square across the entire surgery, to form a room occupancy heat map. Figure 7.1 shows the heat maps for the recorded surgical interventions. The heat maps are normalized across all surgeries. Figure 7.2 shows a sample image for each sequence that is used to compute the heat maps.

One can notice that even with this coarse representation, it is possible to identify a common pattern of room usage for the same kind of procedure and to distinguish different types of surgeries. One observes that clinicians and staff are mainly located at one side of the operating table and that they are often moving in a small area. In vertebroplasties and drainages, we observe that medical staff are using both side of the table. This is because the mobile C-arm device is used during these procedures, which requires at least one nurse at the other side of the operating table to manually adjust the device. In vertebroplasties, the nurse is more often present at the other side of the table, since the C-arm device is used more frequently compared to drainages.

Vertebroplasty Drainage Lung biopsy

Figure 7.2: Samples images from the sequences used to compute the room occupancy heat maps that are presented in Figure 7.1. Images are shown in the same order as the heat maps.

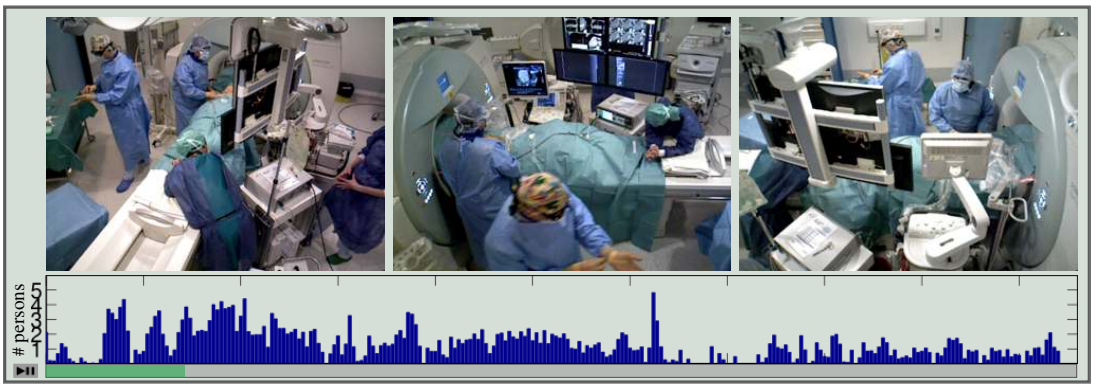


Figure 7.3: Video player tool. The tool shows a bar graph to indicate the number of persons per minute. The bar graph is aligned with the video progress bar, such that the advancement of the progress bar denotes the average number of persons at the corresponding time step.

7.2 Smart Video Browsing

Nowadays, camera recording systems are commonly used in operating rooms to record the performed procedures, mostly for archiving and educational purposes. The recordings are often stored per day, which makes it difficult to browse through the generated files. In order to facilitate browsing though these lengthy videos, we propose to apply our clinician detector to collect statistics about the number of persons present in the room, which is highly informative of the activities taking place.

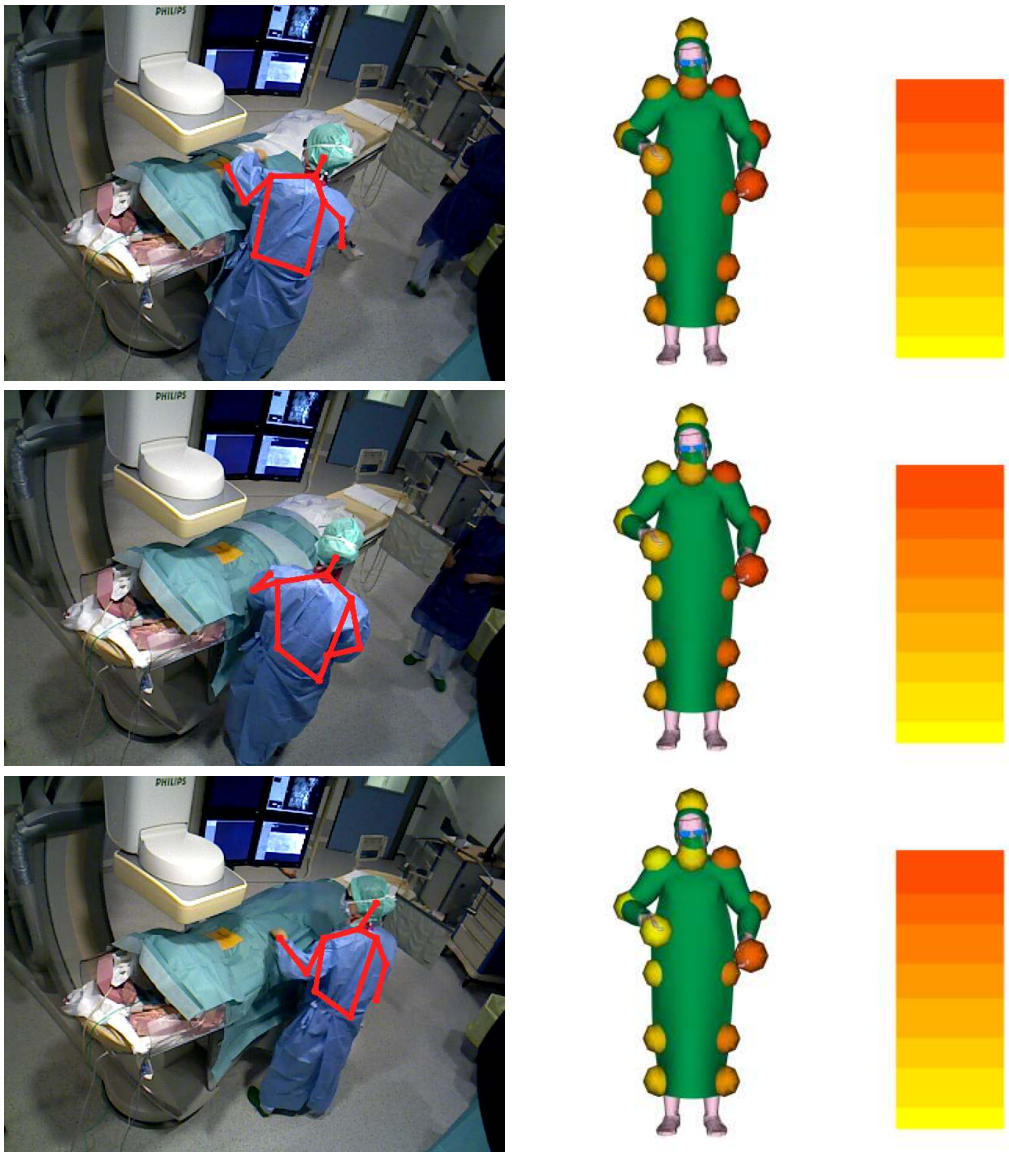
Figure 7.3 shows a video player tool that can be designed to facilitate browsing through lengthy OR video files. Such a player can allow the user to browse through a video while simultaneously showing a bar graph for indicating the average number of persons in the room. To compute the bar graph, the number of persons per frame can be detected using our clinician detection model from Chapter 6 or Chapter 5 depending on the number of views. The bar graph can then be aligned with the video progress bar, such that the advancement of the progress bar indicates the average number of persons at the corresponding time step. For example, the bar graph shown in Figure 7.3 is computed for a full day recording that was captured using a calibrated multi-view camera system at 20 fps. Each bar in this graph represents the average number of persons per minute.

The knowledge of the number of persons per minute allows the user to identify time intervals when the room is empty. This information can also help the user to locate in the video certain types of surgeries and actions efficiently. For example, during a vertebroplasty surgery, at least two medical staff are required (one clinician and one helping nurse operating the C-arm device); and transferring the patient from the gurney to the operating table requires more than two persons. In Figure 7.3, one can observe that no patient is brought to the operating room on a gurney during the last few hours of the day as the number of persons in the room remains below two.

7.3 Radiation Exposure Estimation

As mentioned in Sections 1.2.3 and 4.1, during fluoroscopy-guided surgeries medical staff are regularly exposed to harmful ionizing radiation. Longterm exposure can lead to serious negative effects on the body [Vanhavere 2008]. Currently, medical staff are wearing dosimeters to measure the accumulation of the exposure over time. As each person in the OR wears a single dosimeter at the chest level and the radiation risk varies at different parts of the body, the estimation of the amount of radiation absorbed at different parts of the body is required to correctly assess full-body exposure [Carinou 2011,Loy Rodas 2015]. To that end, we use our multi-view model for estimating 3D body poses of medical staff in combination with the radiation simulation system of [Loy Rodas 2015] for computing the amount of x-ray doses received by each body part.

Figure 7.4(a) shows images from a recording sequence that has been captured during a fluoroscopy-guided surgery. The rows show images from the beginning, the middle and the end of the sequence. To estimate the radiation exposure, we use the simulation model of [Loy Rodas 2015]. Since in our case we do not have the parameters of the C-arm



(a) (b)

Figure 7.4: Estimation of the radiation exposure for frames in the beginning (top row), the middle (middle row) and the end (bottom row) of a sequence: (a) the detected upper-body poses and (b) the estimated radiation exposure per body parts. Each sphere represents a body part and its color denotes the amount of received dose. A clinician mesh in a default posture is shown in each image as a reference. Note that the values are normalized across all frames (red indicates higher dose).

device, we use the model with default parameters. Given parts' 3D positions, this model is used to estimate the amount of doses received by each body part. These values are then normalized across the sequence to bring all values into the range of [0,1], because the simulation model estimates relative radiation exposure up to a scale factor. Figure 7.4 (b) shows the estimated exposure for the selected frames. The exposure is shown for

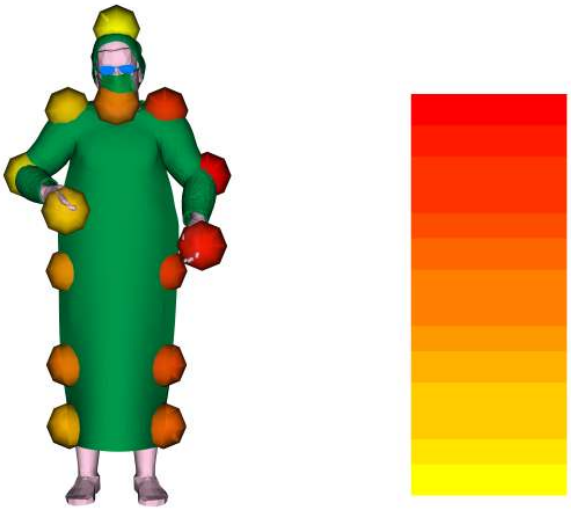


Figure 7.5: Accumulation of radiation exposure. We show a color-coded radiation exposure per body part, which is accumulated over the entire sequence. A clinician mesh in a default posture is shown as a reference.

each body part and a clinician mesh in a default posture is shown as a reference. Note that we estimate the radiation risk for all body parts including the left and right knees as well as ankles although our pose estimation model only predicts upper-body poses. We estimate the positions for the knees and ankles based on the 3D positions of the left and right hips and on an average model that indicates body part lengths in 3D. This is possible for clinical staff because they are always standing and the plane coordinate of the floor can be obtained from the predefined room reference frame.

To compute the accumulation of the radiation exposure over time, we construct a trajectory for the clinician shown in Figure 7.4. The trajectory of the clinician is built by using a greedy algorithm that relies on 3D distances between the heads in consecutive frames. The accumulation of radiation exposure over the entire sequence is presented in Figure 7.5. The accumulation of radiation doses is separately computed for each body part. The left side of the body of the physician is exposed more to the radiation as it is closer to the radiation source.

7.4 Chapter Summary

In this chapter, we apply our approach from Chapter 6 to tackle several applications for the operating room. The results show that detecting clinical staff and estimating their poses can provide important information for different applications. In addition, since our model only relies on camera sensors, it permits us to show these applications on data from real surgeries without disrupting OR workflows.

8 Conclusions and Future Work

Chapter Summary

In this chapter, we conclude by briefly summarizing the contributions of this dissertation. We also discuss the current limitations and possible directions to overcome them. In addition, we outline possible directions for future work on human detection and pose estimation.

8.1 Summary

One of the principal goals of this work was to develop the methods and algorithms necessary for detecting persons in operating rooms and recovering their body part configurations. Even though vision-based people detection and pose estimation are challenging tasks in such an environment, other approaches such as the ones using body-worn markers are not practical options for real surgeries. We have developed our methods for clinician detection and pose estimation by relying on compact RGB-D camera sensors (*Asus Xtion Pro*). Such cameras can be conveniently installed in the OR to record real surgeries. Furthermore, the complementary information provided by the color and depth images enables us to tackle the visual challenges present in the OR.

We introduced in Chapter 4 a new Markov random field energy optimization in order to consistently track upper-body poses in an RGB-D sequence. We used [Buys 2013] as a body part detector to detect body parts in each frame separately. Then, an MRF energy optimization was used to incorporate part detection confidences along with kinematic and temporal smoothness constraints for estimating temporally consistent poses over the

complete sequence. The evaluation results showed that the proposed method is robust to occlusions, presence of multiple persons and failures of the detector that often occur in visually challenging ORs.

In Chapter 5, building on the pictorial structures framework [Felzenszwalb 2005], we presented a novel method for human pose estimation on a pair of aligned color and depth images. We proposed to construct a robust model by using a body part detector based on both color and depth images, and by using 3D pairwise constraints. We also introduced an efficient algorithm to reduce the size of 3D state space and make exact inference tractable. In addition, a new descriptor was proposed for depth images. We showed that the RGB-D detector and the 3D pairwise constraints are essential for detecting people and estimating their poses in cluttered and crowded environments such as ORs. We also demonstrated that the proposed model significantly improves the results for both the tasks of human detection and pose estimation over the dominant approaches [Yang 2013] and [Felzenszwalb 2010].

In Chapter 6, we introduced a multi-view approach for people detection and pose estimation. We revisited the single-view 3DPS model from Chapter 5 and extended it to use a deep convolutional network part detector in order to take a wider image context into account for part detection. The model was called *Deep3DPS*. An approach was also presented to automatically model a priori information about the environment. Finally, a multi-view energy optimization approach was proposed to estimate 3D body poses by incorporating body part detection confidences and depth data across all views. We solved the optimization iteratively to efficiently explore the 3D space. We demonstrated that the Deep3DPS model achieves significant improvements over state-of-the-art methods [Yang 2013] and [Insafutdinov 2016] for single-view pose estimation. We also showed that the proposed multi-view approach reliably detects people and estimates their 3D poses.

Finally, in Chapter 7, we presented the use of our multi-view approach described in Chapter 6 for several applications in the operating room. The applications were room occupancy analysis, smart video browsing and estimation of the accumulation of radiation exposure per body parts.

This is the first work that addresses the problem of human detection and pose estimation on data recorded during real surgeries. We explored different directions to exploit different types of data, namely single-view images, multi-view images and temporal sequences for obtaining robust models. This allowed us to propose approaches with interesting properties and to achieve impressive results on real data. However, due to time constraints, we could not explore all the ideas that have emerged during this work. We would like to discuss some of these ideas in the following section in the hope of coming back to them in future work.

8.2 Discussion and Future Work

Joint multi-person pose estimation. The proposed approaches estimate body poses of each person separately and rely on 3D pairwise constraints to resolve potential

ambiguities among detections belonging to different individuals. It will however be interesting to jointly estimate the body poses of all individuals in the scene. An elegant model has been presented in [Pishchulin 2016] to address this problem, but optimizing such a model is NP-hard and even the approximate optimization for one image takes around 72 hours, which makes it infeasible for a practical application. Their follow-up work in DeeperCut [Insafutdinov 2016] significantly reduces the optimization time. As the approach uses appearance to filter part detection candidates and also to speed up the optimization, it does not generalize well to complex and visually similar environments such as ORs. The 3D information provided by the depth data can be leveraged to generate a better candidate set and also to further speed up the optimization. We envision that a robust and efficient model for joint pose estimation of multiple persons could significantly improve results in crowded ORs.

Body part detection in 3D. In Chapters 5 and 6, we have shown that using depth data always improves the performance of the part detector. However, the depth image is currently treated similarly to the color channels in the body part detectors. It will be interesting to use the depth data to reconstruct the 3D point cloud and develop 3D body part detection models especially in multi-view setups. We envision that such models will have less problems with the foreshortening of the parts and will also speed up the run-time by removing the need for multi-scale part detection.

Context-aware pose estimation. In Chapter 6, we have demonstrated that a priori information about the environment could be used to build robust models. We envision that in modern ORs, signals available from different OR tools and systems could provide information about the context and similarly improve human pose estimation methods. Therefore, we believe that it is an interesting research direction to build human pose estimation methods that incorporate these signals. This will allow to develop context-aware multi-modal pose estimation methods.

Joint training. Our best performance for single-view pose estimation has been obtained using the 3DPS approach with the ConvNet-based body part detector. The deformation model and the body part detector have been trained separately. It would be interesting to learn all parameters jointly. This will allow to have interactions between the part detection and the deformation models to ultimately build a richer model.

Deep deformation model. The impressive results of deep ConvNets for learning robust body part models [Chen 2014, Yang 2016, Insafutdinov 2016] indicate that such a learning paradigm is able to effectively explore training data for building stronger models. Similarly, we envision that current deformation model can be automatically learned and improved by using deep learning.

Human pose estimation combined with other human understanding tasks. In general, human pose estimation is typically studied in isolation from other tasks such as human activity recognition, human behavior analysis and social interaction analysis. But, in practice, there is a high correlation between human body poses and other ways

to study and understand humans. Human pose estimation is often used in the process of building models to perform these tasks. However, improving human pose estimation using the higher-level information from these tasks remains largely unaddressed. For example, the knowledge of the performed surgical action can impose a strong prior on the set of possible body poses. Future research should therefore focus on closing the loop between human pose estimation and other human understanding tasks in a joint manner.

Training data. Current research on human pose estimation is mainly focused on developing supervised approaches to address this problem. However, the ability of the supervised methods are bounded by the quality and the availability of labeled data. Collecting a large amount of data and generating manual annotations are very tedious tasks. The emergence of Amazon Mechanical Turk (AMT) has allowed to distribute the task among many people across the globe. We should however note that preparing the data as well as the annotation tool and ensuring the quality of the annotations are still very demanding tasks to perform. More importantly, in the case of OR data, due to privacy regulations it is not always possible to use AMT. We envision two directions to tackle the problem of shortage of labeled data: synthetic data generation and semi-supervised learning.

Synthetic data generation. With the availability of many 3D body shape models and powerful rendering algorithms, it is possible to generate a substantial number of synthetic images to cover various poses. In order to ensure proper generalizations of models trained on synthetic data to real data, the synthetic data generation process should respect both the range of possible body poses and their appearance in the environment. As shown in [Shotton 2012], such synthetic data can be used to learn robust pose estimation models. We believe that combining small annotated datasets with a large and diverse synthetic dataset is a promising direction to improve pose estimation results. A challenging problem would however be to construct a proper model for loose clinical clothes in order to generate realistic images.

Semi-supervised learning. Semi-supervised learning is a class of learning techniques that makes use of unlabeled and labeled data to build models. Semi-supervised learning has recently enjoyed a great success in the field of rigid object detection [Vijayanarasimhan 2014, Caicedo 2015], while it has remained largely unexplored in the context of human pose estimation. Thus, in the future, we would like to explore this direction to obtain more robust models by relying on a small amount of labeled data and a large amount of unlabeled data recorded using our camera recording system.

List of Publications

[1] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Temporally Consistent 3D Pose Estimation in the Interventional Room Using Discrete MRF Optimization over RGBD Sequences.* In Information Processing in Computer-Assisted Interventions, volume 8498 of *Lecture Notes in Computer Science*, pages 168–177. Springer International Publishing, 2014. Supplementary video

[2] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Pictorial Structures on RGB-D Images for Human Pose Estimation in the Operating Room.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, pages 363–370. Springer International Publishing, 2015.

[3] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Articulated clinician detection using 3D pictorial structures on RGB-D data.* Medical Image Analysis, vol. 35, pages 215 – 224, 2017. Supplementary video

[4] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms.* In Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2017. to appear. Supplementary video

Appendix Part I

A Datasets

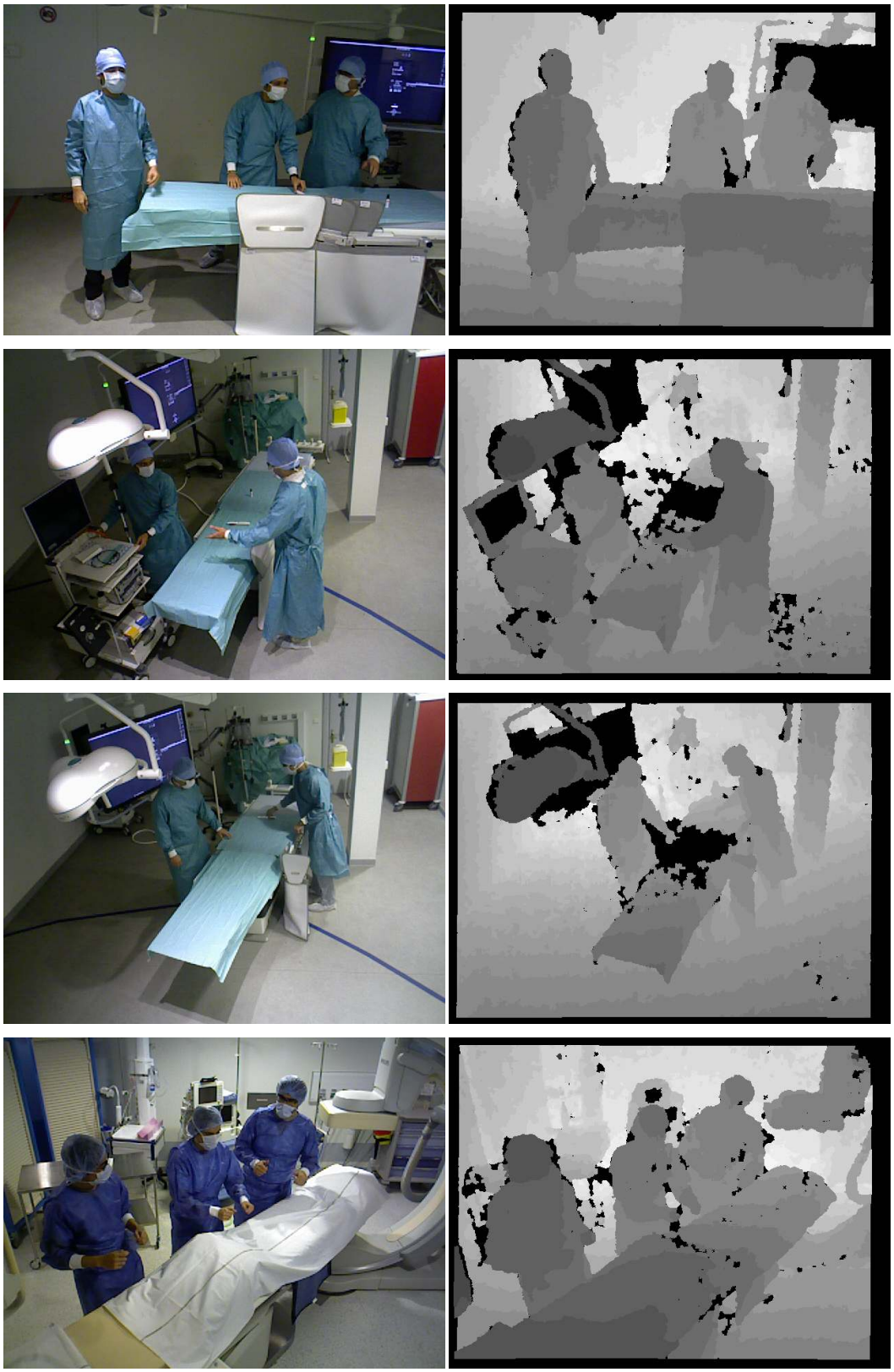
Throughout this work, we have generated several datasets, namely SV-RGBD-Seq, SV-RGBD-CT and MV-RGBD-CT. Next, we summarize the statistics about these datasets and show sample frames for each one.

A.1 SV-RGBD-Seq Dataset

This dataset includes seven RGB-D sequences that have been recorded using an *Asus Xtion Pro* camera. We have captured different simulated medical operations in two different operating rooms. All sequences have been manually annotated to provide ground-truth positions for upper-body skeletons. This dataset has been used to perform the evaluation in Chapter 4.

ID	#Frames	#Persons	Room
S1	50	2	OR1
S2	100	2	OR2
S3	100	3	OR2
S4	110	3	OR2
S5	200	2	OR2
S6	200	2	OR1
S7	200	3	OR1

Table A.1: Presentation of the *SV-RGBD-Seq* dataset (sequence IDs, number of frames and room IDs).



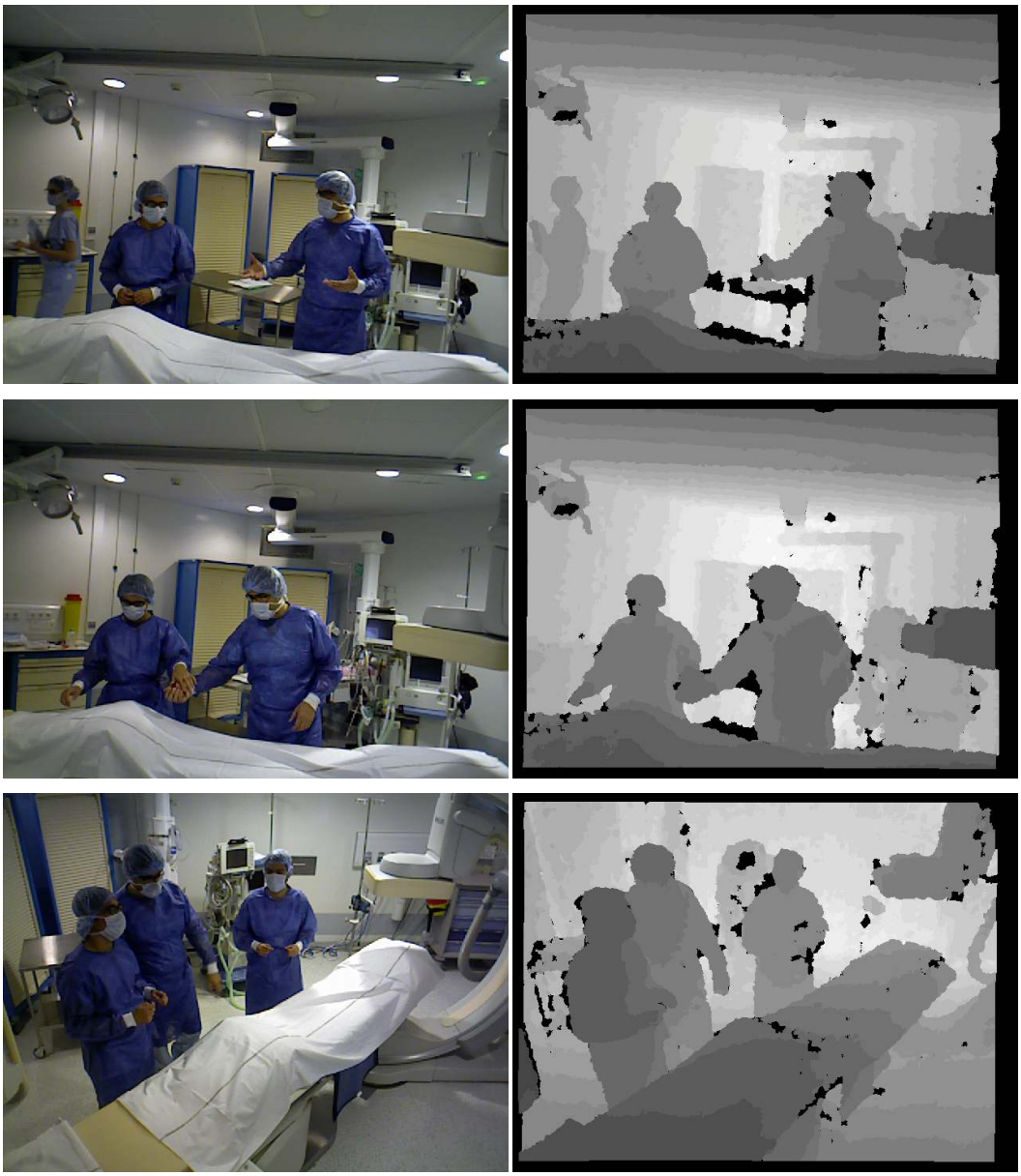


Figure A.1: Sample RGB-D frames from the SV-RGBD-Seq dataset. A sample frame is shown for each sequence. The top three rows show frames from sequences recorded in *OR1*, and the rest of the frames are recorded in *OR2*.

A.2 SV-RGBD-CT Dataset

We have recorded all activities in an operating room containing an inter-operative CT scanner using an *Asus Xtion Pro* camera. The camera position has been changed among three possible locations to capture the room from different viewpoints. From a set of seven half-day recordings, we have constructed a dataset by manually annotating every 500th frame. Two types of annotations have been provided. First, we have annotated all clinical staff with bounding box annotations for the upper-body. A bounding box is also labeled with a *difficult* flag if the head or more than 50% of the upper-body of the person is occluded. Second, we have annotated upper-body poses for clinical staff whose head and more than five upper-body joints are visible. The upper-body joints are the neck, left and right shoulders and hips, as well as left and right elbows and wrists. This dataset has been used to learn model parameters in Chapters 5 and 6. Evaluation in Chapter 5 has also been performed using this dataset.

Half-day	View-id	#Frames	Annotations		
			#Poses	#Bounding boxes	
				Normal	Difficult
1	1	131	216	258	22
2	1	255	306	347	78
3	2	173	277	505	127
4	2	221	278	633	103
5	3	242	349	454	33
6	3	291	350	320	37
7	3	138	215	506	76
		1451	1991	3023	476

Table A.2: Presentation of the SV-RGBD-CT dataset. The dataset includes bounding box and upper-body pose annotations.

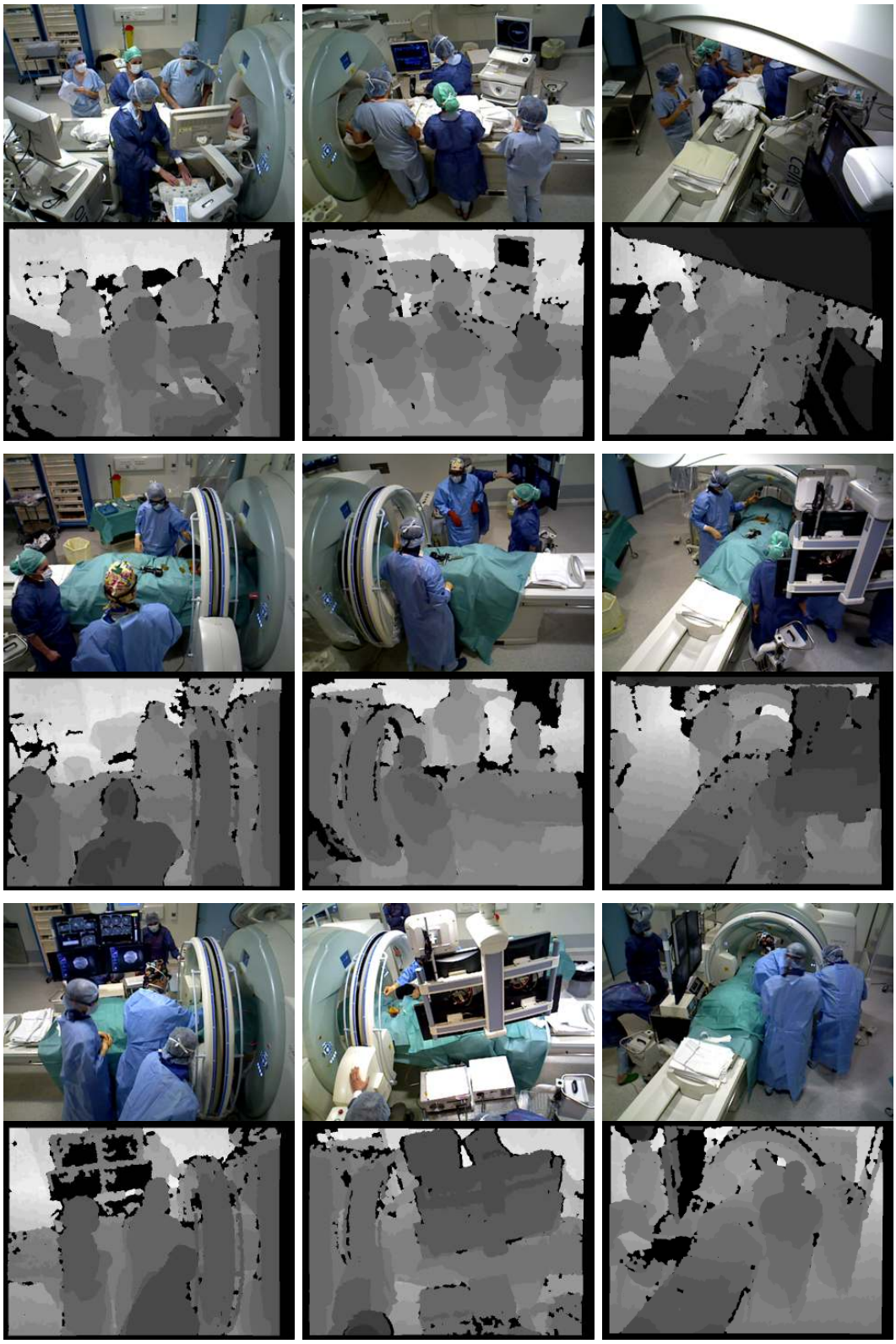
Figure A.2: Sample RGB-D frames from the SV-RGBD-CT dataset. Each column shows images from one of the three possible viewpoints used to capture this dataset.

A.3 MV-RGBD-CT Dataset

We have recorded all activities in an operating room for four days using a multi-view camera system. The camera system consists of three *Asus Xtion Pro* RGB-D cameras and has been fully calibrated using a method similar to [Loy Rodas 2015]. Ground-truth annotations have been provided for the upper-body poses of all members of the clinical team. This dataset has been used in Chapter 6.

Fold no	# Frames	# of persons		
		Visible in 1 view	Visible in 2 views	Visible in 3 views
1	157	124	111	153
2	255	99	166	95
3	235	57	118	145
4	124	46	139	125
	741	326	534	518

Table A.3: Presentation of the MV-RGBD-CT dataset recorded using a calibrated multi-view camera system. The dataset has been annotated for the upper-body poses of clinical staff.



View 1 View 2 View 3

Figure A.3: Sample RGB-D frames from the MV-RGBD-CT dataset. This dataset has been recorded using a three-view RGBD system.

B Résumé en français

B.1 Contexte

Le service de chirurgie joue un rôle essentiel dans le soin des patients au sein d'un hôpital. Il est chargé des consultations préopératoires, des chirurgies et du suivi postopératoire. Les acteurs principaux de ce service sont les chirurgiens et le personnel médical qui collaborent entre eux, prennent des décisions et exécutent les actions nécessaires pour accomplir ces tâches. La salle opératoire est ainsi leur environnement principal de travail, où les procédures qui sont nécessaires au soin des patients sont réalisées à partir de plans préopératoires. Les actions qui sont exécutées ainsi que la façon avec laquelle elles sont exécutées ont un impact direct sur le résultat des traitements. Par conséquent, la modélisation, l'étude et l'amélioration des activités qui se déroulent au sein de la salle opératoire constituent des sujets de recherche importants. Ainsi, des applications comme l'analyse automatique du flux de travail lors d'une chirurgie et des compétences des chirurgiens, permettront l'amélioration du soin des patients et le développement de systèmes intelligents réactifs au contexte pour la salle opératoire.

Localiser le personnel médical ainsi que leurs parties corporelles est fondamental pour accomplir les objectifs mentionnés ci-dessus. Dans le reste ce document, la localisation des cliniciens et la localisation des parties de leurs corps seront respectivement appelées LC et LPC.

Le LC et LPC peuvent fournir des informations très importantes comme l'emplacement

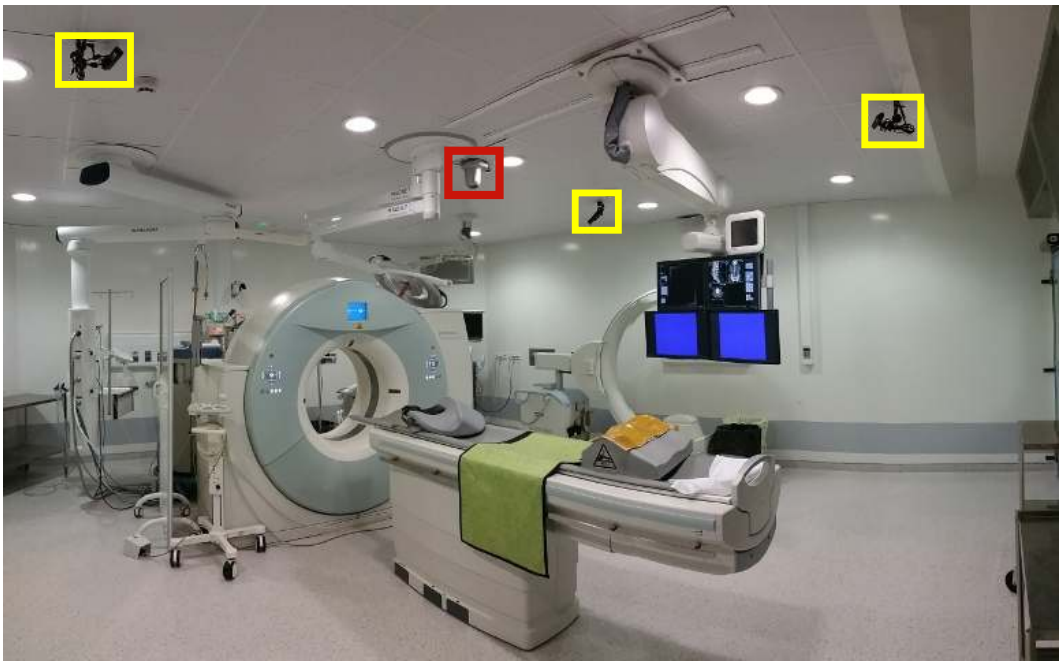


Figure B.1: Vue panoramique d'une salle opératoire au département de radiologie interventionnelle, Nouvel Hôpital civil de Strasbourg. Les positions des caméras sont indiquées en jaune et en rouge. Les boîtes jaunes indiquent les positions des caméras RGB-D que nous avons installé dans la salle pour capturer l'environnement de travail de trois points de vue différents. L'autre est une caméra RGB qui a été installée dans la salle pour de la documentation. Elle se concentre sur le lit près du scanner.

des chirurgiens et du personnel dans la salle, qui est essentiel pour des applications comme l'analyse des activités chirurgicales [Padoy 2009, Twinanda 2016, Bouget 2015], la détection du contexte de la salle opératoire [Meißner 2014, Agarwal 2007] et l'étude du flux de travail d'une intervention [Nara 2015, Agarwal 2007]. De même, d'autres applications comme la collaboration homme-robot [Beyl 2015] et l'analyse automatique des compétences des chirurgiens peuvent y bénéficier [Wanzel 2002, Vedula 2016]. Ces informations peuvent aussi contribuer à l'amélioration de l'analyse des compétences des équipes chirurgicales, en fournissant des informations sur les interactions entre les membres de l'équipe pendant des différentes chirurgies [Reiley 2011].

La détection et l'estimation de la pose des cliniciens dans la salle opératoire nécessite la perception de l'environnement d'une façon qui, non seulement puisse fournir les données nécessaires, mais aussi qui puisse être applicable dans cet environnement si particulier. Deux types de capteurs sont aujourd'hui utilisés pour percevoir l'environnement. D'une part, des capteurs basés sur les caméras infrarouges qui détectent des marqueurs passifs réfléchissants posés sur l'objet d'intérêt. D'autre part, des capteurs basés sur les caméras qui ne nécessitent pas de marqueurs pour percevoir l'environnement ainsi que les objets d'intérêt et le contexte.

Les systèmes à marqueurs peuvent être invasifs et leur installation et utilisation dans

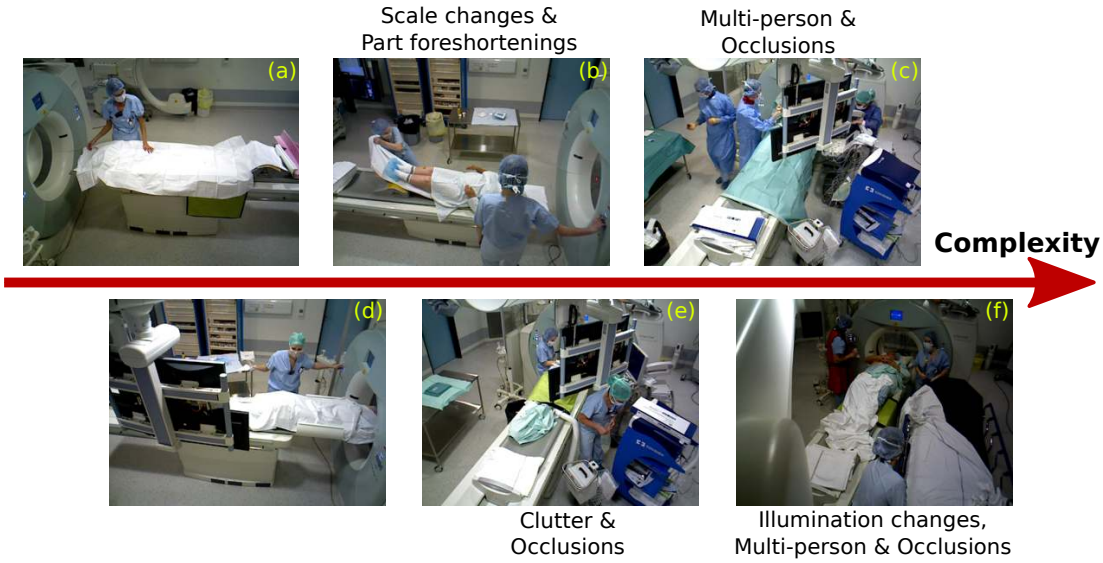


Figure B.2: Exemples d'images d'une salle d'opération montrant des personnes au cours de chirurgies et illustrant certains des défis pour la détection de cliniciens et l'estimation de leurs poses. Les images sont ordonnées de gauche à droite en fonction de leur complexité visuelle.

une salle opératoire difficile. De plus, les processus de stérilisation peuvent compliquer d'avantage la pose des marqueurs. Cependant, des caméras conventionnelles sont souvent déjà installées dans les salles opératoires modernes pour documenter les procédures. Dans le cas contraire, elles peuvent être facilement installées sur le plafond. Les données acquises par ce type de capteur représentent une source d'information riche pour la perception automatique de l'environnement puisqu'elles fournissent une représentation de l'apparence visuelle des objets dans la scène. Dans ce travail, des systèmes d'acquisition multi-capteurs ont été installés dans plusieurs salles opératoires et utilisés pour enregistrer plusieurs chirurgies. La figure B.1 montre notre système d'enregistrement multi-caméras qui est installé dans la salle munie d'un scanner CT au nouvel l'Hôpital Civil de Strasbourg. Ceci nous a permis d'évaluer nos méthodes sur des données provenant de vraies chirurgies.

L'interprétation des données fournies par les caméras n'est toutefois pas simple. En général, la détection humaine et l'estimation de la pose basée sur la vision sont des tâches difficiles qui deviennent encore plus difficiles dans les salles opératoires. Pour illustrer quelques-uns de ces défis, examinons la figure B.2 qui montre des exemples d'images enregistrées lors d'une chirurgie. Ainsi, l'intensité des sources lumineuses qui change au cours de l'opération, l'importante ressemblance entre les couleurs des vêtements et des équipements, les limitations au niveau de l'emplacement des caméras dans la salle et les occlusions auxquelles ces dernières peuvent être soumises font partie de ces contraintes.

Nous proposons d'utiliser des caméras de type RGB-D [Shotton 2012] et de développer des méthodes pour des données RGB-D. Ainsi, ce type de capteurs a plusieurs avantages

Figure B.3: Une paire d'images de couleur et de profondeur capturées à l'aide d'une caméra *Asus Xtion Pro*.

utiles pour le problème qui est considéré ici. Puisqu'ils combinent une caméra couleur avec une caméra capable d'acquérir des cartes de profondeur, ils fournissent deux sources d'informations qui sont complémentaires entre elles. Ceci permet de développer des nouvelles méthodes tout en profitant des progrès récents qui ont été faits dans les approches basées sur des cartes de profondeur. Enfin, l'utilisation des caméras RGB-D facilite aussi la reconstruction 3D de la scène grâce à la fusion des données couleur et profondeur. Dans la figure 3, nous montrons une paire d'images de couleur et profondeur enregistrées avec ce type de capteur.

La détection visuelle des personnes ainsi que l'estimation de leur pose sont des problèmes clefs en Vision par Ordinateur, qui ont été beaucoup abordés dans la littérature ces dernières années [Ramanan 2007, Felzenszwalb 2005, Andriluka 2014, Yang 2013]. Généralement, deux types de méthodes sont proposées. Dans le premier, le corps humain est représenté par un ensemble de parties et l'estimation de la pose est réalisée en deux temps : d'abord, des potentielles parties du corps sont détectées dans l'image, puis, une collection (ou des collections) de parties est trouvée par la vérification de contraintes spatiales entre elles [Kiefel 2014, Insafutdinov 2016, Pishchulin 2016]. Un deuxième type de méthode, les méthodes holistiques, cherchent une correspondance directe entre une image et la position d'une personne et des parties de son corps [Wei 2016, Toshev 2014, Shotton 2012]. Une fonction de correspondance qui peut lire l'image en entier, permet à ce type de méthodes d'apprendre des modèles puissants qui peuvent exploiter le contexte de l'image. Néanmoins, ces modèles font intervenir une large quantité de paramètres et nécessitent alors un jeu de données important pour leur apprentissage. Aussi, elles ne modélisent pas les interdépendances entre les différentes parties de façon explicite, ce qui peut causer des erreurs dans la prédiction de la configuration du corps [Insafutdinov 2016, Yang 2016]. Au contraire, les méthodes basées sur la représentation par parties du corps, nécessitent un jeu de donnés d'entrainement plus petit et peuvent aussi fournir un formalisme puissant pour modéliser de façon

explicite les dépendances entre les différentes parties du corps. Pour ces raisons là, dans ce travail un nouveau modèle basé sur une représentation par parties est proposé pour la détection de cliniciens et pour l'estimation de leur pose.

B.2 Méthodes proposées

Même si les poses des cliniciens peuvent être une source précieuse d'information pour des nombreuses applications dans la salle opératoire, aucun travail n'a encore considéré ce problème clinique. Pour ainsi le faire, trois méthodes sont ici proposées : premièrement, nous présentons une méthode qui utilise un détecteur holistique de parties du corps dans une fonction d'énergie de type champ de Markov, défini sur un ensemble d'images RGB-D pour ainsi obtenir des poses de cliniciens temporellement cohérentes dans des séquences courtes [Kadkhodamohammadi 2014] ; deuxièmement, nous présentons une méthode basée sur des structures pictorielles sur des données RGB-D qui permettent l'entraînement de détecteurs spécifiques à la salle opératoire et aussi d'effectuer l' Estimation de Pose d'un Humain (EPH) par image. Cette méthode utilise des images en couleur et des cartes de profondeur pour développer des détecteurs robustes et aussi pour construire un modèle de déformation plus fiable [Kadkhodamohammadi 2015, Kadkhodamohammadi 2017a] ; troisièmement, nous présentons une approche multi-vue pour l'estimation de la pose de cliniciens, qui incorpore des informations a priori sur l'environnement et optimise les positions des parties en intégrant l'information sur toutes les vues [Kadkhodamohammadi 2017b].

B.2.1 Estimation de pose 3D temporellement cohérente

Le problème du suivi des parties du corps dans la salle opératoire est ici considéré pour la première fois. Ceci peut contribuer à des applications comme la reconnaissance automatique d'activités du personnel médical, l'analyse des compétences des chirurgiens et le suivi de l'exposition aux radiations ionisantes. Dans ce travail, une méthode de classification de parties du corps basée sur des forêts d'arbres décisionnels, inspirée de [Shotton 2012], est utilisée comme détecteur de parties corporelles. Ce détecteur trouve des correspondances entre des caractéristiques visuelles dans la carte de profondeur et des classes de parties du corps. Nous proposons aussi une méthode pour le suivi des parties supérieures du corps dans de séquences courtes, qui applique une optimisation discrète de type champs aléatoire de Markov dans une séquence entière. Les caractéristiques visuelles sont incorporées dans l'optimisation par un terme unaire. La fonction d'énergie proposée impose la cinématique du corps ainsi que des contraintes temporelles afin de rendre la méthode robuste aux ambiguïtés naturelles du suivi et aux possibles erreurs du détecteur dans un environnement si complexe. Ce travail a été publié dans [Kadkhodamohammadi 2014].

La méthode a été évaluée quantitativement sur sept séquences enregistrées dans deux salles opératoires différentes. La figure B.4 résume les performances de notre modèle pour

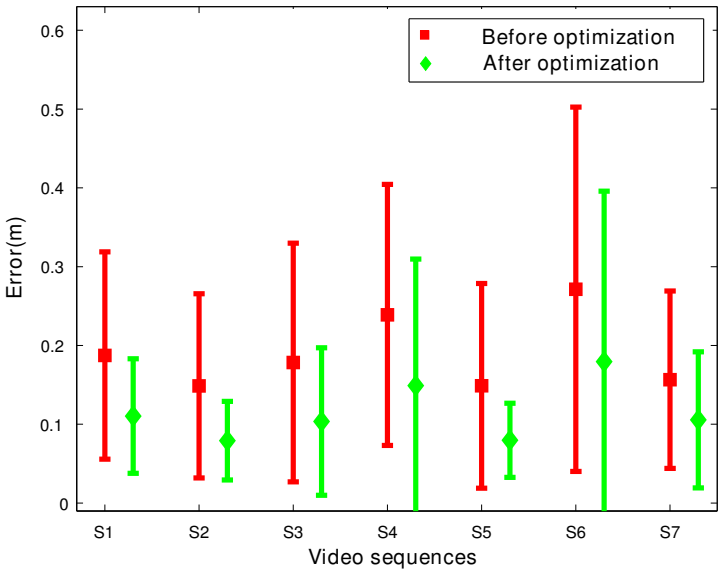


Figure B.4: Erreur moyenne de localisation des parties du corps par séquence. Les erreurs sont affichées avant et après l'optimisation.

la localisation des parties du corps en 3D par séquence. La moyenne et l'écart-type des erreurs de localisation des parties du corps sont indiquées pour chaque séquence avant et après l'optimisation.

Les contraintes cinématiques et temporelles proposées ont diminué l'erreur de façon importante. La diminution de l'erreur dans les séquences où le détecteur des parties du corps échoue est plutôt faible. Mais, même si ce détecteur n'est pas si performant à cause des difficultés visuelles propres à la salle opératoire, il s'agit d'une méthode qui est l'état de l'art pour EPH. Ainsi, des performances prometteuses ont été réussies dans des scènes plus contrôlées comme le bureau ou le salon. Afin d'exploiter au maximum les capacités du détecteur basé sur des forêts aléatoires, il est nécessaire de l'entraîner avec des données réelles provenant de la salle opératoire. Ces données doivent être représentatives des possible variations dans les couleurs des vêtements des cliniciens ainsi que des possibles points de vue dans la salle. Par contre, sa capture est difficile à cause de la quantité importante de données nécessaires à l'entrainement de ce type de méthodes et à cause des régulations d'une salle opératoire. La méthode nécessite en plus des annotations par pixel des parties du corps, ce qui peut être une tâche assez ennuyante à réaliser. En revanche, les méthodes basées sur des structures pictorielles, comme celle qui est introduite ci-dessous, présentent des performances qui sont à l'état de l'art, tout en nécessitant moins de données d'entraînement que les méthodes holistiques. De plus, des annotations des articulations du corps sont nécessaires uniquement et donc aucune annotation par pixel doit être faite.

Figure B.5: Exemples de résultats d'estimation de pose obtenus avec l'approche proposée des structures pictorielles 3D. (Image mieux appréciée en couleur)

B.2.2 Structures pictorielles sur des données RGB-D

L'approche par structure pictorielles formule le problème de l'estimation de la pose d'une personne comme une fonction d'énergie définie sur un graphe à structure arborescente qui consiste de termes unaires et binaires. Les sommets de ce graphe représentent les parties

du corps et les arrêtes indiquent les dépendances par paires qui peuvent exister entre elles. Le terme unaire de la fonction d'énergie incorpore les résultats de la détection de parties du corps et le terme binaire sert à définir un modèle de déformation qui incorpore des contraintes cinématiques du corps [Fischler 1973]. Après les travaux de Felzenswalb et Huttenlocher [Felzenszwalb 2005, Felzenszwalb 2004] en structures pictorielles pour EPH, plusieurs approches ont été proposés afin d'améliorer le détecteur de parties du corps ou le modèle de déformation. La méthode des ensembles flexibles de parties [Yang 2013] (FMP pour ses sigles en anglais) est basée sur des structures pictorielles et obtient des résultats satisfaisants en EPH. Elle figure parmi les méthodes les plus performantes sur plusieurs jeux de données connus et difficiles. FMP étend la méthode de structures pictorielles par l'utilisation de plusieurs ensembles de parties et un modèle de déformation qui apprend les dépendances par paires entre les ensembles.

Afin d'affronter les difficultés propres à la salle opératoire et d'exploiter les bénéfices des données RGB-D, nous proposons deux extensions à la méthode FMP : tout d'abord, des détecteurs de parties du corps robustes et discriminatives sont proposés [Kadkhodamohammadi 2015] ; puis, un modèle de déformation plus fiable a été développé [Kadkhodamohammadi 2017a].

La combinaison de données des modalités couleur et profondeur afin de développer des détecteurs de parties du corps plus performants est ici proposée. Ainsi, l'histogramme de gradient orienté (HOG) est utilisé comme descripteur pour les images en couleur. Pour les cartes de profondeur, trois descripteurs sont utilisés : 1) HOG : le descripteur HOG est souvent utilisé dans des images en couleur et il est ici aussi appliqué à des images de profondeur pour comparer les performances ; 2) Histogramme de vecteurs normaux orientés (HONV) : HONV a été originalement proposé pour la détection d'objets dans des images de profondeur [Tang 2013] ; 3) Histogramme de différences de profondeur (HDD) : ce descripteur innovant introduit dans [Kadkhodamohammadi 2015], encode des changements locaux de profondeur. Il utilise un ensemble de noyaux de convolution afin de capturer des changements relatifs de profondeur et de normaliser les réponses des convolutions pour qu'elles soient invariantes à la profondeur. De plus, les convolutions sont appliquées dans une représentation d'espace échelle et sont discrétisées de façon à que le descripteur soit robuste aux distorsions géométriques et au bruit. Ce travail a été publié dans [Kadkhodamohammadi 2015].

De même, le modèle de déformation des structures pictorielles est étendu pour qu'il puisse inclure des contraintes de plus de deux dimensions [Kadkhodamohammadi 2017a]. Afin de pouvoir effecteur des inférences efficaces et précises dans les structures pictorielles, les conditions suivantes doivent être satisfaites : un graphe cinématique à structure arborescente pour la dépendance par paires, des contraintes de dépendance par paires qui dépendent uniquement de la position, et une évaluation des déformations entre des paires de parties du corps dans l'espace d'état qui s'étend le long d'une grille régulière entièrement connectée. La première condition peut être satisfaite en évitant d'avoir des boucles dans la dépendance entre les parties. Pour remplir les autres conditions, les contraintes

par paires sont définies en se basant sur la distance 2D entre les pixels de l'image dans le cas où il s'agit d'une image 2D. Cependant, ces distances ne sont pas toujours précises à cause du processus de projection 2D. En effet, toutes les méthodes basées sur des structures pictorielles en 2D sont limitées par cette métrique. Contrairement, les approches en 3D requièrent une discrétisation grossière de la grille régulière 3D pour que les exigences de mémoire soient gérables, ce qui a un impact négatif sur leur performance. Dans [Kadkhodamohammadi 2017a], nous proposons une extension 3D aux structures pictorielles en utilisant des données RGB-D. Une méthode capable d'inférer de façon exacte et à basse complexité est présentée. Ceci permet de forcer de contraintes spatiales par paires et d'utiliser une métrique de distance 3D plus réaliste à la place de celle en 2D qui est peu fiable. Cette approche utilise un espace d'état de même taille que celle en 2D, donc, aucune discrétisation est nécessaire. Ce travail a été publié dans [Kadkhodamohammadi 2017a].

Afin d'évaluer notre méthode, un jeu de données manuellement annoté a été généré en enregistrant des chirurgies avec des capteurs RGB-D. Les résultats de l'évaluation montrent que les détecteurs de parties du corps utilisant la profondeur sont beaucoup plus performants que ceux qui utilisent la couleur. La combinaison de deux permet d'améliorer encore plus les résultats. De plus, l'utilisation de l'information de profondeur pour définir des contraintes par paires en 3D améliore les performances dans tous les cas, c'est-à-dire quand le modèle d'apparence dépend uniquement des détecteurs basés couleur, de ceux qui sont basés profondeur et de la combinaison des deux. Les meilleurs résultats sont obtenus quand le détecteur de parties du corps combine HOG sur les données couleur, le descripteur HDD sur les données profondeur et les contraintes par paires en 3D. En conclusion, l'incorporation des informations de profondeur dans les structures pictorielles améliore significativement les performances. La figure B.5 montre des résultats qualitatifs obtenus par l'approche proposée. Les poses estimées du haut du corps sont superposées sur les images originales. Une évaluation extensive de ceci est présentée en [Kadkhodamohammadi 2017a].

B.2.3 Approche RGB-D multi-vue pour l'estimation de pose d'un corps articulé

Dans des environnements encombrés et où évoluent plusieurs personnes comme les salles opératoires, le risque d'occlusion des parties du corps ou même d'une personne en entier est très élevé. Ces occlusions peuvent considérablement dégrader la fiabilité de la détection de la personne et poser des difficultés aux méthodes d'estimation. Nous considérons que dans une salle opératoire, où le volume de travail est connu a priori, les systèmes multi-vues peuvent être utilisés pour percevoir l'environnement à partir de différents points de vue pour ainsi réduire le risque d'occlusions. Suite à nos conclusions à la section B.2.2, nous proposons d'utiliser un système RGB-D multi-vue pour percevoir l'environnement à partir de trois vues complémentaires. Nous introduisons une approche innovante pour profiter des informations a priori sur l'environnement et ainsi faire des

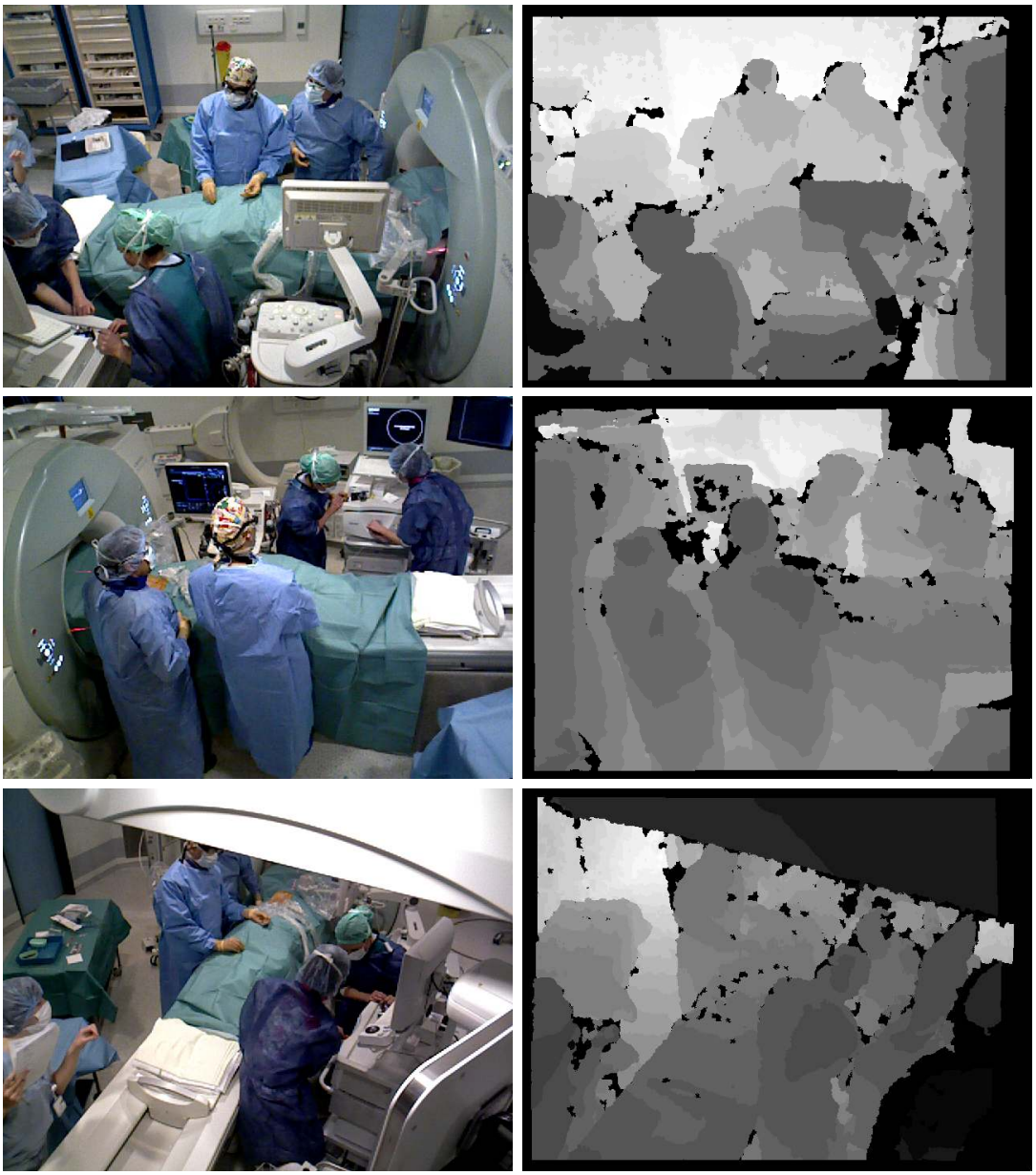


Figure B.6: Paires synchronisées d'images de couleur et de profondeur à partir d'un jeu de données multi-vues. Les images sont enregistrées au cours de chirurgies en direct à l'aide d'un système de multiples caméras RGB-D ayant des vues différentes.

prédictions plus fiables et intégrer des informations provenant de toutes les vues pour localiser les parties du corps. Nous montrons que les avantages d'utiliser de cartes de profondeur dans l'approche multi-vue proposée vont au-delà du fait de fournir des caractéristiques d'apparence améliorées.

L'estimation de pose multi-personne à plusieurs vues est réalisée en deux étapes : premièrement, la détection et la génération de candidats de squelette dans chaque vue

Figure B.7: Exemples de résultats d'estimation de pose multi-vues. Chaque rangée montre un cadre à vues multiples. Les squelettes 3D obtenus après l'optimisation de la fonction d'énergie multi-vue sont projetés sur chaque vue.

; et deuxièmement, la fusion des squelettes à travers les vues et leur raffinement en 3D. Nous étendons également notre approche 3DPS pour utiliser un réseau neuronal convolutif profond comme détecteur des parties du corps sur les images RGB-D pour être moins sensible aux fausses détections qui peuvent introduire des erreurs sur les méthodes de fusion et d'optimisation multi-vues ultérieures. De plus, nous proposons une forêt aléatoire afin de modéliser automatiquement les informations a priori sur l'environnement et de filtrer les fausses détections de squelettes. Une nouvelle fonction d'énergie à vues multiples est introduite pour mettre à jour des positions de pièces 3D basées sur des indices multi-vues, qui a un certain nombre de propriétés intéressantes:

1) l'utilisation d'informations de profondeur pour une estimation plus efficace et fiable des correspondances entre les vues, 2) estimer les coûts de reprojection en fonction de la profondeur plutôt que de la similitude d'apparence, ce qui n'est pas fiable dans les environnements et 3) l'optimisation itérative de la fonction d'énergie multi-vue pour explorer efficacement un espace large en 3D.

Les méthodes de détection de cliniciens et de parties de leur corps en 3D ont été évaluées quantitativement dans un jeu de données multi-vues capturé pendant des chirurgies. Un exemple de ce jeu de données est donné dans la figure B.6. Nous avons évalué quantitativement l'approche proposée pour les tâches de détection humaine mono-vue ainsi que pour l'estimation de pose et pour la localisation multi-vues de parties de corps en 3D. Notre approche obtient des meilleurs résultats que les méthodes les plus avancées sur ce jeu de données difficile pour la tâche de l'estimation de la pose humaine mono-vue. L'approche atteint des résultats prometteurs pour la localisation 3D de la partie du corps. La figure B.7 montre des projections 2D des poses 3D obtenues en utilisant l'optimisation multi-vues proposée pour quelques images. Une évaluation plus détaillée est présentée dans [Kadkhodamohammadi 2017b].

B.3 Perspectives

A ce stade, nous avons introduit des méthodes pour l'estimation de la pose de personnes dans une ou plusieurs vues RGB-D. Le travail futur va se concentrer à l'utilisation de ces méthodes pour des applications dans la salle opératoire comme le suivi de l'exposition corporelle des cliniciens aux radiations ionisantes pendant les interventions guidées par rayons X, la modélisation du contexte de la salle et le parcours intelligent de vidéos.

References

[Agarwal 2007] S. Agarwal, A. Joshi, T. Finin, Y. Yesha and T. Ganous. *A Pervasive Computing System for the Operating Room of the Future*. Mobile Networks and Applications, vol. 12, no. 2-3, pages 215–228, 2007.

[Amin 2013] S. Amin, M. Andriluka, M. Rohrbach and B. Schiele. *Multi-View Pictorial Structures for 3D Human Pose Estimation*. In British Machine Vision Conference (BMVC), September 2013.

[Amin 2014] S. Amin, P. Müller, A. Bulling and M. Andriluka. *Test-Time Adaptation for 3D Human Pose Estimation*. In X. Jiang, J. Hornegger and R. Koch, editeurs, Pattern Recognition, volume 8753 of *Lecture Notes in Computer Science*, pages 253–264. Springer, 2014.

[Andrieu 2003] C. Andrieu, N. de Freitas, A. Doucet and M. I. Jordan. *An Introduction to MCMC for Machine Learning*. Machine Learning, vol. 50, no. 1, pages 5–43, 2003.

[Andriluka 2009] M. Andriluka, S. Roth and B. Schiele. *Pictorial structures revisited: People detection and articulated pose estimation*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1014–1021. IEEE, June 2009.

[Andriluka 2012a] M. Andriluka, S. Roth and B. Schiele. *Discriminative Appearance Models for Pictorial Structures*. International Journal of Computer Vision, vol. 99, no. 3, 2012.

[Andriluka 2012b] M. Andriluka and L. Sigal. *Human Context: Modeling human-human interactions for monocular 3D pose estimation*. In F. J. Perales, R. B. Fisher and T. B. Moeslund, editeurs, Articulated Motion and Deformable Objects : 7th International Conference, AMDO 2012, volume 7878, pages 260–272. Springer, 2012.

[Andriluka 2014] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele. *2D Human Pose Estimation: New Benchmark and State of the Art Analysis*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3686–3693. IEEE, June 2014.

[Baak 2009] A. Baak, B. Rosenhahn, M. Mueller and H.-P. Seidel. *Stabilizing motion tracking using retrieved motion priors.* In Proceedings of the International Conference on Computer Vision, pages 1428–1435, 2009.

[Baak 2011] A. Baak, M. Müller, G. Bharaj, H. Seidel and C. Theobalt. *A data-driven approach for real-time full body pose reconstruction from a depth camera.* In Proceedings of the International Conference on Computer Vision, pages 1092–1099. IEEE, 2011.

[Barber 2012] D. Barber. Bayesian reasoning and machine learning. Cambridge University Press, 2012.

[Bardram 2011] J. E. Bardram, A. Doryab, R. M. Jensen, P. M. Lange, K. L. G. Nielsen and S. 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. IEEE, 2011.

[Belagiannis 2014a] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab and S. Ilic. *3d pictorial structures for multiple human pose estimation.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 1669–1676. IEEE, 2014.

[Belagiannis 2014b] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic and N. Navab. *Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures.* In ChaLearn Looking at People Workshop, European Conference on Computer Vision (ECCV2014). IEEE, September 2014.

[Belagiannis 2016] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic, H. Feussner and N. Navab. *Parsing human skeletons in an operating room.* Machine Vision and Applications, pages 1–12, 2016.

[Benenson 2013] R. Benenson, M. Mathias, T. Tuytelaars and L. V. Gool. *Seeking the Strongest Rigid Detector.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 3666–3673. IEEE, June 2013.

[Benenson 2014] R. Benenson, M. Omran, J. Hosang and B. Schiele. *Ten years of pedestrian detection, what have we learned?* In European Conference on Computer Vision, pages 613–627. Springer, 2014.

[Berclaz 2011] J. Berclaz, F. Fleuret, E. Turetken and P. Fua. *Multiple Object Tracking Using K-Shortest Paths Optimization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 9, pages 1806–1819, Sept 2011.

[Beyl 2015] T. Beyl, P. Nicolai, M. Comparetti, J. Raczkowsky, E. De Momi and H. Wörn. *Time-of-flight-assisted Kinect camera-based people detection for intuitive human*

robot cooperation in the surgical operating room. International Journal of Computer Assisted Radiology and Surgery, pages 1–17, 2015.

[Bishop 2006] C. M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer, Secaucus, NJ, USA, 2006.

[Bouget 2015] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele and P. Jannin. *Detecting Surgical Tools by Modelling Local Appearance and Global Shape*. IEEE Transactions on Medical Imaging, vol. 34, no. 12, pages 2603–2617, Dec 2015.

[Bourdev 2009] L. Bourdev and J. Malik. *Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations*. In International Conference on Computer Vision. IEEE, 2009.

[Bourdev 2010] L. Bourdev, S. Maji, T. Brox and J. Malik. *Detecting People Using Mutually Consistent Poselet Activations*. In European Conference on Computer Vision, ECCV'10, pages 168–181. Springer, 2010.

[Bourdev 2014] L. D. Bourdev, F. Yang and R. Fergus. *Deep Poselets for Human Detection*. CoRR, vol. abs/1407.0717, 2014.

[Burenius 2013] M. Burenius, J. Sullivan and S. Carlsson. *3D Pictorial Structures for Multiple View Articulated Pose Estimation*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3618–3625. IEEE, 2013.

[Butt 2013] A. A. Butt and R. T. Collins. *Multi-target Tracking by Lagrangian Relaxation to Min-cost Network Flow*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1846–1853. IEEE, June 2013.

[Buys 2013] K. Buys, C. Cagniart, A. Baksheev, T. D. Laet, J. D. Schutter and C. Pantofaru. *An adaptable system for RGB-D based human body detection and pose estimation*. Journal of Visual Communication and Image Representation, 2013.

[Caicedo 2015] J. C. Caicedo and S. Lazebnik. *Active Object Localization with Deep Reinforcement Learning*. In International Conference on Computer Vision, pages 2488–2496. IEEE, Dec 2015.

[Carinou 2011] E. Carinou, M. Brodecki, J. Domienik, L. Donadille, C. Koukorava, S. Krim, D. Nikodemová, N. Ruiz-Lopez, M. Sans-Merce, L. Struelens and F. Vanhavere. *Recommendations to reduce extremity and eye lens doses in interventional radiology and cardiology*. Radiation Measurements, vol. 46, no. 11, pages 1324 – 1329, 2011. International Workshop on Optimization of Radiation Protection of Medical Staff, {ORAMED} 2011.

[Chekuri 2001] C. Chekuri, S. Khanna, J. S. Naor and L. Zosin. *Approximation algorithms for the metric labeling problem via a new linear programming formulation*. In Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, pages 109–118. Society for Industrial and Applied Mathematics, 2001.

[Chen 2014] X. Chen and A. Yuille. *Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations.* In Advances in Neural Information Processing Systems (NIPS), 2014.

[Cooper 1990] G. F. Cooper. *The computational complexity of probabilistic inference using bayesian belief networks.* Artificial Intelligence, vol. 42, no. 2, pages 393 – 405, 1990.

[Dalal 2005] N. Dalal and B. Triggs. *Histograms of oriented gradients for human detection.* In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, pages 886–893 vol. 1. IEEE, June 2005.

[Dalal 2006] N. Dalal, B. Triggs and C. Schmid. Human detection using oriented histograms of flow and appearance, pages 428–441. Springer, 2006.

[Dollar 2009] P. Dollar, Z. Tu, P. Perona and S. Belongie. *Integral Channel Features.* In Proceedings of the British Machine Vision Conference, pages 91.1–91.11. BMVC Press, 2009.

[Dollar 2012] P. Dollar, C. Wojek, B. Schiele and P. Perona. *Pedestrian Detection: An Evaluation of the State of the Art.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pages 743–761, 2012.

[Eichner 2009] M. Eichner and V. Ferrari. *Better Appearance Models for Pictorial Structures.* In Proceedings of the British Machine Vision Conference, pages 1–11. BMVC Press, 2009.

[Eichner 2012a] M. Eichner and V. Ferrari. *Human Pose Co-Estimation and Applications.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pages 2282–2288, Nov 2012.

[Eichner 2012b] M. Eichner, M. Marin-Jimenez, A. Zisserman and V. Ferrari. *2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images.* International Journal of Computer Vision, vol. 99, no. 2, pages 190–214, 2012.

[Felzenszwalb 2004] P. F. Felzenszwalb and D. P. Huttenlocher. *Distance transforms of sampled functions.* Technical report, Cornell Computing and Information Science, 2004.

[Felzenszwalb 2005] P. F. Felzenszwalb and D. P. Huttenlocher. *Pictorial Structures for Object Recognition.* International Journal of Computer Vision, vol. 61, no. 1, pages 55–79, 2005.

[Felzenszwalb 2010] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan. *Object Detection with Discriminatively Trained Part-Based Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1627–1645, September 2010.

[Ferrari 2008] V. Ferrari, M. Marin-Jimenez and A. Zisserman. *Progressive search space reduction for human pose estimation.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, June 2008.

[Fischler 1973] M. A. Fischler and R. A. Elschlager. *The Representation and Matching of Pictorial Structures.* IEEE Transactions on Computers, vol. C-22, no. 1, pages 67–92, Jan 1973.

[Gall 2010] J. Gall, B. Rosenhahn, T. Brox and H.-P. Seidel. *Optimization and Filtering for Human Motion Capture.* International Journal of Computer Vision, vol. 87, no. 1, pages 75–92, 2010.

[Gavrila 1999] D. M. Gavrila and V. Philomin. *Real-time object detection for "smart" vehicles.* In International Conference on Computer Vision, volume 1, pages 87–93 vol.1. IEEE, 1999.

[Gavrila 2000] D. M. Gavrila. *Pedestrian Detection from a Moving Vehicle.* In D. Vernon, editeur, European Conference on Computer Vision, pages 37–49. Springer, 2000.

[Gkioxari 2014] G. Gkioxari, B. Hariharan, R. Girshick and J. Malik. *Using k-Poselets for Detecting People and Localizing Their Keypoints.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, pages 3582–3589. IEEE, 2014.

[Grimmer 2011] J. Grimmer. *An Introduction to Bayesian Inference via Variational Approximations.* Political Analysis, vol. 19, no. 1, pages 32–47, 2011.

[Haque 2016] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung and F. Li. *Viewpoint Invariant 3D Human Pose Estimation with Recurrent Error Feedback.* CoRR, vol. abs/1603.07076, 2016.

[He 2015] K. He, X. Zhang, S. Ren and J. Sun. *Deep Residual Learning for Image Recognition.* CoRR, vol. abs/1512.03385, 2015.

[Hofmann 2011] M. Hofmann and D. M. Gavrila. *Multi-view 3D Human Pose Estimation in Complex Environment.* International Journal of Computer Vision, vol. 96, no. 1, pages 103–124, 2011.

[Huo 2012] F. Huo and E. A. Hendriks. *Multiple people tracking and pose estimation with occlusion estimation.* Computer Vision and Image Understanding, vol. 116, no. 5, pages 634 – 647, 2012.

[Insafutdinov 2016] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka and B. Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model, pages 34–50. Springer, 2016.

[Jafari 2014] O. H. Jafari, D. Mitzel and B. Leibe. *Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras.* In IEEE International Conference on Robotics and Automation (ICRA), pages 5636–5643, May 2014.

[Jain 2014] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor and C. Bregler. *Learning Human Pose Estimation Features with Convolutional Networks.* In International Conference on Learning Representations (ICLR), April 2014.

[Jain 2015] A. Jain, J. Tompson, Y. LeCun and C. Bregler. Modeep: A deep learning framework using motion features for human pose estimation, pages 302–315. Springer, 2015.

[Jia 2014] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell. *Caffe: Convolutional Architecture for Fast Feature Embedding.* arXiv preprint arXiv:1408.5093, 2014.

[Joo 2015] H. Joo, H. Liu, L. Tan, L. Gui, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara and Y. Sheikh. *Panoptic Studio: A Massively Multiview System for Social Motion Capture.* In International Conference on Computer Vision, pages 3334–3342. IEEE, 2015.

[Judkins 2008] T. N. Judkins, D. Oleynikov and N. Stergiou. *Objective evaluation of expert performance during human robotic surgical procedures.* Journal of robotic surgery, vol. 1, no. 4, pages 307–312, 2008.

[Kadkhodamohammadi 2014] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Temporally Consistent 3D Pose Estimation in the Interventional Room Using Discrete MRF Optimization over RGBD Sequences.* In Information Processing in Computer-Assisted Interventions, volume 8498 of *Lecture Notes in Computer Science*, pages 168–177. Springer, 2014.

[Kadkhodamohammadi 2015] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Pictorial Structures on RGB-D Images for Human Pose Estimation in the Operating Room.* In Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science. Springer, 2015.

[Kadkhodamohammadi 2017a] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Articulated clinician detection using 3D pictorial structures on RGB-D data.* Medical Image Analysis, vol. 35, pages 215 – 224, 2017.

[Kadkhodamohammadi 2017b] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms.* In Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2017. to appear.

[Khoshelham 2012] K. Khoshelham and S. O. Elberink. *Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications*. Sensors, vol. 12, no. 2, page 1437, 2012.

[Kiefel 2014] M. Kiefel and P. Gehler. *Human Pose Estimation with Fields of Parts*. In European Conference on Computer Vision, volume 8693 of *Lecture Notes in Computer Science*, pages 331–346. Springer, 2014.

[Kohn 2000] L. T. Kohn, J. M. Corrigan, M. S. Donaldson*et al.* To err is human:: building a safer health system, volume 6. National Academies Press, 2000.

[Koller 2007] D. Koller, N. Friedman, L. Getoor and B. Taskar. *Graphical Models in a Nutshell*. In L. Getoor and B. Taskar, editeurs, Introduction to Statistical Relational Learning. MIT Press, 2007.

[Komodakis 2007] N. Komodakis and G. Tziritas. *Approximate Labeling via Graph Cuts Based on Linear Programming*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 8, pages 1436–1453, Aug 2007.

[Komodakis 2008] N. Komodakis, G. Tziritas and N. Paragios. *Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies*. Computer Vision and Image Understanding, vol. 112, no. 1, pages 14 – 29, 2008. Special Issue on Discrete Optimization in Computer Vision.

[Kumar 2015] S. Kumar, P. Singhal and V. N. Krovi. *Computer-Vision-Based Decision Support in Surgical Robotics*. IEEE Design Test, vol. 32, no. 5, pages 89–97, Oct 2015.

[Ladikos 2008] A. Ladikos, S. Benhimane and N. Navab. *Real-Time 3D Reconstruction for Collision Avoidance in Interventional Environments*. In Medical Image Computing and Computer-Assisted Intervention, volume 5242, pages 526–534. Springer, 2008.

[Ladikos 2010] A. Ladikos, C. Cagniart, R. Ghotbi, M. Reiser and N. Navab. *Estimating radiation exposure in interventional environments*. In Medical Image Computing and Computer-Assisted Intervention, volume 13, pages 237–244. Springer, 2010.

[Lalys 2014] F. Lalys and P. Jannin. *Surgical process modelling: a review*. International Journal of Computer Assisted Radiology and Surgery, vol. 9, no. 3, pages 495–511, 2014.

[Lasota 2014] P. A. Lasota, G. F. Rossano and J. A. Shah. *Toward safe close-proximity human-robot interaction with standard industrial robots*. In IEEE International Conference on Automation Science and Engineering, pages 339–344. IEEE, 2014.

[Liu 2013] J. Liu, Y. Liu, Y. Cui and Y. Q. Chen. *Real-time human detection and tracking in complex environments using single RGBD camera*. In IEEE International Conference on Image Processing, pages 3088–3092, Sept 2013.

[Liu 2015] Z. Liu, J. Zhu, J. Bu and C. Chen. *A survey of human pose estimation: The body parts parsing based methods.* Journal of Visual Communication and Image Representation, vol. 32, pages 10 – 19, 2015.

[Loy Rodas 2015] N. Loy Rodas and N. Padoy. *Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, Monte Carlo simulations and wireless dosimeters.* International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 8, pages 1181–1191, 2015.

[Luo 2010] X. Luo, B. Berendsen, R. T. Tan and R. C. Veltkamp. *Human Pose Estimation for Multiple Persons Based on Volume Reconstruction.* In International Conference on Pattern Recognition, pages 3591–3594, Aug 2010.

[Makary 2016] M. A. Makary and M. Daniel. *Medical error—the third leading cause of death in the US.* BMJ, vol. 353, page i2139, 2016.

[Medina 2013] C. Medina, J. C. Segura and A. De la Torre. *Ultrasound Indoor Positioning System Based on a Low-Power Wireless Sensor Network Providing Sub-Centimeter Accuracy.* Sensors, vol. 13, no. 3, pages 3501–3526, 2013.

[Meißner 2014] C. Meißner, J. Meixensberger, A. Pretschner and T. Neumuth. *Sensor-based surgical activity recognition in unconstrained environments.* Minimally Invasive Therapy & Allied Technologies, vol. 23, no. 4, pages 198–205, 2014.

[Michalos 2015] G. Michalos, S. Makris, P. Tsarouchi, T. Guasch, D. Kontovrakis and G. Chryssolouris. *Design Considerations for Safe Human-robot Collaborative Workplaces.* Procedia CIRP, vol. 37, pages 248–253, 2015.

[Mikolajczyk 2005] K. Mikolajczyk and C. Schmid. *A performance evaluation of local descriptors.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pages 1615–1630, Oct 2005.

[Mitchelson 2003] J. R. Mitchelson and A. Hilton. *Simultaneous Pose Estimation of Multiple People using Multiple-View Cues with Hierarchical Sampling.* In Proceedings of the British Machine Vision Conference, pages 1–10. BMVC Press, 2003.

[Munaro 2014] M. Munaro and E. Menegatti. *Fast RGB-D people tracking for service robots.* Autonomous Robots, vol. 37, no. 3, pages 227–242, 2014.

[Murphy 1999] K. P. Murphy, Y. Weiss and M. I. Jordan. *Loopy Belief Propagation for Approximate Inference: An Empirical Study.* In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[Nara 2011] A. Nara, K. Izumi, H. Iseki, T. Suzuki, K. Nambu and Y. Sakurai. Surgical workflow monitoring based on trajectory data mining, pages 283–291. Springer, 2011.

[Nara 2015] A. Nara, C. Allen and K. Izumi. *Surgical Phase Recognition using Movement Data from Video Imagery and Location Sensor Data.* In Proceedings of the 13th International Conference on GeoComputation, September 2015.

[Newell 2016] A. Newell, K. Yang and J. Deng. Stacked hourglass networks for human pose estimation, pages 483–499. Springer, 2016.

[OpenNI 2016] OpenNI. *PrimeSense NiTE library*, Janvier 2016. Last access: Jan 2016.

[Oren 1997] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio. *Pedestrian detection using wavelet templates.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 193–199, Jun 1997.

[Ouyang 2012] W. Ouyang and X. Wang. *A discriminative deep model for pedestrian detection with occlusion handling.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 3258–3265. IEEE, 2012.

[Ouyang 2013] W. Ouyang, X. Zeng and X. Wang. *Modeling Mutual Visibility Relationship in Pedestrian Detection.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 3222–3229. IEEE, June 2013.

[Padoy 2009] N. Padoy, D. Mateus, D. Weinland, M.-O. Berger and N. Navab. *Workflow Monitoring based on 3D Motion Features.* In Computer Vision Workshops (ICCV Workshops), pages 585–592, 2009.

[Padoy 2011] N. Padoy and G. D. Hager. *3D thread tracking for robotic assistance in tele-surgery.* In Intelligent Robots and Systems, pages 2102–2107. IEEE, 2011.

[Padoy 2012] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger and N. Navab. *Statistical modeling and recognition of surgical workflow.* Medical Image Analysis, vol. 16, no. 3, pages 632 – 641, 2012. Computer Assisted Interventions.

[Papageorgiou 1999] C. Papageorgiou and T. Poggio. *Trainable pedestrian detection.* In International Conference on Image Processing, volume 4, pages 35–39 vol.4, 1999.

[Pearl 1988] J. Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Pishchulin 2016] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler and B. Schiele. *DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation.* In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016.

[Ramanan 2007] D. Ramanan. *Learning to parse images of articulated bodies.* In Advances in Neural Information Processing Systems (NIPS), pages 1129–1136. MIT Press, 2007.

[Reiley 2011] C. E. Reiley, H. C. Lin, D. D. Yuh and G. D. Hager. *Review of methods for objective surgical skill evaluation.* Surgical Endoscopy, vol. 25, no. 2, pages 356–366, 2011.

[Sapp 2010] B. Sapp, C. Jordan and B. Taskar. *Adaptive pose priors for pictorial structures.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 422–429. IEEE, 2010.

[Sapp 2011] B. Sapp, D. Weiss and B. Taskar. *Parsing human motion with stretchable models.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 1281–1288. IEEE, 2011.

[Schlesinger 2006] D. Schlesinger and B. Flach. *Transforming an arbitrary minsum problem into a binary one.* Technical report, Dresden University of Technology, 2006.

[Schmidhuber 2015] J. Schmidhuber. *Deep learning in neural networks: An overview.* Neural Networks, vol. 61, pages 85 – 117, 2015.

[Sermanet 2013] P. Sermanet, K. Kavukcuoglu, S. Chintala and Y. Lecun. *Pedestrian Detection with Unsupervised Multi-stage Feature Learning.* In IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, pages 3626–3633. IEEE, 2013.

[Shotton 2012] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman and A. Blake. *Efficient Human Pose Estimation from Single Depth Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pages 2821 – 2840, 2012.

[Sigal 2009] L. Sigal, A. O. Balan and M. J. Black. *HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion.* International Journal of Computer Vision, vol. 87, no. 1, pages 4–27, 2009.

[Spinello 2011] L. Spinello and K. O. Arras. *People detection in RGB-D data.* In IEEE International Conference on Intelligent Robots and Systems, pages 3838–3843. IEEE, Sept 2011.

[Stoll 2011] C. Stoll, N. Hasler, J. Gall, H. P. Seidel and C. Theobalt. *Fast articulated motion tracking using a sums of Gaussians body model.* In International Conference on Computer Vision, pages 951–958. IEEE, Nov 2011.

[Sutherland 2006] J. Sutherland and W. van den Heuvel. *Towards an Intelligent Hospital Environment: Adaptive Workflow in the OR of the Future.* In Hawaii International Conference on Systems Science, 2006.

[Tang 2013] S. Tang, X. Wang, X. Lv, T. Han, J. Keller, Z. He, M. Skubic and S. Lao. *Histogram of Oriented Normal Vectors for Object Recognition with a Depth Sensor.* In K. Lee, Y. Matsushita, J. Rehg and Z. Hu, editeurs, Asian Conference on Computer Vision, volume 7725 of *Lecture Notes in Computer Science*, pages 525–538. Springer, 2013.

[Tokola 2013] R. Tokola, W. Choi and S. Savarese. *Breaking the chain: liberation from the temporal Markov assumption for tracking human poses.* In International Conference on Computer Vision. IEEE, 2013.

[Tompson 2014] J. J. Tompson, A. Jain, Y. LeCun and C. Bregler. *Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation.* In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, editeurs, Advances in Neural Information Processing Systems (NIPS), pages 1799–1807. Curran Associates, Inc., 2014.

[Tompson 2015] J. Tompson, R. Goroshin, A. Jain, Y. LeCun and C. Bregler. *Efficient object localization using Convolutional Networks.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 648–656. IEEE, June 2015.

[Toshev 2014] A. Toshev and C. Szegedy. *DeepPose: Human Pose Estimation via Deep Neural Networks.* In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.

[Tran 2016] D. T. Tran, R. Sakurai and J.-H. Lee. An improvement of surgical phase detection using latent dirichlet allocation and hidden markov model, pages 249–261. Springer, 2016.

[Twinanda 2015] A. Twinanda, E. Alkan, A. Gangi, M. de Mathelin and N. 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.

[Twinanda 2016] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin and N. Padoy. *EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos.* CoRR, vol. abs/1602.03012, 2016.

[Vanhavere 2008] F. Vanhavere, E. Carinou, L. Donadille, M. Ginjaume, J. Jankowski, A. Rimpler and M. S. Merce. *An overview on extremity dosimetry in medical applications.* Radiation protection dosimetry, 2008.

[Vapnik 1995] V. N. Vapnik. The nature of statistical learning theory. Springer, New York, NY, USA, 1995.

[Vedula 2016] S. S. Vedula, A. Malpani, N. Ahmidi, S. Khudanpur, G. Hager and C. C. G. Chen. *Task-Level vs. Segment-Level Quantitative Metrics for Surgical*

Skill Assessment. Journal of Surgical Education, vol. 73, no. 3, pages 482 – 489, 2016.

[Vijayanarasimhan 2014] S. Vijayanarasimhan and K. Grauman. *Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds*. International Journal of Computer Vision, vol. 108, no. 1, pages 97–114, 2014.

[Wang 2013] C. Wang, N. Komodakis and N. Paragios. *Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey*. Computer Vision and Image Understanding, vol. 117, no. 11, pages 1610 – 1627, 2013.

[Wanzel 2002] K. R. Wanzel, E. D. Matsumoto, S. J. Hamstra and D. J. Anastakis. *Teaching technical skills: training on a simple, inexpensive, and portable model*. Plastic and reconstructive surgery, vol. 109, no. 1, page 258—263, January 2002.

[Wei 2016] S.-E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh. *Convolutional pose machines*. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016.

[Wong 2015] C. Wong, Z. Q. Zhang, B. Lo and G. Z. Yang. *Wearable Sensing for Solid Biomechanics: A Review*. IEEE Sensors Journal, vol. 15, no. 5, pages 2747–2760, May 2015.

[Xu 2016] W. Xu, P. c. Su and S. c. S. Cheung. *Human pose estimation using two RGB-D sensors*. In 2016 IEEE International Conference on Image Processing (ICIP), pages 1279–1283, Sept 2016.

[Yang 2013] Y. Yang and D. Ramanan. *Articulated human detection with flexible mixtures of parts*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pages 2878–2890, 2013.

[Yang 2016] W. Yang, W. Ouyang, H. Li and X. Wang. *End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation*. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016.

[Yao 2012] A. Yao, J. Gall and L. Van Gool. *Coupled Action Recognition and Pose Estimation from Multiple Views*. International Journal of Computer Vision, vol. 100, no. 1, pages 16–37, 2012.

[Ye 2011] M. Ye, X. Wang, R. Yang, L. Ren and M. Pollefeys. *Accurate 3d pose estimation from a single depth image*. In International Conference on Computer Vision, pages 731–738. IEEE, 2011.