Object Detection for Contactless Vital Signs Estimation

Fan Yang

Thesis submitted to the University of Ottawa
In partial fulfillment of the requirements for the
Master of Applied Science in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

This thesis explores the contactless estimation of people's vital signs. We designed two camera-based systems and applied object detection algorithms to locate the regions of interest where vital signs are estimated. With the development of Deep Learning, Convolutional Neural Network (CNN) model has many applications in the real world nowadays. We applied the CNN based frameworks to the different types of camera based systems and improve the efficiency of the contactless vital signs estimation.

In the field of medical healthcare, contactless monitoring draws a lot attention in the recent years because the wide use of different sensors. However most of the methods are still in the experimental phase and have never been used in real applications. We were interested in monitoring vital signs of patients lying in bed or sitting around the bed at a hospital. This required using sensors that have range of 2 to 5 meters. We developed a system based on the depth camera for detecting people's chest area and the radar for estimating the respiration signal. We applied a CNN based object detection method to locate the position of the subject lying in the bed covered with blanket. And the respiratory-like signal is estimated from the radar device based on the detected subject's location.

We also create a manually annotated dataset containing 1,320 depth images. In each of the depth image the silhouette of the subject's upper body is annotated, as well as the class. In addition, a small subset of the depth images also labeled four keypoints for the positioning of people's chest area. This dataset is built on the data collected from the anonymous patients at the hospital which is substantial.

Another problem in the field of human vital signs monitoring is that systems seldom contain the functions of monitoring multiple vital signs at the same time. Though there are few attempting to work on this problem recently, they are still all prototypes and have a lot limitations like shorter operation distance. In this application, we focused on contactless estimating subjects' temperature, breathing rate and heart rate at different distances with or without wearing the mask. We developed a system based on thermal and RGB camera and also explore the feasibility of CNN based object detection algorithms to detect the vital signs from human faces with specifically defined RoIs based on our thermal camera system. We proposed the methods to estimate respiratory rate and heart rate from the thermal videos and RGB videos. The mean absolute difference (MAE) between the estimated HR using the proposed method and the baseline HR for all subjects at different distances is 4.24 ± 2.47 beats per minute, the MAE between the estimated RR and the reference RR for all subjects at different distances is 1.55 ± 0.78 breaths per minute.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Miodrag Bolic who has been crucial in defining the research problem and directing me in this research topic. I also appreciate professor Bolic for guiding me through research practices consistently and teaching me to ask important research questions.

I thank my colleagues Shan He, Zixiong Han, Rajitha Hathurusinghe, Varun Mehta and Cristóvão Iglesias at Health Devices Research Group (HDRG) for working together in solving various interesting research problems and their ideas in developing my abilities.

I thank Zixiong Han for introducing me to Prof Bolic and HDRG. I like to thank everybody at HDRG where I could spend the rest of my time doing my research work. I like to remind Angel, Alexandra, Deepak, Ziqiao, Runzhi, Wenbin, Lingfeng, Junhao. Discussions and knowledge sharing sessions with you all have been crucial in my thinking while proceeding with the research work. I am also thankful for the friendly and motivating atmosphere at HDRG.

I thank Prof. Bolic for generously funding my research through Natural Sciences and Engineering Research Council of Canada (NSERC). I am beyond grateful in receiving such tremendous support for my research.

I also thank all the friends I met for the first time in Ottawa including many friendly Chinese students who made me feel like back home here and being supportive in many ways.

Finally I would like to thank my parents for their support throughout my education at University of Ottawa.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Due to the availability of computing resources and developments in deep learning, deep neural networks have been proven to have the power to solve real world tasks efficiently. The neural network models rely on a vast number of data to implement acceptable predictions on unseen data. Especially in the field of computer vision, convolutional neural networks (CNNs) have changed the way how people view and solve the problems. Since Alex et al proposed "ImageNet Classification with Deep Convolutional Neural Networks" in 2012 [1], the CNNs have been widely used to address the tasks like image classification and object detection. In addition, with the improvement of the computing capabilities, for example Nvidia has released generations of graphic cards i.e. GPUs, it is now common to apply the CNNs and train the model based on massive amounts of data. Deep learning methods for computer vision problems have the advantages of robustness, scalability, and acceptable accuracy over other algorithms. However, the CNNs based methods are not guarantee for solving all the problems, and they also have some limitations and boundaries in the application. For example, the dataset is always one of the problem that has huge impact on the generated models. A biased dataset that does not accurately represent a model's use case, will train to get some biased models resulting in skewed outcomes [2]. A small dataset will not lead to a high generalization ability of the model. In addition, the privacy of the data source would potentially cause issues due to not having consent from data owners or other unavoidable ethical conflicts. General Data Protection Regulation (GDPR) [3] imposed by European Union which is also adopted by North America in the recent past is a major law governing the data privacy.

Object detection is one of the research direction of computer vision, dealing with detecting instances of semantic objects of a certain class like humans or cars in the digital images and videos. It is a complicated task in the sub-field of computer vision, involving many other computer vision tasks such as classification [4], segmentation [5] - [6], and tracking [7]. To figure out the location of the object in the view and what kind of class it is, really helps

researchers to have better understanding of the whole scene. For example, in the research of self-driving cars, pedestrians and other cars on the road must be detected by the vision system and then the safe path would be generated to drive. Another application of object detection is in the medical research. Researchers present the medical images segmentation of tubular anatomical structures such as the aortic arc and the spinal cord by applying parametric object detection and tracking [8].

Traditionally, the research in object detection is mainly using frames or streams collected from common RGB cameras. Many of the algorithms and methods are designed for the purpose of RGB image-based dataset. It is very common to see well-built dataset that is composed of massive number of images with well-annotated labels like PASCAL Visual Object Classes dataset [9]. The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. These challenges and dataset contribute a lot to the development of object detection algorithm. With the availability of advanced imaging devices, there are also some task specific dataset built on special images collected from devices other than normal RGB cameras like RGB-D Object Dataset [10]. The RGB-D Object Dataset is a large dataset of 300 common household objects. The objects are organized into 51 categories arranged using WordNet hypernym-hyponym relationships (similar to ImageNet). This dataset was recorded using a Kinect style 3D camera that records synchronized and aligned 640x480 RGB and depth images. Each object was placed on a turntable and video sequences were captured for one whole rotation. However, there is still a lack of datasets like these for research.

Our research is mainly focused on using object detection based on the different imaging devices like depth camera and thermal camera to monitor the vital signs of the subjects in a contactless way. With the attention to the personal privacy increasing, currently most of the surveillance systems are putting the security in the first place especially in the places like hospital and nursing home. Privacy and confidentiality are the biggest barriers in utilizing many diverse devices in creating a dataset to learn useful tasks in the field of healthcare. As a result, we design and install a depth camera based system in the hospital ward to collect depth frames according to the ethic requirements pointed out in Section 4.1. In addition, object detection is working for the locating of the subjects before we attempt to extract people's vital signs like respiratory rate and heart rate. We have tested object detection algorithms on depth camera, thermal camera, and RGB camera.

The reason behind using non-RGB cameras such as depth camera and thermal camera is that those devices can provide us with extra useful information. For example, depth camera

can give out the distance between the camera and the object that could be used to locate position of the object in the real world. In addition thermal camera can provide the temperature distribution of the object. In order to utilize these meaningful information from the depth frames or thermal frames, object detection needs to be done to help find our regions of interest. The data like distance and temperature can be analyzed and processed for further signal extraction. With the development of optical chips and imaging technology, there are more possibilities that we can obtain vision information from different devices rather than the common RGB cameras. We want more information from not only what we could see and sense by ourselves, but also the invisible and insensible by human eyes under the allowed conditions. Meanwhile, the study of many years on image processing provide us opportunity to have a better understanding on this data. Many related methods like object detection algorithms can also help us to process these special depth images or thermal images with different formats and contents. We learned from object detection based on usual RGB images and then transfer the knowledge and apply it to unusual images.

Although there are many researches both in the object detection area and contactless vital signs monitoring field, many problems still remain unsolved. For example, Microsoft Kinect depth camera has the algorithms to detect the human postures [11], but they cannot detect the complex postures like when the person is lying in bed. The algorithms require the targets to be facing in front of the camera and partly occluded human body cannot be detected. This is more suitable for the situation like somatic games rather than for the purpose of monitoring. However in our depth camera project, the subject is lying in the bed covered with a blanket or clothes so that the typical posture estimation algorithms does not work. The region of interest (RoI) detection is always the biggest problem in the research of contactless vital signs monitoring because the sensors usually have a larger sensing area than the targeted areas we actually need. Moreover, limited by the environmental factors like light source and room layout, the methods that work under the controlled conditions might not work in the real world application. Other problems like low estimation accuracy and short detection range also need further study.

To find the solutions to the above research problems, we designed camera-based collection systems. One system is using a depth camera and a radar device, another system is using RGB camera and a thermal camera. The depth camera based system was installed in a hospital to collect data from volunteer patients. We have created datasets not only including different types of images but also extra data collected from radar, breathing belt, and heart rate monitoring device at the same time. The depth image dataset was well annotated with classes and labels to train the deep learning model. We trained a CNNs based model to detect subject lying in bed at the hospital ward as illustrated in Figure 1.1. We present

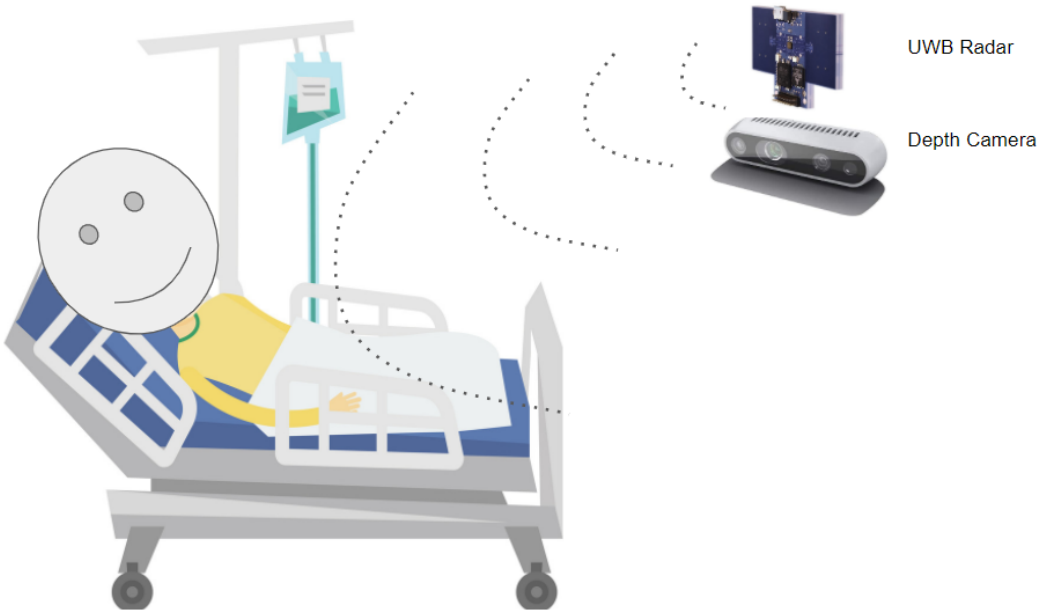


Figure 1.1: Schematic diagram of the depth camera research.

the test results of the trained model to evaluate the performance produced by this end-to-end CNN based approach. In addition, we applied different face detectors in the thermal camera system to extract vital signs of the subject as illustrated in Figure 1.2. We also employed image enhancement algorithms to improve the weak signal from the thermal videos.

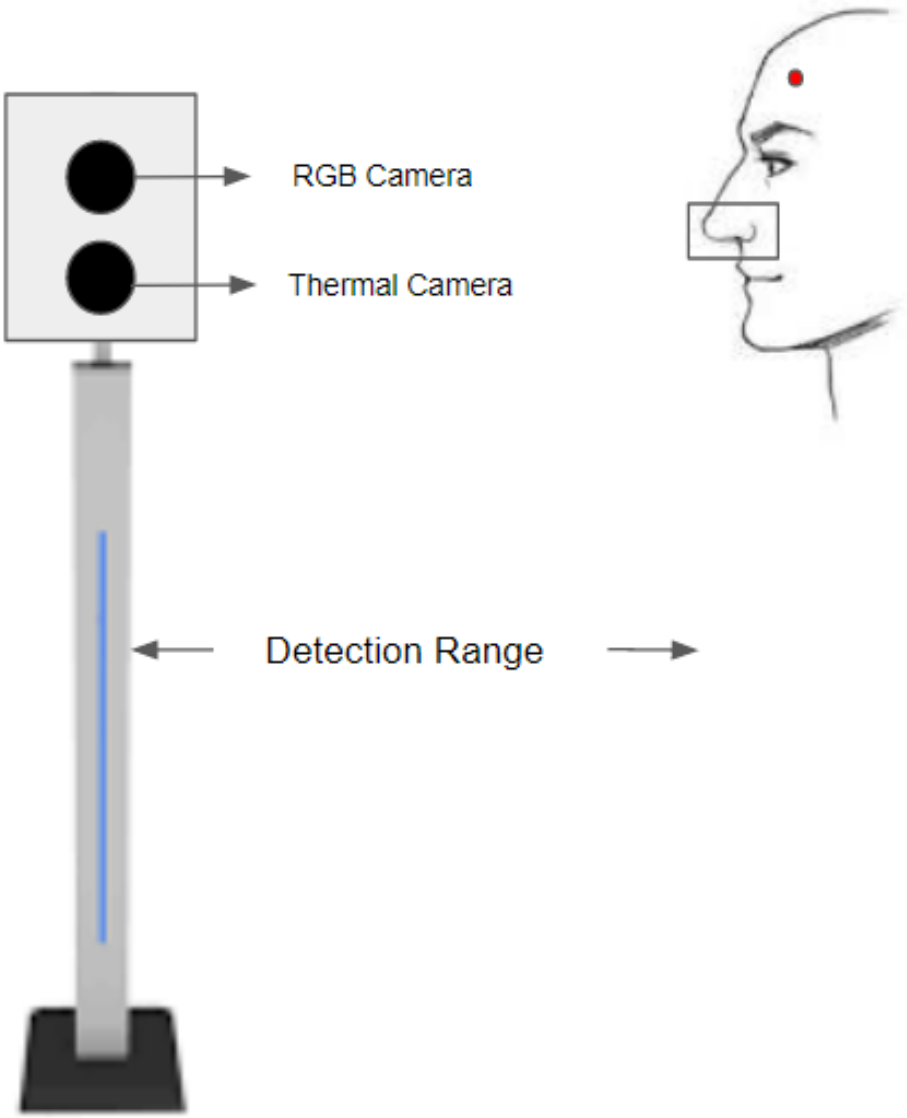


Figure 1.2: Schematic diagram of the thermal camera research.

1.1 Motivation

Monitoring vital signs without any contact is a crucial task in healthcare. Camera-based vital sign estimation allows the contactless assessment of important physiological parameters [12]. The vital signs like respiratory rate, heart rate, and body temperature are useful in detecting or monitoring medical problems no matter in a medical setting, at home, at the site of a medical emergency, or elsewhere [13]. The respiration rate is the number of breaths a person takes per minute. The rate is usually measured when a person is at rest and simply involves counting the number of breaths for one minute by counting how many times the chest rises. Respiration rates may increase with fever, illness, and other medical conditions. When checking respiration, it is important to also note whether a person has any difficulty breathing. Normal respiration rates for an adult person at rest range from 12 to 16 breaths per minute. The normal heart rate, or the number of times the heart beats per minute, for healthy adults ranges from 60 to 100 beats per minute. The heart rate may fluctuate and increase with exercise, illness, injury, and emotions. The normal body temperature of a person varies depending on gender, recent activity, food and fluid consumption, time of day, and, in women, the stage of the menstrual cycle. Normal body temperature can range from 97.8 degrees F (or Fahrenheit, equivalent to 36.5 degrees C, or Celsius) to 99 degrees F (37.2 degrees C) for a healthy adult. A person's body temperature can be taken by skin. The special thermometer can quickly measure the temperature of the skin on the forehead. The remote and contactless measurement makes it possible to use the equipment in more scenarios under any conditions and improves the efficiency of the diagnosis if the accuracy can be assured. Additionally, it provides a potential that multiple persons could be monitored at the same time using cameras.

With motivation to create a well-labeled depth image dataset, we collect data in the real world using the system designed by ourselves and manually annotate each image using the existing annotation tool. We fine-tune and train the state-of-the-art pre-trained object detection model with transfer learning which applies the knowledge gained from one problem to solve a different but related problem. We use the model to estimate the distance between the subject and the device, and extract the matched respiratory signal with the help of radar data collected at the same time.

Another motivation is about testing existing CNNs based face detection algorithms and apply one of the best to the real-time scenario. We collected data using the combination of thermal camera and RGB camera. Meanwhile, the devices for collecting reference signals also collected data synchronously as a baseline.

Healthcare research with deep learning is novel and driven to explore ways to utilize

sensitive data in training neural network based models. Dataset is a vital need in the domain of deep learning. Most research done on medical or healthcare based tasks are done in isolation due to the sensitive nature of the utilized data. Research on privacy guaranteeing mechanisms such as de-anonymization published on a common dataset can provide general baselines and comparable results to drive towards improvements. This could probably explain one of the reason why there is lack of professional datasets in the field of healthcare. And another reason is due to the difficulty of collecting and labelling data.

1.2 Current Challenges and Existed Problems

Our final and overall goal of this research is to have an automated contactless monitoring system that it would achieve the following with contactless monitoring:

1) Detect people and track them in the field of view

2) Detect the regions of interest for vital signs monitoring

3) Classify activities of the subject so that we can obtain their vital signs when they do not move

4) Extract the vital signs only when the subject does not move

5) Assign the signal and the class to the particular subject

However, there are some challenges and problems for the whole system at this point. Since we are trying to apply the contactless monitoring, the estimation accuracy and detection range are always the two big problems compared to the currently widely used wearable devices. While there has been extensive work on contactless monitoring of the vital signs of humans using cameras, to the best of the best of our knowledge, there has been no system solving all of these problems. Different from the common RGB applications, we are using special cameras here capturing the non-RGB images. Therefore, the existing approaches that process this data are not effective and feasible anymore in our use case. In this thesis, we proposed some methods to address the existing problems and challenges and our system also implements the goal 1, 2, and 4 mentioned above.

Object Detection

We are using special cameras to detect subjects and monitor their vital signs. However, most of the existed algorithms and methods for object detection that will be introduced in

the Section 2.2 are based on the normal RGB images or videos. There are few examples that apply the object detection algorithms to the area of depth images. The state-of-the-art method like what Microsoft Kinect depth sensor uses [11] has trouble detecting the person under complicated conditions as illustrated in Figure 1.3 where the person is lying in the bed and is covered with blanket, without having the full body exposed. Similarly, Nuitrack [14] which is a 3D Skeleton Tracking Middleware also cannot handle the situation where the person is not facing the camera or a part of the human body is occluded. The typical object detection methods rely on the well-selected features to detect the desired object. These features, no matter designed by researchers or generated by the convolutional neural network, are representatives of the certain RGB patterns. While in our use case, we only have one-channel pixel values from the depth images that represent the distance from the camera to the object. New features or new CNN needs to be redesigned to detect the person in the depth images.

Figure 1.3: Example of a subject lying covered with a blanket where Microsoft Kinect system was not able to detect the subject.

Lack of Dataset

Even though there are many related researches studying the remote monitoring, there are few researches collecting data in the realistic occasions like hospital and retirement home.

There is also a lack of this kind of datasets that are collected from real-world scenes. For example, there are some datasets about RGB-D images capturing objects like furniture and utensils in the indoor scene [15]. Besides, there is also one Depth Images with Humans (DIH) dataset for human body landmark detection and human pose estimation from depth images [16]. However, the DIH dataset contains the synthetic images and images acquired with a Kinect 2 depth sensor totally under the controlled laboratory conditions. There are no depth image datasets that have well labeled object like people in the real world scenes like in a ward of a hospital. The creation of such a dataset is critical to the research area like medical healthcare. Especially the labels and class annotations of the subject help us to analyze and detect the potential conditions like sleep apnea. In addition there are many requirements and limitations when obtaining people's personal information which is a key problem about the privacy. So the privacy issue of the dataset that is composed of people's private information such as health data is also a big challenge in our research.

New Requirements for the Application

Current researches have already made huge progress and developed solutions to address some of the problems mentioned before. However with the time changing, there are new problems emerging and new requirements for the system. For example, with the explosion of Covid-19 pandemic in 2020, there are policies that require us to excersize social distancing and to wear masks in the public area. These policies raised challenges to our typical behaviors, and also add new requirements to many of the healthcare devices. The previous successful methods might have trouble fitting in the new requirements and even cannot work anymore under the new conditions because they did not consider the more complicated and difficult situations. For example, previous face detectors have problem detecting the faces occluded by the mask while wearing the mask is a necessary requirement for the people now. As a result, the approaches should be updated with the change of application scenarios.

1.3 Contributions

In this thesis, methods are introduced to solve the described problems in Section 1.2 using depth camera based system and thermal camera based system. Since we use different types of the camera, we contribute to integrating the convolutional neural networks based object detection methods to our camera-based systems as illustrated in Figure 1.4. Additionally we also explore the complicated situations where the previous approaches have never studied or the problems even state-of-the-art methods could not address yet. We attempt to enhance the

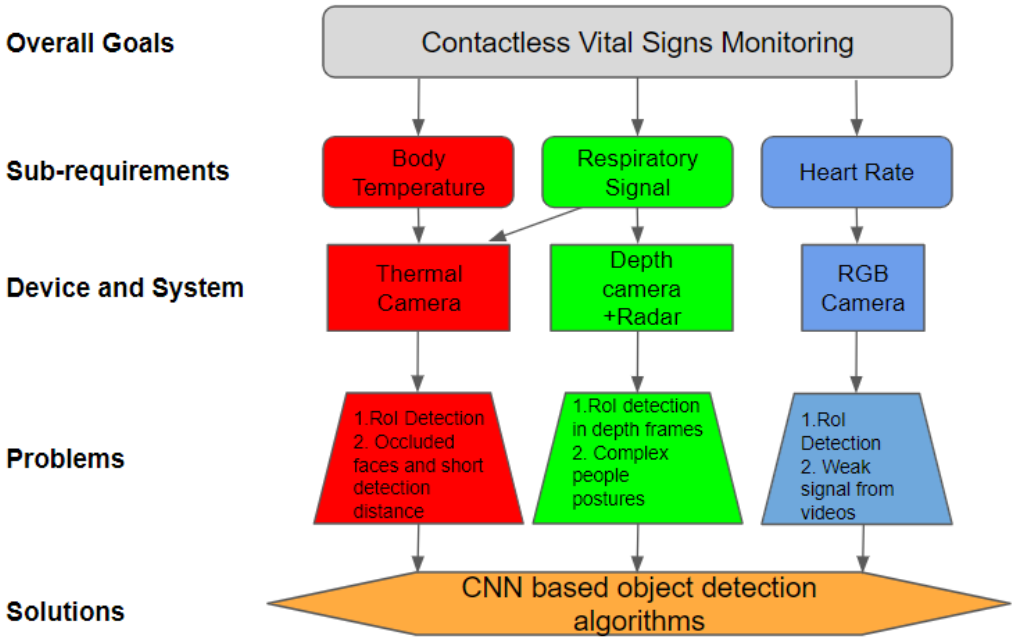


Figure 1.4: Overall introduction of the thesis.

performance of contactless vital signs estimation algorithms and apply the approaches to the real world situation by the following contributions:

Depth Camera Based Collection System

We design the collection system and use it to collect depth frames and radar data at the same time. The system works in real time and stores the data in an organized way. The system has been installed in the hospital ward and collected data from multiple subjects comforting to the related regulations described in Section 3.1.

Manually Labeled Dataset

We formed two datasets. The first dataset included not only frames from the depth camera but also radar data. The created dataset was manually labeled with annotation tools. One of the depth image dataset contains more than one thousand mask annotations and another one is composed of mask and keypoint labels. We present the details about the data source and created dataset information in Section 3.1.2 and Section 3.1.3 separately. Another dataset is composed of thermal camera video frames, RGB frames, baseline respiratory signals,

and baseline heart rate. The formats and details information of this dataset will be further introduced in Section 4.1.2.

Human Body Detection Based on Depth Frames

We train the CNNs based object detection model with transfer learning from the existing well-trained model based on our own dataset. Transfer learning is a process where a model trained on one problem is used in some way on a second related problem. We observe the performance on manually annotated validation dataset and the test set, satisfying our requirements for detection. In our problem, what we want to detect is the human body in the depth frames while the related problem is detecting the human in RGB images. We resolve the problem by learning from the knowledge of previously well-studied human posture estimation problem, and by modifying the CNN based detection model which will be detailed in 4.3.1. Besides, the new trained model is applied for addressing the problem of distance estimation.

Contactless Vital Signs Estimation

We compare different CNN based face detection models by testing with our use case and our own data, from which we find the best one to solve the problems. We integrate the state-of-the-art face detectors to our contactless vital signs estimation methods to address the challenges described in the Section 1.2. The combination of the approaches proves to meet the latest healthcare requirements like detection of breathing when wearing the mask as pointed in Section 1.2.

The following paper was published based on the research leading to this thesis:

- F. Yang, Z. Han and M. Bolic, "Detection of Respiratory Signal Based on Depth Camera Body Tracking," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 481-484

In addition to this one, another paper titled "Vital Signs Estimation Using Thermal Camera" is ready to be submitted to IEEE Access.

1.4 Organisation of the Thesis

In Chapter 2, we review the relevant background literature on the problems about object detection and camera based vital signs estimation. Section 2.1 gives the details of the different devices used in our depth camera based system and thermal camera based system separately as well as the overall systems' composition. Section 2.2 describes the literature review of object detection algorithms in recent years. Section 2.3 gives a brief literature on different remote vital signs monitoring methods.

Chapter 3 details the methods involved in achieving the contributions. Section 3.1 introduces how data collected through our depth camera system and the format of collected data as well as the details of the self-created dataset. The section also introduces the methods that the system takes to detect the human body and extract breath-like signal from the depth images. Section 3.2 gives the detail of the approaches that thermal camera system takes to extract vital signs.

Next, we present results and related analysis from our experiments in Chapter 4. These results closely follow the experiments mentioned in the Section 4.1. We provide the discussion of results and analysis under each subsection in the Chapter 3. In Section 4.2, we present the related information for our self-created dataset. Detailed statistics on dataset are presented in this section. Later in Section 4.3, we present the results and analysis of our depth image body tracking methods as well as respiratory signal estimation. Also, we present the comparison of different face detectors and the situation where the algorithms run on different devices in Section 4.4. Finally in Section 4.5, we describe the results of vital signs estimation based on our methods and the analysis by comparing with the reference data.

Chapter 5 presents our conclusion, summary of contributions and future work. Some of the future work are based on the result analysis in the last chapter.

Chapter 2

Background and Related Work

This chapter explains the background for the research, which spans over devices description and related algorithms literature review. Section 2.1 introduces the composition of depth camera based system and thermal camera based system. Section 2.2 reviews the object detection algorithms as well as the face detectors in recent years. Section 2.3 reviews the contactless monitoring methods for the human vital signs including respiratory signal and heart rate.

2.1 Devices

2.1.1 Depth Camera

Standard digital cameras output images as a 2D grid of pixels. Each pixel has values associated with it – usually the Red, Green and Blue, or RGB. Each attribute has a number from 0 to 255, so black, for example, is (0,0,0) and a pure bright red would be (255,0,0). Thousands to millions of pixels together create the kind of photographs we are all very familiar with. A depth camera on the other hand, has pixels which have a different numerical value associated with them, that number being the distance from the camera, or "depth." Some depth cameras have both an RGB and a depth system, which can give pixels with all four values, or RGBD [17].

The output from a depth camera can be displayed in a variety of ways. In the example as illustrated in Figure 2.1, the color image is shown side by side with the depth image, where each different color in the depth map represents a different distance from the camera. In this case, cyan is closest to the camera, and red is furthest. It does not really matter what color values the depth map uses, this is just displayed in this way to make it easy to visualize.

There are a variety of different methods for calculating depth, all with different strengths

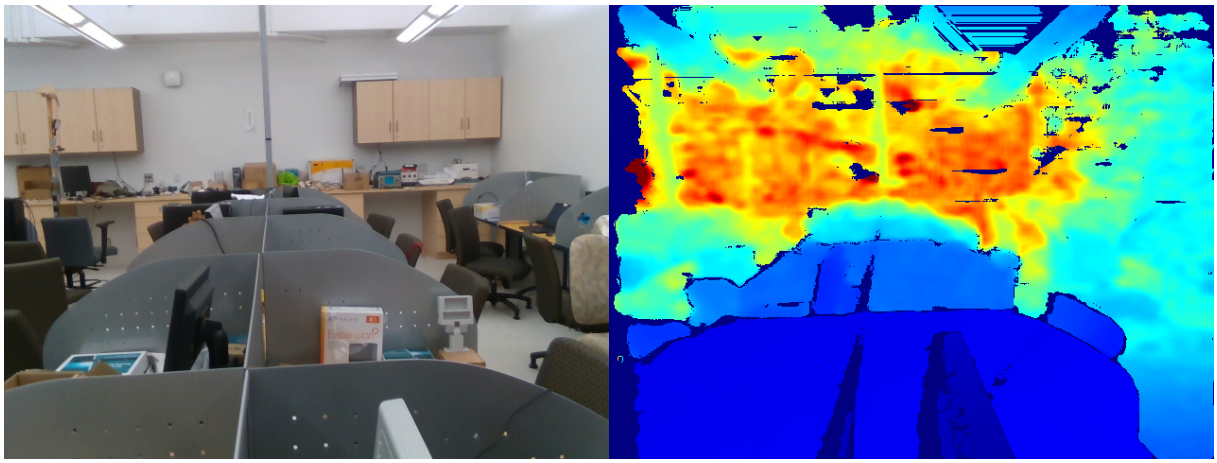


Figure 2.1: Color image and its depth map captured at the same time.

and weaknesses and optimal operating conditions. Depending on the application, here are some questions to consider: How far does the sensor need to detect? What sort of accuracy is needed? Does the sensor need to operate outdoors? And there are several specifications that needed to consider and compare when chose depth camera for application like the sensor range, image resolution, field of view (FOV), frame rate, etc.

Structured light and coded light depth cameras are similar technologies. They rely on projecting light (usually infrared light) from some kind of emitter onto the scene. The projected light is patterned, either visually or over time, or some combination of the two. Because the projected pattern is known, how the sensor in the camera sees the pattern in the scene provides the depth information. Using the disparity between an expected image and the actual image viewed by the camera, distance from the camera can be calculated for every pixel [17]. Because this technology relies on accurately seeing a projected pattern of light, coded and structured light cameras do best indoors at relatively short ranges. Another issue with systems like these is that they are vulnerable to other noise in the environment from other cameras or devices emitting infrared light. Similarly, all types of time of flight device emit some kind of light, sweep it over the scene, and then time how long that light takes to get back to a sensor on the camera. Depending on the power and wavelength of the light, time of flight sensors can measure depth at significant distances – for example, being used to map terrain from a helicopter.

Stereo depth cameras also often project infrared light onto a scene to improve the accuracy of the data, but unlike coded or structured light cameras, stereo cameras can use any light to measure depth. Stereo depth cameras have two sensors, spaced a small distance apart. A stereo camera takes the two images from these two sensors and compares them. Since the distance between the sensors is known, these comparisons give depth information.

Because stereo cameras use any visual features to measure depth, they will work well in most lighting conditions including outdoors. The addition of an infrared projector means that in low lighting conditions, the camera can still perceive depth details. The other benefit of this type of depth camera is that there are no limits to how many you can use in a particular space where the cameras do not interfere with each other in the same way that a coded light or time of flight camera would. The distance these cameras can measure is directly related to how far apart the two sensors are, where the wider the baseline is, the further the camera can see [17].

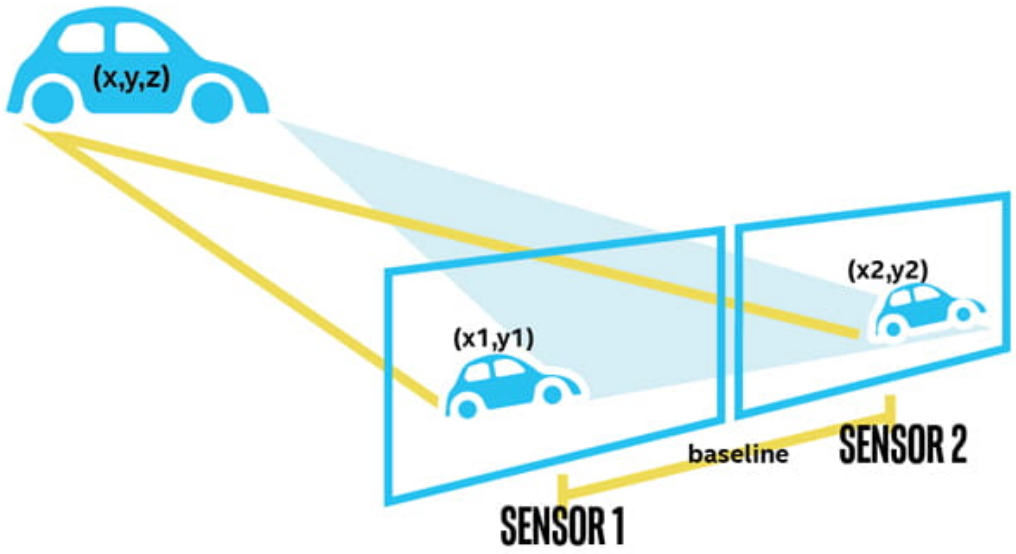


Figure 2.2: Operation of the stereo depth camera cited from [17]

Each kind of depth camera relies on known information in order to extrapolate depth. For example, in stereo, the distance between sensors is known. In coded light and structured light, the pattern of light is known. In the case of time of flight, the speed of light is the known variable used to calculate depth. LiDAR sensors are a type of time of flight camera which use laser light to calculate depth. All types of time of flight devices emit some kind of light, sweep it over the scene, and then measure the time the light takes to get back to a sensor on the camera. However, any situation where the light hitting the sensor may not have been the light emitted from the specific camera but could have come from some other source like the sun or another camera can degrade the quality of the depth image. All depth cameras provide the advantage of additional understanding about a scene, and it gives any device or system the ability to understand a scene in ways that do not require human intervention [17]. While

it is possible for a computer to extract information from a 2D image, that requires significant investment and time in training a machine learning network. A depth camera inherently gives some information without the need for training, for example, it is easier to distinguish foreground and background objects from a scene. This becomes useful in applications like background segmentation – a depth camera can remove background objects from an image, allowing a green-screen free capture. Besides, depth cameras are also very useful in the field of robotics and autonomous devices like drones.

2.1.2 Thermal Camera

A thermographic camera (also called an infrared camera or thermal imaging camera or thermal imager) is a device that creates an image using infrared radiation, similar to a common camera that forms an image using visible light. Instead of the 400–700 nanometre range of the visible light camera, infrared cameras are sensitive to wavelengths from about 1,000 nm to about 14,000 nm. The practice of capturing and analyzing the data they provide is called thermography. Basically the higher an object's temperature, the more infrared radiation is emitted as black-body radiation. A thermal camera can detect this radiation in a way similar to the way an ordinary camera detects visible light. It even works in total darkness because ambient light level does not matter. This makes it useful for rescue operations in smoke-filled buildings and underground. For use in temperature measurement the brightest (warmest) parts of the image are customarily colored white, intermediate temperatures reds and yellows, and the dimmest (coolest) parts black. A scale should be shown next to a fake color image to relate colors to temperatures. Their resolution is considerably lower than that of optical cameras, mostly only 160×120 or 320×240 pixels, although more expensive cameras can achieve a resolution of 1280×1024 pixels.

Similar to the common RGB camera, what sets a thermal camera apart are the specifications like resolution, field of view (FOV), frame rate, thermal sensitivity, etc. Thermal resolution is one important specification that reflects how many pixels the camera has on the scene. Spatial resolution is based on detector pixels and the field of view (FOV) spec, combining them to define the area the imager sees at any given moment. Higher resolution means that each image contains more information and more details, and also a greater possibility of obtaining an accurate measurement. Spatial resolution can be used to help define the smallest object size that can be detected. A lower spatial resolution value means better detail and image quality. It is getting much harder to measure with low resolution when the distance between the object and the camera increases. Thermal sensitivity or Noise Equivalent Temperature Difference (NETD) describes the smallest temperature difference that the

Figure 2.3: Thermal image with 256×192 pixels collected in a laboratory at the University of Ottawa.

camera can detect. The highest and lowest temperature that is encountered in the detection range determines the temperature range you need from your thermal imager. Another important specification is the spectral range, which is the range of wavelengths that the sensor can detect. The price of the camera is rising up with the improvement of the different specifications.

Thermographic cameras are much more expensive than their visible-spectrum counterparts, though low-performance add-on thermal cameras for smartphones became available for hundreds of dollars in 2014 [18]. It is not likely to apply expensive thermal camera to the daily life because of their commercial cost. For example, one with 640×512 resolution is over 10,000 USD. On the other side, the thermal camera with low resolution is not capable of detecting the accurate temperature or respiratory features with the distance at around 1m. As a result, we select Rakinda FT20 thermal camera with moderate thermal resolution 256×192, and relatively acceptable price 1000 Canadian dollar as our thermographic system.

Table 2.1: Specs and Prices of Thermal Cameras

Product Name	Thermal Resolution	Frame Rate	NETD	FoV	Price
SEEK Compact	206×156	9Hz		$36°$	249USD
SEEK Compact Pro	320×240	9Hz	70mK	$32°$	499USD
FLIR One GEN3	80×60	8.7Hz	100mK	$55° \times 43°$	249USD
FLIR One LT	80×60	8.7Hz	70mK	$55° \times 43°$	395USD
FLIR One Pro	160×120	8.7Hz	70mK	$55° \times 43°$	530USD
FLIR Boson 320 (core)	320×256	60Hz	60mK	adjustable	1680USD
FLIR Boson 640 (core)	640×512	60Hz	60mK	adjustable	3520USD
FLIR A35	320×256	60Hz	50mK	$13° \times 10°$	5000USD
Optris PI 400i	382×288	80Hz	75mK	adjustable	5150USD
Optris PI 640	640×480	80Hz	75mK	adjustable	8750USD
ICI 8320 P-Series	320×240	60Hz	20mK	$40° \times 30°$	6974USD
ICI 9640 P-Series	640×512	60Hz	20mK	$80° \times 60°$	10552USD

The specifications of the camera are obtained from the official website of the camera producer. The prices of the camera are referred and viewed from the dealer website like Amazon and eBay.
Both the first parameter of Thermal Resolution and FoV is horizontal, and the second parameter is vertical. The single value means it is same in horizontal and vertical direction.

2.2 Object Detection Algorithm

The progress of object detection has gone through a long period and has established some norms to solve the problem. Meanwhile some evaluation metrics have been created to prove the qualities of the architecture. Generally speaking, it is widely accepted that the progress of object detection has generally gone through two historical periods: "traditional object detection period (before 2014)" and "deep learning based detection period (after 2014)" [19].

2.2.1 Machine learning based Detectors

Traditional object detection tasks are usually seen as part of the problems in image processing, and it is common to first find some features based on images or videos. Then the feature representations in combination with machine learning algorithm were designed to detect the target object.

For example, P. Viola and M. Jones achieved real-time detection of human faces for the first time without any constraints [20], where the detector was hundreds of times faster than any other algorithms in its time under comparable detection accuracy. This machine learning

approach for visual object detection is then called Viola-Jones detectors and is widely used because of its very high true-positive rate and very low false-positive rate while running in real time for practical applications. The Viola-Jones detector designed an image representation method called integral image that allows for very fast feature evaluation. The integral image can be computed from an image using a few operations like Haar feature selection filters per pixel. A common Haar feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region. Once computed, any one of these features can be computed at any scale or location in constant time. The second contribution of Viola–Jones object detection framework is a method for building a classifier by selecting a small number of important features using Adaptive Boosting (AdaBoost) formulated by Yoav Freund and Robert Schapire [21]. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner. Viola-Jones detector combines the concepts of Haar-like features, the AdaBoost algorithm, and the cascade classifier to create a system for object detection that increases the speed of the detector by focusing on promising regions of the image.

Another widely using detector applies Histogram of Oriented Gradients (HOG) features descriptor [22]. HOG decomposes an image into small squared cells, computes the histogram of oriented gradients in each cell, normalizes the result using a block-wise pattern, and returns a descriptor for each cell. The HOG detector has long been an important basis of many object detectors [23] and a large variety of computer vision applications for many years. However, HOG has some disadvantages such as that it is difficult to deal with the occlusion problem, and it is difficult to detect when the range of object motion is wide or the object's direction is changing. Consequently, Deformable Part-based Model (DPM) was first proposed by Felzenszwalb to extend HOG feature descriptors and improved the object detection algorithm [24]. DPM is based on HOG detector and uses a sliding window approach, where a filter is applied at all positions and scales of an image to represent an object category. DPM follows the philosophy that objects are composed of parts at specific relative locations, and so it takes star-structured part-based model, defined by a root filter plus a set of parts filters and associated deformation models. Another improvement is a representation of the class of models by a mixture of star models. The DPM detector was also dramatically sped up with a cascade algorithm in [23]. However, these methods highly rely on different features design for different object detection tasks.

2.2.2 CNNs based Two-stage Detectors

Since the rise of convolutional neural networks in 2012 [1], CNNs based approaches has been started to apply for the object detection task. As a result, these CNNs based methods largely enhanced the performance of object detection and then changed the landscape of research in computer vision especially in object detection.

Convolutional neural networks were inspired by biological processes [25] [26] in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. A convolutional neural network consists of an input layer, hidden layers and an output layer. In the feed-forward neural networks, any middle layers are called hidden because their inputs and outputs are masked by the activation function and final convolution. In a convolutional neural network, the hidden layers include layers that perform convolutions. Typically this includes a layer that does multiplication or other dot product, and its activation function is commonly ReLU. ReLU is the abbreviation of rectified linear unit, which applies the non-saturating activation function $f(x) = \max(0, x)$ [1]. This is followed by other convolution layers such as pooling layers, fully connected layers and normalization layers. It effectively removes negative values from an activation map by setting them to zero [27]. It increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer. Another important concept of CNNs is pooling, which is a form of non-linear down-sampling. There are several non-linear functions to implement pooling among which max pooling is the most common. It partitions the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. After several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers where neurons have connections to all activations in the previous layer.

What listed below is a table about different CNNs based two-stage object detection methods. The reason why they are called "two-stage" is because they take coarse to fine strategy and generally solve the object detection problem by two steps. Two-stage methods first generate thousands of region proposals that probably include the objects and then select the best one from multiple candidates to detect the object.

R-CNN

R. Girshick et al. took the lead to apply CNNs by proposing the Regions with CNN features (R-CNN) for object detection in 2014 [28, 29]. R-CNN takes one image as the input, and

Table 2.2: Overview of two-stage object detection methods

Index	Method Name	Features	Limitations
1	R-CNN	Selective search + SVM classifier	Large time cost on region proposals selection for each image.
2	SPPNet	Spatial Pyramid Pooling strategy	Multi-stage training is complex.
3	Fast R-CNN	SPPNet + RoI Pooling (train end-to-end)	Large time cost on selective search for proposals.
4	Faster R-CNN	Region Proposal Network	Low efficiency on selecting proposal candidates.
5	FPN	Feature Pyramid Network	Complex network architecture because of two-stage detection.

The table lists main features and limitations of the two-stage CNN based methods for object detection including the sort-of-the-art methods. The details of the solutions are given in this section.

firstly extracts a set of object proposals as candidate by using selective search [30]. Selective search is a common algorithm that seeks to merge together the perceptual grouping of pixels in a bottom-up hierarchical grouping. It is based on color similarity, texture similarity, size similarity, shape compatibility and their combination to find about 2000 regions of the image. Then each candidate proposals is rescaled to a fixed size (227×227) and fed into a CNN model trained on ImageNet (AlexNet [1]) to extract features. Finally, linear support vector machine (SVM) classifiers are used to predict the presence of an object within each region and to recognize object categories. In addition to predicting the presence of an object within the region proposals, the algorithm also predicts four values which are offset values to increase the precision of the bounding box using regressor. R-CNN yields a significant performance boost on PASCAL VOC 2007, with a large improvement of mean Average Precision (mAP) from 33.7% (DPM-v5 [31]) to 58.5% [29].

The problems with R-CNN is that it still costs a large amount of time to train the networks since around 2000 region proposals per image need to be classified. Besides, the huge number of region proposals also takes a lot space during the training of the model. In the testing process, R-CNN takes about 47 seconds detect each image which is too slow to implement real time. In addition, the selective search algorithm is a fixed algorithm, therefore, some good candidates cannot be shared to save time and space.

SPPNet

In 2014, K. He et al. proposed Spatial Pyramid Pooling Networks (SPPNet) [32]. The main contribution of SPPNet is the introduction of a Spatial Pyramid Pooling (SPP) strategy, which enables a CNN to eliminate the requirement that the input image must be fixed size by generating a fixed-length representation regardless of image size. SPP can maintain spatial information by pooling in local spatial bins and is robust to object deformations. These spatial bins have sizes proportional to the image size like 4×4, 2×2, 1×1, and totally 21 different spatial bins, so the number of bins is fixed regardless of the image size. The outputs of the spatial pyramid pooling are kM-dimensional vectors with the number of bins denoted as M (k is the number of filters in the last convolutional layer). The fixed-dimensional vectors are the input to the fully-connected layer. When using SPPNet for object detection, the feature maps can be computed from the entire image only once. After that, the fixed length representations of arbitrary regions can be generated for training the detectors, which avoids repeatedly computing the convolutional features.

Although SPPNet has effectively improved the detection speed, there are still some drawbacks. First, the training is still multi-stage which is complicated, and it still takes a lot space to store the parameters and weights during the process of training. Second, SPPNet only fine-tunes its fully connected layers while simply ignores all previous layers. Later in the next year, a new CNNs based method was proposed and solved these problems.

Fast R-CNN

R. Girshick proposed a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection in 2015 [33]. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. One of the innovation is that Fast R-CNN enables us to simultaneously train a classifier and a bounding box regressor under the same network configurations. A Fast R-CNN network takes an entire image and multiple regions of interest (RoI) as input. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector. The fully connected layers finally branch into two output layers: one that predicts probabilities of all classes including the background, and another generates bounding box regression offsets for each of the object classes. Besides, all the features are temporally stored in the memory during the training to save extra spaces. Another innovation of Fast R-CNN is the RoI pooling layer. The RoI pooling layer uses max pooling to convert the features into a feature map with a fixed spatial size of $H \times W$ rather than multiple fixed size feature maps in the

SPPNet [32] which can sharply cuts off the computational amounts in the training.

On VOC07 dataset, Fast R-CNN increased the mean Average Precision (mAP) from 58.5% (RCNN) to 70.0% while with a detection speed over 200 times faster than R-CNN [34]. Although Fast R-CNN successfully integrates the advantages of previous R-CNN and SPPNet and makes some improvements, the detection speed is still limited by the time-consuming selective search algorithm that takes around 2 seconds per image runs on CPU for proposal detection.

Faster R-CNN

In 2015, S. Ren et al. proposed Faster R-CNN [35] shortly after the Fast R-CNN. The main contribution of Faster-RCNN is the introduction of Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. RPN is a fully convolutional network that simultaneously predicts object bounds and class scores at each position. This network uses an $H \times W$ spatial window as input from the feature map. Each sliding window is mapped to a lower dimensional feature. Anchor boxes are defined to capture the scale and aspect ratio of specific object classes to be detected. The main reason to use anchor boxes is that all object predictions can be evaluated at once. They help to speed up and improve efficiency for the detection portion of a deep learning neural network framework. Anchor boxes also help to detect multiple objects, objects of different scales, and overlapping objects without the need to scan an image with a sliding window. The RPN is trained end-to-end to generate high-quality region proposals, which are also used by Fast R-CNN. Besides, the RPN and Fast R-CNN are merged into a single network by sharing their convolutional features.

Although Faster RCNN breaks through the speed bottleneck of Fast R-CNN, there is still computation redundancy at subsequent detection stage. The methods of selecting proposal candidates remain potential to improve the efficiency.

Feature Pyramid Networks

In 2017, T.-Y. Lin et al. proposed Feature Pyramid Networks (FPN) [36] on basis of Faster R-CNN. The previous CNNs based detection framework focus on low-resolution and semantically strong features from the convolutional layers to detect objects. FPN combines low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. A top-down architecture with lateral connections is developed for building high-level semantic feature maps at all scales. The single-scale feature map in RPN [35] is replaced by FPN. Thus, it is not necessary to have

Table 2.3: Overview of one-stage object detection methods

Index	Method Name	Features	Limitations
1	YOLO	Regression problem + NMS compression	More localization errors and low localization accuracy mAP 57.9%.
2	SSD	Using a set of bounding boxes	Worse performance on small objects and still low localization accuracy mAP 74.3%.
3	RetinaNet	FPN + a new Focal Loss function	Imbalanced classes.

The table lists main features and limitations of the one-stage CNN based methods for object detection including the sort-of-the-art methods. The details of the solutions are given in this section.

multi-scale anchor boxes on a specific level. Since a CNN naturally forms a feature pyramid through its forward propagation, the FPN shows great advances for detecting objects with a wide variety of scales. Using FPN in a basic Faster R-CNN system, FPN has now become a basic component of the network of many latest object detectors.

Although the merge of FPN and Faster R-CNN achieves state-of-the-art single model detection results on the MSCOCO dataset [36], it is still not fast enough and highly rely on the high performance of GPU because of the complex networks.

2.2.3 CNNs based One-stage Detectors

Different from the two-stage object detectors that generate regions of interests (RoIs) in the first stage and then send the region proposals down the pipeline for object classification and bounding-box regression, one-stage detectors treat object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates as listed in Table 2.2.3. Two-stage detectors like Faster R-CNN can reach the highest accuracy rate but are typically slower in detection. One-stage detectors, on the other side, reach lower accuracy rate but are much faster than two-stage detectors [37].

You Only Look Once (YOLO)

YOLO was proposed by R. Joseph et al. in 2015. It was the first one-stage detector in deep learning era [38]. YOLO treats the object detection as a regression problem by designing a single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. YOLO also takes non-maximum

suppression (NMS) to select one entity out of many overlapping proposal candidates. NMS sorts the candidates according to their confidence scores and update the candidates list by calculating the Intersection over Union (IoU) and keeping the ones that are smaller than the threshold. The detection network has 24 convolutional layers that extract the features from the input followed by 2 fully connected layers that aim at predicting the class probabilities and bounding box coordinates. The base YOLO model processes images in real-time at 45 frames per second. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is far less likely to predict false detections where nothing exists. Later, R. Joseph has made a series of improvements on basis of YOLO and has proposed its v2 and v3 editions [39, 40], which further improves the detection accuracy while keeping very high detection speed.

In spite of its great improvement of detection speed, YOLO suffers from a drop of the localization accuracy compared with two-stage detectors, especially for some small objects. Besides YOLO also struggles to detect close objects because each grid can only detect one object. YOLO's subsequent versions v2 and v3 [39, 40] have paid more attention to this problem.

Single Shot MultiBox Detector (SSD)

SSD [41] was proposed by W. Liu et al. in 2015. The main contribution of SSD is that the network separates the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. SSD combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. The core of SSD is predicting category scores and box offsets for a fixed set of default bounding boxes which are similar to the anchor boxes in the Faster R-CNN [35]. As a result, SSD has comparable accuracy to aforementioned two-stage detectors like Fast R-CNN and is much faster, while providing a end-to-end framework for both training and inference. Compared to other single stage methods, SSD has much better accuracy, even with a smaller input image size and is more robust to various object sizes in the input.

However, SSD still has some drawbacks. For example, SSD produces worse performance on smaller objects, as they may not appear across all feature maps. Increasing the input image resolution alleviates this problem but does not completely address it. Besides, SSD confuses objects with similar categories like animals.

RetinaNet

Though one-stage detectors are faster and simpler compared to the two-stage methods, they do not have the comparable accuracy of two-stage detectors thus far. T.-Y. Lin et al. have discovered the reasons behind and proposed RetinaNet in 2017 [42]. They discovered that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. Because training is inefficient as most locations are easy negatives that contribute no useful learning signal, and the easy negatives can overwhelm training and lead to degenerate models. Consequently they addressed this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. They proposed a novel loss termed as *Focal Loss*, which adds a factor to the standard cross entropy criterion. The proposed *Focal Loss* enables training highly accurate dense object detectors in the presence of vast numbers of easy background examples by focusing more on the hard, misclassified examples and reducing the weights of easy negative samples. *Focal Loss* enables the one-stage detectors to achieve comparable accuracy of two-stage detectors while maintaining very high detection speed.

2.2.4 Face Detection

Deep learning applies multiple processing layers to learn representations of data with multiple levels of feature extraction. This emerging technique has reshaped the research landscape of face detection that could be seen as one of the sub-area of object detection since 2012, AlexNet [1] launched by Alex Krizhevsky et al. Since then, deep learning technique, characterized by the hierarchical architecture to stitch together pixels into invariant face features, has dramatically improved the state-of-the-art performance and fostered successful real-world applications.

Cascade CNN

A cascade architecture built on convolutional neural networks (CNNs) was proposed in 2015 to detect faces [43]. Compared to previous classical Viola-Jones dace detector [20], cascade CNN could apply in real-world face detection, handling large visual variations, such as those due to pose, expression, and lighting. The proposed Cascade CNN operates at multiple resolutions, quickly rejects the background regions in the low resolution stages, and carefully evaluates a small number of challenging candidates in the high resolution stage. The detection stages also take NMS strategy to select the best from overlapping candidates. To improve localization effectiveness, and reduce the number of candidates at later stages, a CNN-based

Table 2.4: Overview of deep learning based face detection methods

Index	Method Name	Features	Limitations
1	Cascade CNN	cascade CNN + NMS strategy	Not an end-to-end training.
2	DDFD	SSD network + Feature Enhance Module	Imbalanced negative samples.
3	MTCNN	cascaded structure + NMS + Facial landmarks detection	Cannot handle the occluded faces.
4	FAN	Face Attention Module	Cannot estimate the facial landmarks.
5	SSH	Context Module	Bad performance on occluded faces and facial landmarks detection.
6	S3FD	A scale-equitable face detection framework	Cannot estimate the facial landmarks.

The table lists main features and limitations of the deep learning based face detectors including the sort-of-the-art methods. The details of the methods are given in this section.

calibration stage was introduced after each of the detection stages in the cascade. The proposed method runs at 14 FPS on a single CPU core for VGA-resolution images and 100 FPS using a GPU. Although using convolutional neural networks to extract the features, Cascade CNN was trained sequentially instead of end to end, which may not be desirable.

Deep Dense Face Detector (DDFD)

In 2015, S. Farfade et al. proposed Deep Dense Face Detector (DDFD) [44] to address the problem of multi-view face detection. The proposed method does not require pose or facial landmark annotation and is able to detect faces in a wide range of orientations using a single model based on deep convolutional neural networks. DDFD uses the same backbone network as SSD network [41], and uses a Feature Enhance Module on top of a feedforward VGG [45] or ResNet [46] architecture to generate the enhanced features. The method, unlike other deep learning object detection methods, does not require additional components such as segmentation, bounding-box regression, or SVM classifiers. In addition, the proposed method is able to detect faces from different angles and can handle occlusion to some extent. However in the dataset, the number of negative samples are 100 times of the positive one, so the imbalance problem within the dataset couldn't be addressed efficiently.

Multi-task Cascaded CNN (MTCNN)

In 2016, Zhang et al. proposed Multi-task Cascaded CNN (MTCNN) to detect and align the face [47]. The proposed framework adopts a cascaded structure with three stages of carefully designed convolutional neural networks that predict face and landmarks location in a coarse-to-fine manner. In the first stage it uses a fully convolutional network to quickly produce candidate windows and their bounding box regression vectors. After obtaining the bounding box vectors, some refinement is done to combine overlapping regions. The final output of this stage is all candidate windows after refinement. In the second stage it refines the proposed candidate windows through a more complex CNN. The Refine Network further reduces the number of candidates, performs calibration with bounding box regression and employs NMS to merge overlapping candidates. Lastly in the third stage it uses a third CNN, more complex than the others, to further refine the result and output facial landmark positions for eyes, nose and mouth. MTCNN achieves superior accuracy over the state-of-the-art techniques on the challenging FDDB [48] and WIDER FACE [49] benchmark for face detection, and AFLW [50] benchmark for face alignment, while keeps real time performance.

Face Attention Network (FAN)

Wang et al. proposed Face Attention Network (FAN) [51] in 2017 to resolve the challenging problem in the face detection that is the occlusion issue due to mask and sunglasses. FAN designed a specific anchor setting together with the attention function based on RetinaNet [42]. To address the occlusion issue, FAN proposed a novel anchor level attention based on the ResNet and FPN network structure. The attention supervision information is obtained by filling the ground-truth box. Meanwhile supervised heatmaps are associated to the ground-truth faces assigned to the anchors in the current layer. The attention mask can enhance the feature maps in the facial area, and diminish what's not in the area. The proposed FAN can significantly improve the recall of the face detection problem in the occluded case without compromising the speed.

Single Stage Headless Face Detector (SSH)

Also in 2017, Najibi et al. proposed the Single Stage Headless (SSH) face detector [52] where the head of the networks has been removed. Most CNN based detector, whether detecting objects or only faces, converts classification network into two stage detection systems. In the first stage convolutional layers proposes a set of bounding boxes for the object and then the remaining layers of classification network referred as 'head' are used to classify the proposals. The head of the classification network can be computationally very expensive

and had to be performed for every proposed bounding box. SSH designed a context module that can incorporate context by increasing the receptive field proportional to the stride of corresponding layer. Additionally, instead of relying on an image pyramid to detect faces with various scales, SSH is scale-invariant by design. SSH can simultaneously detect faces with different scales in a single forward pass of the network, but from different layers. Unlike the current state-of-the-art, SSH does not use an image pyramid and is 5 times faster.

Single Shot Scale-invariant Face Detector (S3FD)

Single Shot Scale-invariant Face Detector (S3FD) [53] proposed by Zhang et al. in 2017, performs superiorly on various scales of faces with a single deep neural network, especially for small faces. In the paper, the authors discussed about the reasons behind the problem of anchor-based methods like few features, anchor scale mismatch, and background from small anchors generating more negative anchors etc. As a result, S3FD proposed a scale-equitable face detection framework to handle different scales of faces well. The method tiles anchors on a wide range of layers to ensure that all scales of faces have enough features for detection. Besides, the authors design anchor scales based on the effective receptive field and a proposed equal proportion interval principle. The methods guarantee that different scales of anchor have the same density on the image, so that various scales face can approximately match the same number of anchors. As a consequence, S3FD achieves state-of-the-art detection performance on all the common face detection benchmarks, including PASCAL face, FDDB and WIDER FACE datasets.

2.2.5 Evaluation Metrics

To evaluate the performance of object detection architectures, both the classification and detection tasks were evaluated as a two-class tasks: e.g. for classification "is there a person in the image?", and for detection "where are the subjects in the image (if any)?". A separate score is computed for each of the classes and there are many metrics like precision and recall to evaluate the object detection results.

For the classification task, a confidence score for each image and each class is computed. For the detection task, the concept of Intersection over Union (IoU) needs to be introduced first. IoU computes intersection of the ground truth and the predicted bounding box over the union of these two bounding boxes. A threshold value is set for the IoU to determine if the object detection is valid or not. A detection is considered a true positive (TP) only if it satisfies that the predicted class matches the class of the ground truth, the confidence score is larger than the threshold, and the predicted bounding box has an IoU greater than

the threshold value. Violation of either the first or third condition makes a false positive (FP). When the confidence score of a detection that is supposed to detect a ground-truth is lower than the threshold, the detection counts as a false negative (FN).

Accordingly, we define precision as the number of true positives divided by the sum of true positives and false positives:

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

Similarly, recall is defined as the number of true positives divided by the sum of true positives and false negatives that is just the number of ground-truths:

$$recall = \frac{TP}{TP + FN} \tag{2.2}$$

As a result, the interpolated average precision (AP) was used to evaluate both classification and detection [54]. The AP summarises the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels [0,0.1,...,1]:

$$AP = \frac{1}{11} \sum_{r \in (0,0.1,...,1)} P_{interp}(r) \tag{2.3}$$

The interpolated precision P_{interp} at a certain recall level *r* is defined as the highest precision found for any recall level $r' \geq r$:

$$P_{interp}(r) = \max_{r' \geq r} p(r') \tag{2.4}$$

where $p(r')$ is the measured precision at recall r'. The calculation of AP only involves one class. However, in object detection, there are usually $K > 1$ classes. Mean average precision (mAP) is defined as the mean of AP across all K classes:

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{2.5}$$

2.3 Vital Signs Monitoring

Monitoring the vital signs of human being is an important task for medical diagnosis. Current methods require the sensing device to be attached to the subjects' body thereby constraining or causing them discomfort and thus potentially affecting the measurement. Contact based measurements of respiratory rate (RR) typically consists of electrophysiological measure-

ments [55] and pressure sensors [56]. For example, to measure the respiration rate of the subject, typically a breathing belt is attached to the chest of the subject, which makes it inconvenient for the subject to move freely and even causes the discomfort. Similarly, to measure the heart rate of the subject, a fingertip monitor is needed to check the pulse. Or one can check his or her own heart rate by pressing the first and second fingertips firmly but gently at the wrist and counting the pulse for 60 seconds. As a result, the contact-based measurements require the wearing operations from the subject, and it relies on specific sensors. So the non-contact methods for vital signs monitoring are worthy of further study especially now during the Covid-19 pandemic. Camera-based vital sign estimation allows the contactless assessment of important physiological parameters including respiration rate, heart rate, etc.

2.3.1 Respiration Rate Estimation

Radar-based methods

One approach of remote respiratory rate estimation is using the Doppler radar. Non-contact detection characteristic of Doppler radar provides an unobtrusive means of respiration detection and monitoring [57]. Robustness of Doppler radar against environmental factors, such as light, ambient temperature, interference from other signals occupying the same bandwidth, fading effects, and other environmental constraints strengthen the possibility of employing Doppler radar in long-term respiration detection and monitoring applications such as sleep studies.The Doppler effect [58] occurs when there is change in frequency in the radiated or reflected radio wave due to the movement of an object. When a continuous wave is transmitted towards an object, the reflected signal is either frequency modulated or phase modulated due to the movement of the object. By comparing the received and transmitted signals, the change in frequency and phase can be derived. Changes in the reflected signal due to the motion of the subject's chest during the respiratory cycles can be extracted and then used to estimate the breathing rate. In addition, Fourier and wavelet transform approaches were used in extracting the respiration rate in each particular breathing cycle. The extracted peak frequency can be used to approximate the breathing rates especially under a normal breathing condition. Doppler radar is highlighted as an alternative approach not only for determining respiration rates, but also for identifying breathing patterns and tidal volumes as a preferred non-wearable alternative to the conventional contact sensing methods.

However, one problem of respiration rate estimation using radar is that the distance between the subject's chest and radar must be measured before each respiration rate estimation. So distance estimation should be correct, and the distances were manually measured

by researchers in the experiment.In addition, radar-based approaches are restricted to monitor physiological parameters in stationary settings and indoors, where it is easier to ensure stillness of both the person and of the hardware installation, limiting their application in real-world deployments.

Camera-based methods

Another contactless approach to monitor the respiration is based on cameras. Plethysmography measures changes in volume in different areas of human's body, and it can be done by means such as variations in air pressure, impedance, or strain. Photo-plethysmography (PPG), introduced in the 1930's [59] uses light reflections or transmission and is the least expensive method and simple to use. PPG is based on the principle that blood absorbs light more than surrounding tissue so variations in blood volume affect transmission or reflections of light correspondingly. Applications of PPG include monitoring of oxygen saturation (pulse oxymetry), heart rate (HR), respiration rate (RR), blood pressure, cardiac output, assessment of autonomic functions and detection of peripheral vascular diseases [60]. Remote, non-contact pulse oxymetry and PPG imaging have been explored only relatively recently [61, 62].

The body surface movements caused by respiration modify the path length of the illumination light, and the subsequent changes of the reflected light indicate the timing of respiration events. By capturing the images with camera, the image sensors collect the reflected light signal along with noise due to artifacts. As a result, the corresponding variation of the brightness of the moved chest and abdomen area due to breathing indicate the respiratory events. Thus, the desired signal can be formed over time from a series of captured images [63]. Wu et al. proposed the method "Eulerian Video Magnification (EVM)" [64] to reveal the small motions in videos that are difficult or impossible to see with the naked eye and display them in an indicative manner. With the help of EVM, the motion of chest or abdomen area could be enhanced and so the corresponding breathing behavior could be extracted from the thermal videos or infrared videos [65, 66].

However, there are many limitations within this method. For example, the image acquisition device is a near-IR enhanced camera that is sensitive to light in the visible and near infrared region. Also, the distance between the device and the subject has not been made clear in the paper [65]. Another issues is that the methods based on RGB camera require an ambient source of light, and does not work properly in dark places or under varying lighting conditions [63]. Detecting the movement of shoulder or chest caused by the respiratory cycles may have much space to improve since the corresponding variations are sometimes too subtle to be captured by cameras especially when the subject is wearing thick clothes.

Sometimes this kind of brightness variation can be easily interrupted by the noises from environment and so the extracted signals are not purely the reflections of the breathing cycles. Another problem remained unsolved is that the distance between the subject and camera must be within short range or the mentioned variations are difficult to extract even by advanced algorithms such as EVM. The distance has real impact on the potential real-world applications of the approach.

In addition to the motion-based detection like chest based or abdomen based methods, respiration could be detected using thermal-based methods. The work on thermography has shown that it is possible to track respiration in a contactless manner by monitoring the temperature changes around the nostrils which are caused by inhalation and exhalation during breathing cycles [67, 68]. Thermal imaging technique can measure temperature in a passive way and does not require light sources. Similarly, Bennett et al. used EVM to detect breathing rate from thermal video [69]. With the development of imaging devices and decrease of the price of cameras like infrared and thermal cameras, using these special cameras to detect respiration rate becomes feasible and promising.

For example, Cho et al. proposed a method to track respiratory rate in high-dynamic range scenes using mobile thermal imaging [70]. The main contribution of the paper is that the optimal quantization of the mapping of temperature to pixels was designed, and the thermal voxel-based respiratory rate estimation to enhance the respiratory signal quality was proposed. Similarly, Jiang et al. used both visible light imaging and infrared imaging to detect respiratory infections [71]. The authors proposed a portable non-contact method to screen the health condition of people wearing masks through analysis of the respiratory characteristics. The device mainly consists of a FLIR one thermal camera and an Android phone. The device was designed to help identify those potential patients of COVID-19 under practical scenarios such as pre-inspection in schools and hospitals. The health screening is performed through the combination of the RGB and thermal videos obtained from the dual-mode camera and deep learning architecture. A respiratory data capture technique for people wearing masks by using face recognition is firstly proposed. Then, a bidirectional recurrent neural network with attention mechanism is applied to the respiratory data to obtain the health screening result.

2.3.2 Heart Rate Estimation

Non-contact and low-cost measurements of heart rate (HR) are highly desirable for telemedicine [63]. HR is often measured using contact based optical sensors that use PPG i.e. the variation of transmissivity and/or reflectivity of light through the finger tip as a function

of arterial pulsation [72, 73], followed by different signal post-processing approaches [74]. This approach works due to the different absorption of certain frequency by hemoglobin in the blood, compared to the surrounding tissue such as flesh and bone. Non-contact based optical sensors, have been used to measure HR from a video of a human face [73], by looking at the variation of average pixel value of the green channel in the subject's forehead [75]. Though attractive in principle, many of these methods accomplish noise reduction using linear filters, which are ineffective in the event that background noise falls within the same frequency band as the physiological signal of interest [76].

Wu et al. in 2000 proposed the experimental setup and preliminary results of a charge-coupled device (CCD) based Photoplethysmographic Imager (PPGI) which has been shown to be capable of assessing various disorders of the peripheral venous system by standard test methods derived from the classical photophethysmographic practice in a noninvasive and non-contact way [77]. The PPGI is a computer-based imaging system to visualize the skin vessels and analyze the local changes of dermal blood volume. Their experiment results show that the system performs as well as the currently available commercial PPG system. With this pioneering system, Wu et al. demonstrated that camera-based technology could indeed be used for non-contact measurement of heart rate.

Poh et al. [78] explored the possibility to measure HR from face videos recorded by a web-cam in 2010. They detected the region of interest (RoI, i.e. the face area) using Viola-Jones face detector and computed the mean pixel values of the RoI of each frame from three color channels. Then Independent Component Analysis (ICA) was applied to separate the PPG signal from the three color traces, and the PPG signal was transferred into frequency domain to find the HR frequency. Wu et al. in 2012 proposed Eulerian Video Magnification (EVM) [64] that can visualize the flow of blood as it fills the face which enables the researchers to measure the HR [79]. Bennett et al. also used the EVM to extract HR from thermal video with a wide temporal band-pass filter and low amplification factor as well as with a narrower, targeted temporal band-pass filter and higher amplification factor [80].

In 2014, Li et al. proposed a framework which utilizes face tracking and Normalized Least Mean Square adaptive filtering methods to counter the influences caused by the subjects' motions and illumination variations when the videos are recorded [81]. The framework was designed to measure heart rate remotely from face videos under more challenging conditions. The proposed framework for HR measurement from facial videos in realistic human computer interaction (HCI) situations includes RoI detection and tracking, illumination rectification, non-rigid motion elimination and temporal filtering. The researchers also used the method for long term heart rate monitoring in a game evaluation scenario and achieved

promising results.

He et al. used EVM to analyze the HR and applied the method to detect the falls [82] in 2018. The proposed system could be used for remote patient monitoring and health informatics for improved detection and identification of the causes of falls. However, these methods using camera to estimate the HR have the high requirement on the light condition and illumination. Besides, the motion artifacts and noises generated during the detection have a large impact on the final HR measurement [83]. One of the most common issues for contactless HR measurement is about the RoI selection. Traditional methods like manual selection and tracking algorithms are inefficient. Additionally, most of the recent face detectors cannot handle the difficult situations. Our methods, as a result, are proposed as a potential solution for these problems based on our experiments which will be discussed in the later chapters.

Chapter 3

Methodology

This chapter guides through the methodology developed for respiratory signal detection using depth camera system and vital signs monitoring using thermal camera in details. Section 3.1 explains the method of using deep leaning based object detection from depth camera to estimate respiratory signals. The details about how the whole system works and the dataset creation will be introduced in this section. In the second part, Section 3.2 introduces the thermal camera based system aiming at monitoring several different vital signs of the subject. We present the comparisons between different face detection algorithms based on our own dataset.

3.1 Depth Camera Based System

As one of our major contributions, we present the depth camera based system which is designed to remotely monitor the respiratory signals of the subjects at hospital. We attempt to locate the position of the subject in the ward based on the depth frames, and then extract the respiratory-like signal from the radar collection. The detailed experiment information will be introduced in Section 4.1. This chapter follows the method involved in collecting data and creating the dataset. Our primary motive for building the dataset was to use it to train the deep learning model and detect the objects in an automatic way. Our method is based on existed deep learning networks and trained on our own dataset.

3.1.1 Depth-Radar System

The whole depth-radar monitoring system is composed of one depth camera, one radar and one mini-computer. The mini-computer is responsible for running the program and storing the data. Both the depth camera and the radar are connected to the mini-computer and they

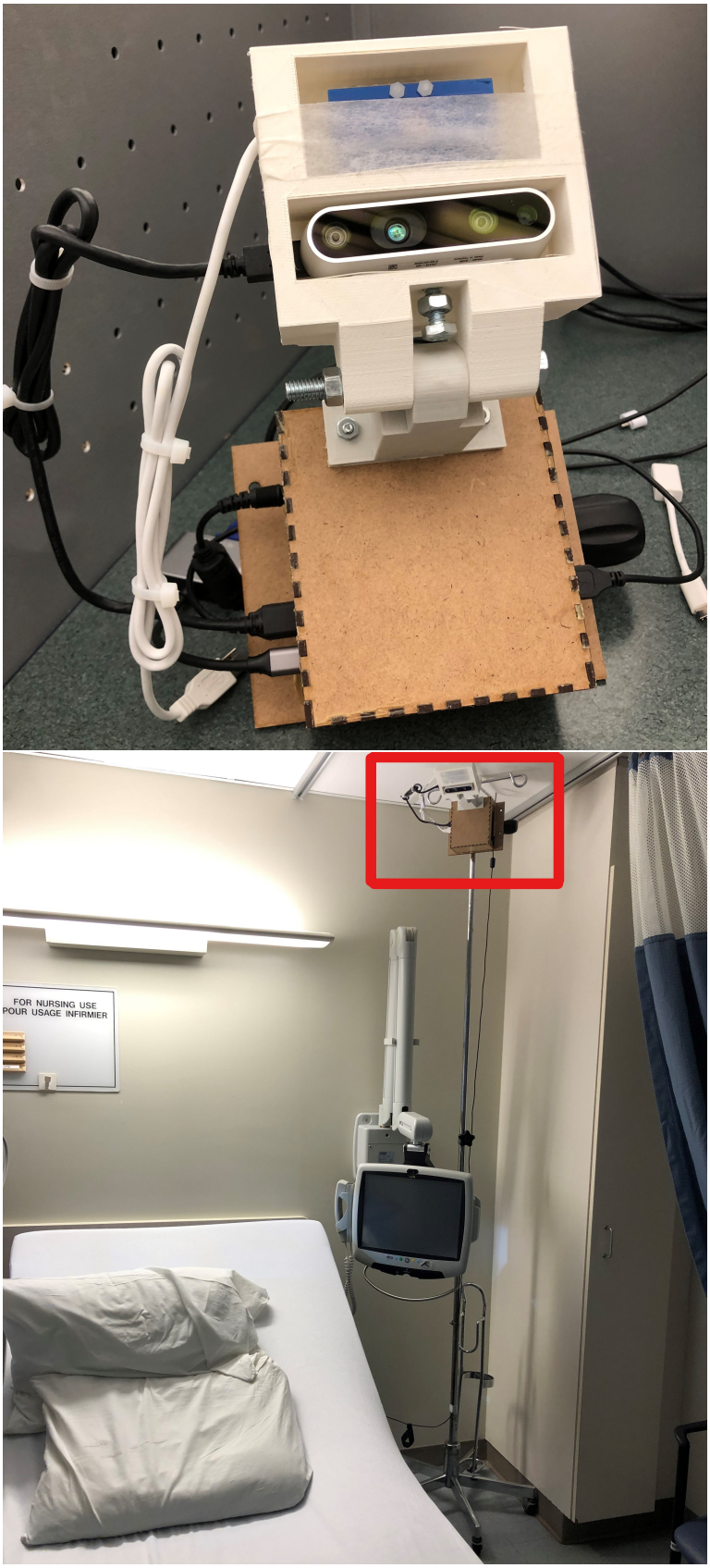
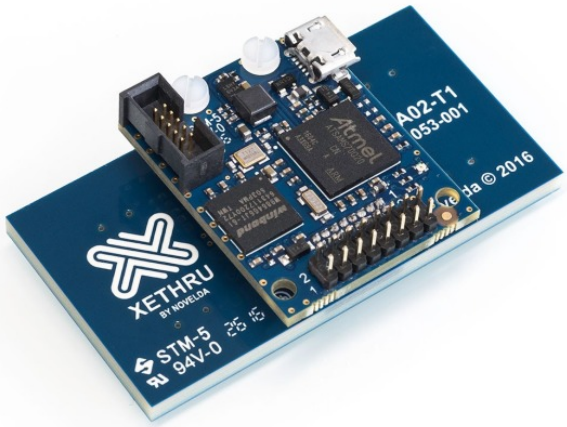


Figure 3.1: The depth-radar monitoring system.

(a) RealSense D435 depth camera



(b) X4-M03 radar

Figure 3.2: Depth camera and radar device used in our monitoring system.

are running simultaneously to have the same timeline. The depth camera collects the depth frame to predict the distance between the person and the radar. While the breath-like signal is collected by the radar.

We are using the Intel RealSense D435 depth camera, illustrated in Figure 3.2a, to capture the depth frames in the scene like a ward of a hospital or an elderly home. Before we try to get the respiratory signal from the patients, we determine patient positions relative to the camera at the very beginning. With the help of the depth camera, it is easy to obtain the relative position in the ward. According to [84], the depth quality of the RealSense D435 depth camera is really high with an absolute error less than 2 percent when the target is up to 2 meters. So, once we find the target in the view, the depth extracted can be directly used for other purposes. Another reason of using the depth camera is to take care of the identity privacy of the patients. It is vital that the human facial details cannot be obtained and recovered from the original depth data, so the patients have nothing to worry about during the treatment. Usually we set the device close to the bed and make the view of camera cover

the area of patient's movement just like illustrated in Figure 3.1. Our research is focused on the estimation of the subject's respiratory signal during the experiment.

As a part of the monitoring system, ultra wide band (UWB) radar is widely used in biomedical field nowadays. It is a short-range high-resolution radar emitting electromagnetic waves with ultra-wide frequency band using relatively low frequencies. The exploitation of the ultra-wide frequency band provides the high resolution of the radar, which enables UWB radars to detect respiration motion. The electromagnetic waves emitted in the frequency bandwidth up to 1.5GHz can penetrate through standard building materials with acceptable attenuation. Moreover, UWB radar systems can be really light and have small form factor. We use Xethru X4, illustrated in Figure 3.2b, that is an UWB impulse radar. It provides sub-mm movement sensing accuracy at distances from 0 to 10 meters depending on target size. Its accuracy is about 1mm, which means a ultra-high spatial resolution for simultaneous multi-object tracking and its human presence simultaneous tracking range up to 10m.

3.1.2 Data Collection

The depth camera based system includes a Intel RealSense D435 depth camera, a Xethru X4 ultra wide band (UWB) radar, and a minicomputer Intel Nuc as introduced in Section 3.1.1. The system is controlled by the minicomputer to execute the commands to collect the data and store the data in the hard drive. In order to keep both the depth frames from depth camera and radar data from UWB radar collecting simultaneously, both radar and camera data are recorded together with time indexes. Since that we set the depth camera collecting data in 3 frames per second (fps), and so we have around 180 frames in one minute. Similarly, the sampling rate of radar is set to 17 fps, and duration of each data package is 1 minute. Even though the depth camera's sampling rate is different from the UWB radar sampling rate, we encapsulate the data minute by minute which means that both the depth frames and radar data are collected and stored in the format of one minute package.

Depth Data

In order to develop a system for continuous monitoring of people, we first had to collect data for the period of several days or weeks and storing depth images requires a lot of space. The original depth data collected by the camera are in 16 binary bits and the size of each frame is 1280×720 [84] which takes around 14 Mb for each frame. While collecting the depth frame, the imencode function in OpenCV [85] is used to compress the image and store it in the memory buffer that is resized to fit the result. Then we decided to use lossless compression Portable Network Graphics (PNG) format in the compression encoding because

we want to keep as much depth information of each pixel in the image as possible. However, PNG format compresses digital images at a larger file size while reserves the details so the size of each original depth frame is around 700 Kb saving half of the space compared to the uncompressed original data. Similarly, to save the space that each frame will take of the storage while maintaining the basic function, the sampling rate is set to 3 fps. Though the sampling rate cannot change too much in a long continuous period, the number of depth images collected in each one-minute packages that we compress and store varies a little bit due to the system delay. It is hard to directly process the invisible distance data unless we change the format to a way that is visible to us human being. Illustrated in Fig. 3.1, we transfer the 16 bits data into 8 bits data and finally get a depth-based grayscale image. We use the proportionate matching strategy to proportionally match the 16-bit depth distance into the 8-bit value ranging from 0 to 255 pixel by pixel. And the grayscale image is lossyly compressed using JPEG format to minimize the image size and save the space.

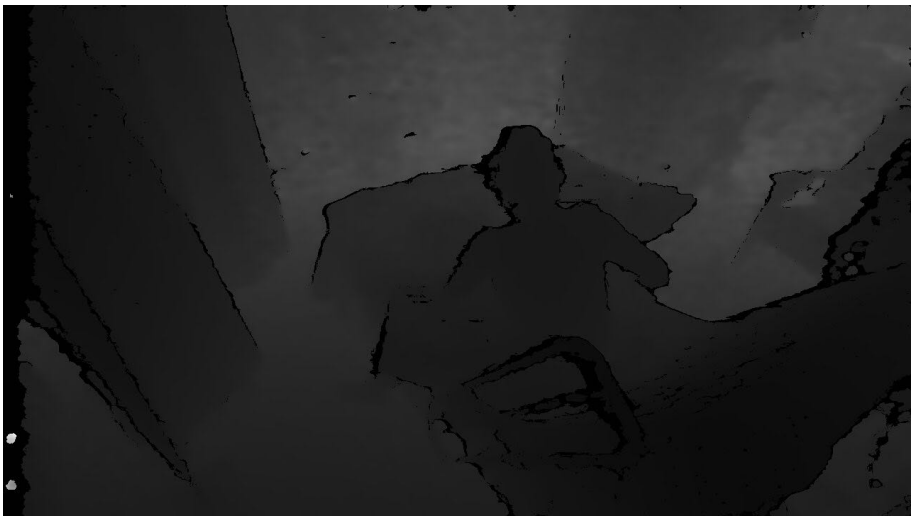


Figure 3.3: Grayscale depth image (1280×720)

Radar Data

The original data from Xethru X4M03 radar modules are received as individual frames and each frame consists of an array of numbers corresponding to the amplitude of the received signal at given time instances. These time instances correspond to distances as explained above at Section 3.1.1 and each frame is saved as a 2-dimension matrix with the first column corresponding to the timestamp when the radar data was received. For the remainder of this thesis, the time instances within each frame are referred to as fast time and the times between each frame are referred to as slow time. An example of the data format corresponding to the

X4 radar can be seen in Figure 3.4. The radar data for a specific time window are organized as an M $\times$ N matrix where M is the number of recorded frames (slow time) and N is the number of points (range bins) in the fast time direction. Each range bin indicates 5 cm length in real world space, and Xethru X4M03 radar has a detect range up to 10 m, so there are close to 200 range bins in radar record. The sampling rate of radar is set to 17 frames per second, and duration of each package is 1 minute. So the radar data found within the 180 range bin have a fixed length of 60 seconds.

		Fast time					
Slow time	Timestamp(ms)	Distance d1	d2	...	...	d(N-1)	d(N)
	254.12	7.0195e-06+2.1557e-06i	5.7537e-07+6.0299e-06i	...	...	-0.00021081+0.00019418i	-3.9731e-05+0.000357i
	269.9	6.1373e-08-2.1711e-05i	-2.6851e-07-2.3958e-05i	...	...	-0.00020656+0.00023399i	-2.6444e-05+0.00040573i
	276.64	3.7821e-06+5.3087e-06i	-1.6601e-05+5.3394e-06i	...	...	-0.00020371+0.00024688i	-3.7606e-05+0.0004245i
	281.08	-2.4434e-05-4.8791e-06i	-2.2593e-05-3.8435e-06i	...	...	-0.00021681+0.0002391i	-3.1691e-05+0.00041293i
	311.97	3.4522e-07-2.089e-05i	-9.812e-06-1.1362e-05i	...	...	-0.0001923+0.00022037i	-2.8845e-05+0.00039308i
	357.11	6.6052e-06+1.0679e-05i	-9.8196e-07+1.8542e-05i	...	...	-0.00020202+0.00022329i	-3.4054e-05+0.00038985i
	415.68	-5.5466e-06-2.57e-06i	-2.8431e-05-6.3751e-06i	...	...	-0.00019217+0.00021679i	-1.855e-05+0.00038089i
	...	...	...	...	...	...	...
	...	...	...	...	...	...	...
	65039	-2.4733e-05-2.9689e-06i	-2.7579e-05+4.9712e-06i	...	...	-0.00019696+0.00023495i	-1.5489e-05+0.00039882i
	65097	-3.4752e-06-1.4453e-05i	-1.723e-05-1.7576e-05i	...	...	-0.00016443+0.00021139i	-2.6467e-06+0.00037095i
	65159	-1.0142e-05-1.4162e-05i	-1.9816e-05-2.5777e-06i	...	...	-0.00020235+0.00023793i	-2.6275e-05+0.0003949i
	65222	-1.9946e-07+2.2585e-05i	-2.3467e-05+1.4215e-05i	...	...	-0.00016588+0.00021666i	-1.5059e-05+0.00038757i

Figure 3.4: Data format used for collecting data from the Xethru X4M03 module.

3.1.3 Dataset Creation

VGG Image Annotator (VIA)

We start with building our own dataset based on the images collected from the depth camera system after obtaining the data from the hospital. The depth image dataset is designed to train the deep learning based neural network for the task of object detection. We annotated the depth frames through VGG Image Annotator (VIA) tool [86] and labeled more than thousands of depth images.

In our training dataset, polygon is selected to draw the outline of the subject just like illustrated in Figure 3.5. The training dataset includes different postures of the 6 subjects from the local hospital, mainly lying in the bed. Since that the subject is usually covered with the blanket when he or she is lying in the bed, we draw the polygon outline based on the upper body which is obviously visible to the human eyes. Another reason behind this is that the chest and shoulder part of the body play an important role in defining the interested region to detect the respiratory rate via UWB radar. Plus, the class of the selected region is labeled as person so there are totally two different classes of the depth image. One of the class is person, and another one is the background.

VGG Image Annotator is a simple and standalone manual annotation software for image, audio and video. The VIA software allows human annotators to define and describe

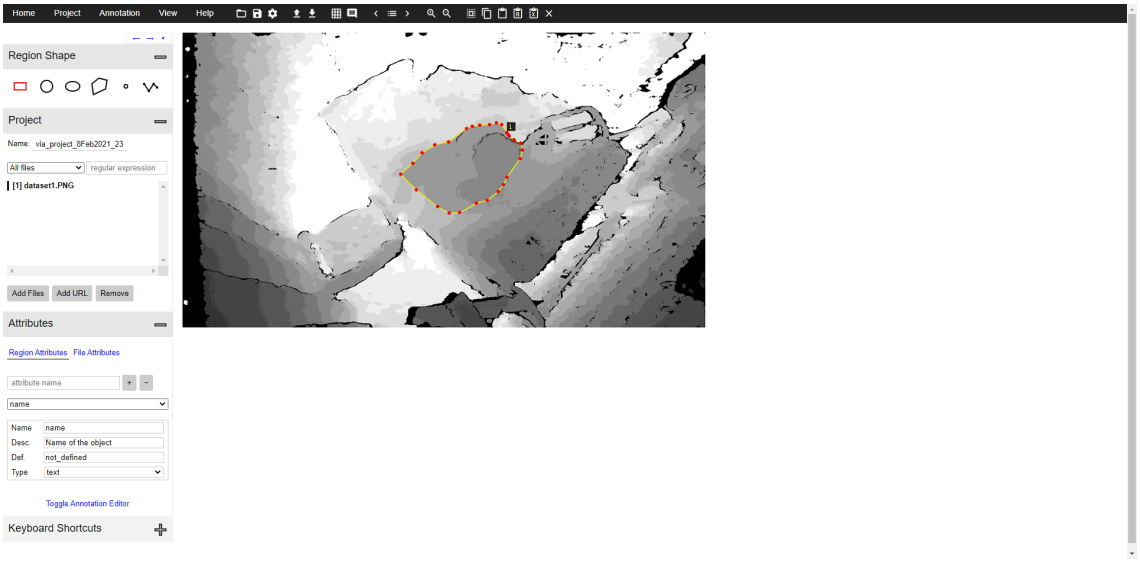


Figure 3.5: Screenshot of the VIA annotation tool.

regions in an image. The manually defined regions can have one of the following shapes like rectangle, polygon and point. Rectangular shaped regions are very common and are mostly used to define the bounding box of an object. Polygon shaped regions are used to capture the boundary of objects having a complex shape.

COCO Annotator

In addition to the VIA tool used to create our own datatset mentioned previously, we also used COCO Annotator [87] to label and build a smaller dataset with keypoint labels based on depth images. Because at the time we created our first dataset, the aforementioned VIA tool was not capable of labeling the point shape that is used to define keypoints in the depth images. Besides, we decided to firstly detect the mask of the human body and then to try the keypoints detection only if the mask segmentation performance satisfies the requirements for our task. After the testing on our first mask dataset, we found that COCO Annotator provides many distinct features including the ability to label an image segment (or part of a segment), labeling objects with disconnected visible parts and discrete keypoints, efficiently storing and exporting annotations in the COCO format. The annotations are all stored using JSON where the detailed information of COCO format dataset will be introduced in Section 4.2. COCO Annotator is also a web-based image annotation tool designed for efficiently label images to create training data for image localization and object detection.

The annotation process is delivered through an intuitive and customizable interface and provides many tools for creating accurate datasets. COCO Annotator allows us to annotate images using curves or polygons. Besides, we add the keypoints annotation in our new

dataset, mainly labeling the neck, left shoulder, right shoulder, and check points on the depth images. While the mask segmentation is almost the same with VIA tool.

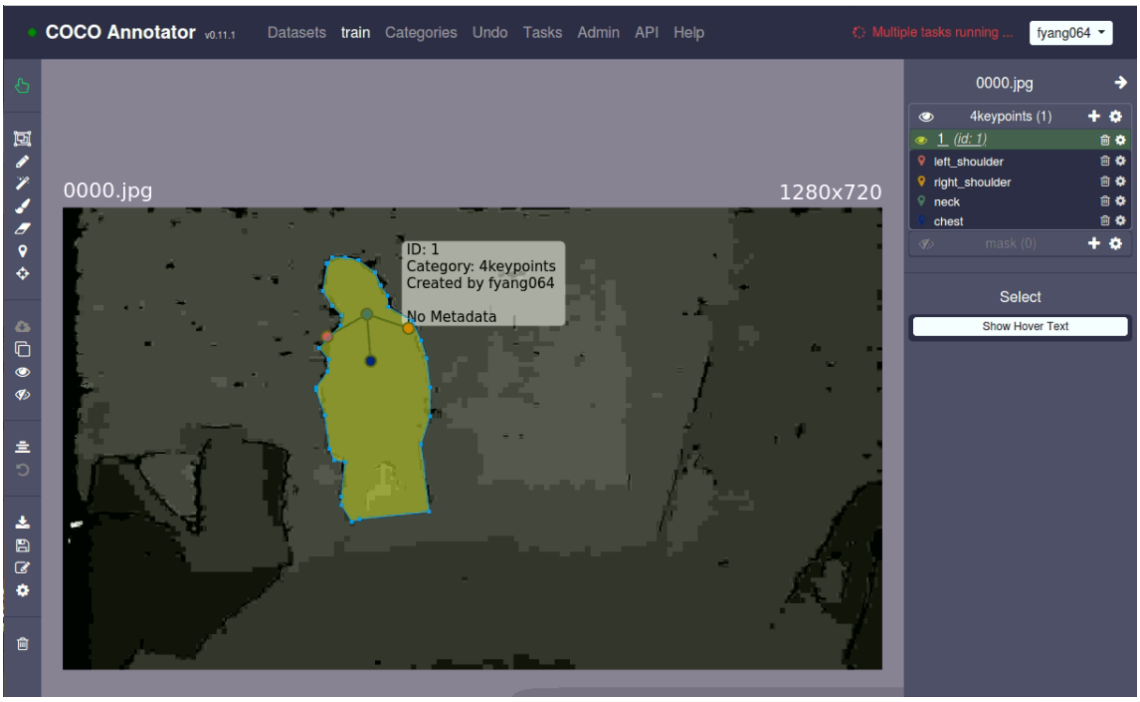


Figure 3.6: Interface of the COCO Annotator.

3.1.4 Body Tracking Method

The whole depth camera based system aims at monitoring and estimating the respiratory rate of the subjects in the hospital ward. The first part of the method is detecting object from the blurry grayscale depth images. While the second part is using radar data to extract the desired signals. Figure 3.7 illustrated that two streams of the method are connected, and the second one is depending on the predicted result of the first.

To predict the location of the subject, we applied a deep learning based neural network Mask R-CNN to detect the human body in the first stream. Since the subject is lying in the bed and covered with a blanket, the contour of the body must be first found and predicted rather than only the bounding box. Mask R-CNN proposed by He et al. in 2018 [88], is a extended and improved version of the framework of aforementioned Faster R-CNN [35]. Compared to the previous Faster R-CNN, Mask R-CNN adds a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. The approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The framework of Mask R-CNN is illustrated as Figure

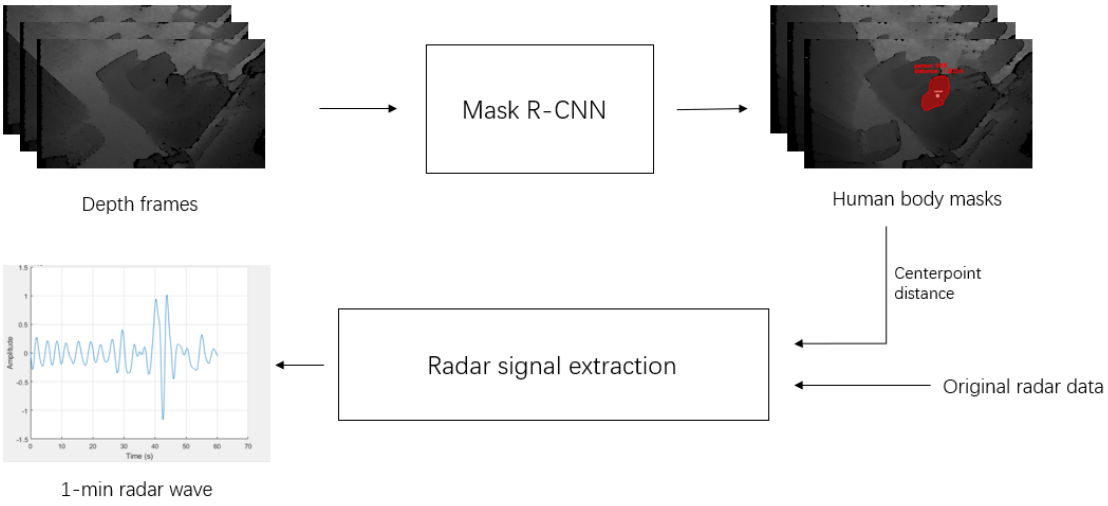


Figure 3.7: Workflow of the whole system including body tracking and RR detection.

3.8.

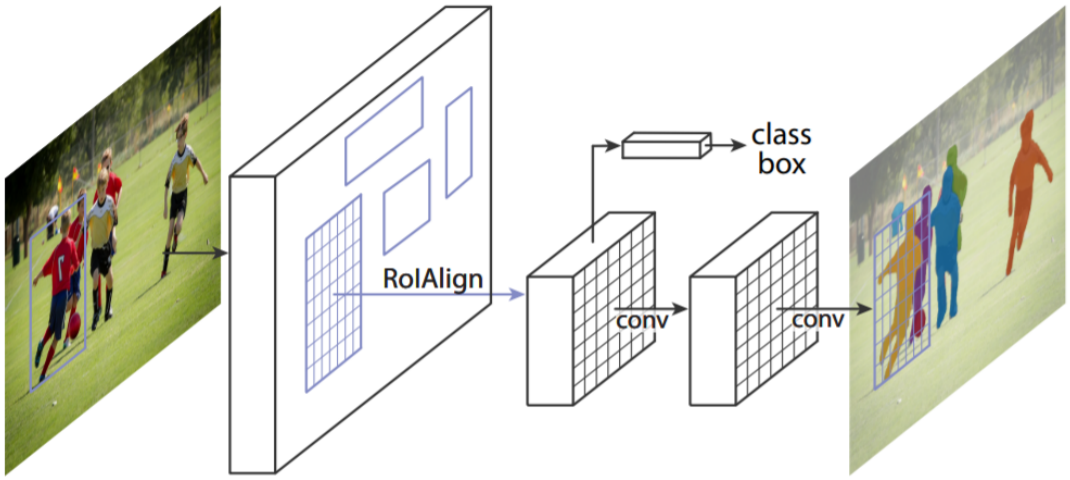


Figure 3.8: The Mask R-CNN framework for instance segmentation cited from [88].

We applied ResNet-50 as the network backbone which extracted features from the final convolutional layer of the 4th stage. The extracted feature maps are fed into a region proposal network (RPN) based on ResNet-50 to create RoIs. Mask R-CNN predicts an $m \times m$ mask that encodes an input object's spatial layout from each RoI using a fully convolutional networks [89]. The author proposed an RoIAlign layer to properly align the extracted features with the input using bilinear interpolation strategy so that the pixel-accurate masks could be predicted. After the operation of aligning the features with the original image, the network

head deconvolves and enlarge the feature map to the input size facilitating the accurate masks prediction. Mask R-CNN adopts the same two-stage procedure as aforementioned Faster R-CNN [35], with an identical first stage which is RPN proposing candidate bounding box. In the second stage, in parallel to predicting the class and box offset, Mask R-CNN also outputs a binary mask for each RoI. In training process, a multi-task loss on each sampled RoI is defined as:

$$L = L_{cls} + L_{box} + L_{mask} \tag{3.1}$$

where the classification loss L_{cls} and bounding-box loss L_{box} are identical as those defined in Faster R-CNN [35]. The classification loss L_{cls} is log loss over two classes that is object versus not object. The bounding box regression adopts the box's center coordinates and its width and height as parameters and uses the robust loss function smooth L1 to regress L_{box}. Mask loss L_{mask} is defined as the average binary cross-entropy loss and applies a per-pixel sigmoid.

Since that the input of our method to track the human body is depth-based grayscale images, it needs to modify and customize the network settings to transfer learning from the previous prediction model trained on common RGB images which will be introduced in Section 4.3.1.

Besides, the purpose of our detection model is not only tracking the human body continuously but also estimating the distance between the target and the sensor. Mask R-CNN generates the mask of the incomplete human body that is much like the silhouette of the person mainly including the head, the shoulders, and chest of the subject since the subject's body are often covered with blanket in our scenario. The mask of subject represents a region that could be distinguished in the image. We can calculate the centroid point of the mask, as known as image moments, by recreating a binary image based on the masks generated from the Mask R-CNN.

An image moment is a certain particular weighted average (moment) of the image pixels' intensities, or a function of such moments. For a 2D continuous function $f(x, y)$ the moment (sometimes called "raw moment") of order $(p + q)$ is defined as:

$$M_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \tag{3.2}$$

for p, q = 0,1,2,... Adapting this to scalar (greyscale) image with pixel intensities $I(x, y)$, raw image moments M_{ij} are calculated by:

$$M_{ij} = \sum_{x} \sum_{y} x^i y^j I(x, y) \tag{3.3}$$

The image properties like centroid point coordinates derived via raw moments is defined as :

$$(\bar{x}, \bar{y}) = (\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}) \tag{3.4}$$

In the end of body tracking stream, the coordinates of the geometric center point calculated from the Mask R-CNN is then back-forwarded to the original depth image and extract the depth information of the pixel.

3.1.5 Respiratory Signal Detection

Mask R-CNN framework helps us to find the location of the subject in the room and the depth image give out the distance between our monitoring system and the subject. The estimated center point that could be seen as the position of the RoI consisted of chest, is what we are going to search in the radar detection range bin. X4-M03 UWB radar in the system collects data varied in 180 range bins. The key of radar respiration monitoring is to identify the range bin where the subject's chest is located and to extract the breath like signal. Once we get the distance from the body tracking framework, we can look into the corresponding bin by using the estimated distance, extract and process the radar signal from the estimated range bin and the radar signals from the range bins in its proximity, as illustrated in Figure 3.7. In addition the algorithm is applied for screening the real respiratory signal that has the certain curve shape or the number of peaks corresponding to the respiratory rate.

3.2 Thermal Camera based System

The thermal camera based system is designed for measuring subjects' vital signs including body temperature, respiratory rate (RR), and heart rate (HR). The remote vital signs measurement consists of ROI detection and signal processing. The steps of remotely estimating HR, RR, and body temperature is shown in Figure 3.9. In this study, both RGB and thermal camera are used simultaneously to measure HR, RR and body temperature. The RGB camera is used to detect and track human subject's face, and thus locate ROIs from the video sequence to estimate HR. And with the help of image alignment process, these ROIs can be located on the simultaneous thermal video frames and then HR, RR, and the body temperature can be

estimated.

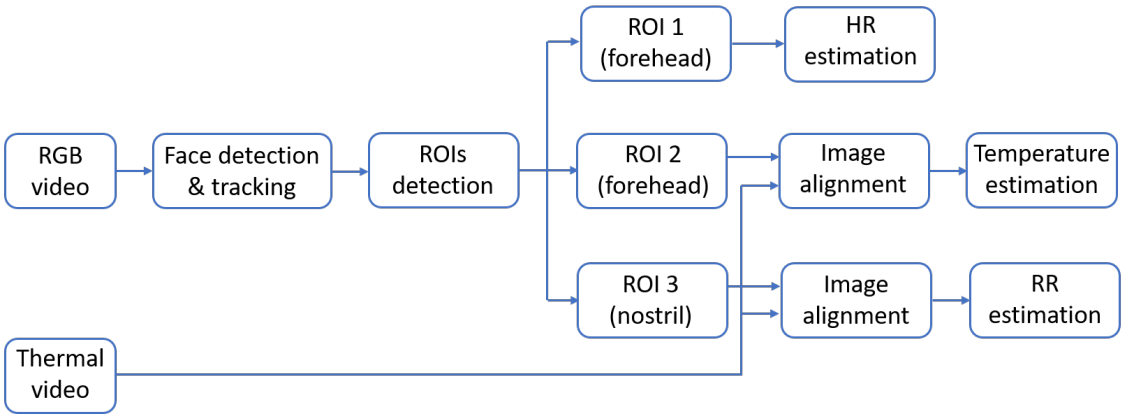


Figure 3.9: Workflow of the thermal camera based system.

3.2.1 Face Detection

We apply different face detectors to detect faces from the RGB frames. By comparing different face detectors like MTCNN [47], S3FD [53], PyramidBox [90], and the RetinaFace [91] framework in our scenario, we find the best one that is suitable for detecting the faces with the masks. Because in most of the time more than half of the faces are occluded with the mask in our use case. In addition we also need to detect the facial landmarks of each face. We find that RetinFace can handle the problems very well in various aspects which the result will be given and discussed in Chapter 4.

RetinaFace [91] proposed by Deng et al in 2019, ranks second in the competition of WIDER FACE (Hard). It is a practical single-stage deep learning based face detector working on each frame of the video. We choose it because its advantage in harder face detection like occluded faces over other frameworks which will be introduced in next chapter. Additionally, RetinaFace can detect multiple faces in one image which makes it possible for our method to detect multiple subjects' vital signs at the same time.

RetinaFace is designed based on the feature pyramids with independent context modules as illustrated in Figure 3.10. Feature pyramid is just like the FPN mentioned before [36]. Context modules on feature pyramids is applied to enlarge the receptive field from Euclidean grids and enhance the model's contextual reasoning power for capturing tiny faces [92]. Because convolutional features at higher layers tend to have larger receptive fields, smaller receptive fields necessitate the use of lower layer features. Besides smaller receptive fields do better for small faces, because the entire face is visible. Adding context helps to accurate

detection on small instances and find low-resolution faces in our case. Following the context modules, we calculate a multi-task loss for each anchor.

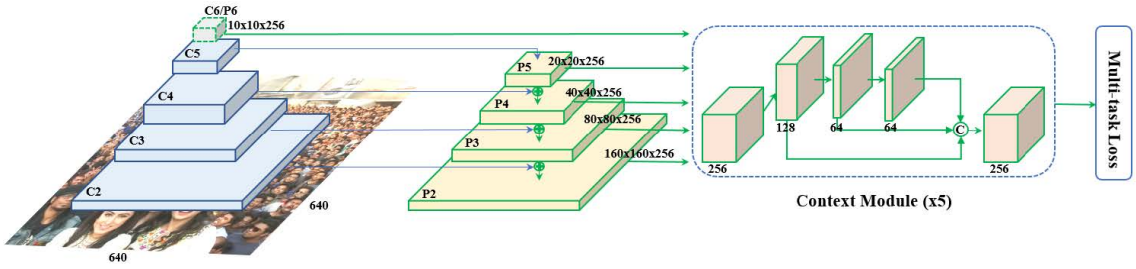


Figure 3.10: An overview of the single-stage RetinaFace localisation approach cited from [91].

RetinaFace is training with manually annotated five facial landmarks on the WIDER FACE dataset including eye centers, nose tip, and mouth corners. Besides RetinaFace used multi-task learning in the training with the facial landmark regression loss:

$$\begin{aligned} L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) \\ + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) + \lambda_3 p_i^* L_{pixel} \end{aligned} \tag{3.5}$$

Face box regression loss $L_{box}(t_i, t_i^*)$, where t_i and t_i^* represent the coordinates of the predicted box and ground-truth box associated with the positive anchor. Facial landmark regression loss $L_{pts}(l_i, l_i^*)$, where l_i and l_i^* represent the predicted five facial landmarks and ground-truth associated with the positive anchor. Both face box regression and the five facial landmark regression employ the target normalisation based on the anchor centre using smooth-L1 function defined in [35]. The facial landmarks are important for the division of the face and detection of our interested regions as illustrated in Figure 3.11. The framework applied single-stage methods that is much faster compared to the two-stage methods, and used context modules to enhance the model's contextual reasoning ability. In addition, RetinaFace also employed light-weight MobileNet as the backbone networks which could run considerable real-time speed of 20 fps at multi-thread CPU for 1920×1080 images, and 60 fps at single-thread CPU for 640×480 images. As a result, it is capable of being used in real time detection with the image size in our scenario and has much potential to deploy the whole framework on the mobile devices like smart phones.

3.2.2 RoIs detection

After detecting the face and facial landmarks, three different ROIs are identified based on the previously detected facial landmarks and extracted from RGB videos frame by frame for

vital signs' estimation as illustrated in Figure 3.12.

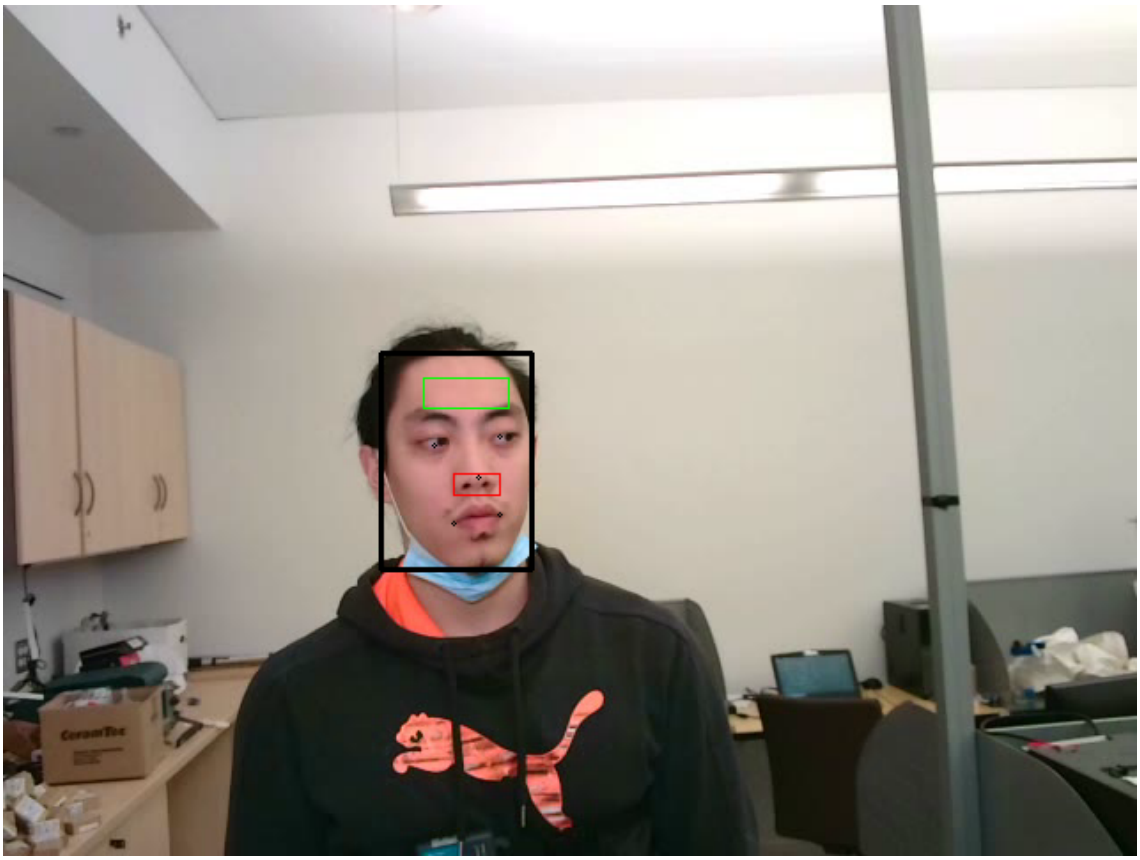


Figure 3.11: Facial landmarks and partitions detected at 1.2m.

Most of the non-contact thermometers measure body temperature on forehead, as our forehead emits heat in the form of infrared radiation. In this study, a single point on forehead is identified as the interested point (T_x, T_y) for body temperature estimation. After detecting the bounding box of the face and locating the positions of the eyes, the height of the forehead can be divided as $\frac{1}{3}$ of the height between the upper bound of the face box and y coordinates of the eyes. We set T_x as the average of x coordinates of the left eye and right eye. And T_y is set as half the height of forehead.

The forehead area is also selected as the ROI for remote heart rate estimation. Blood circulates from the heart to the head through the carotid arteries during each cardiac cycle. This periodic inflow of blood effects both the optical properties of facial skin and the mechanical movement of the head which enables remote HR measurement on forehead. In this study, the forehead ROI used for HR measurement is based on the center-point defined above as (T_x, T_y). The width of the forehead area is set as 80 percent of the width of the bounding box of the detected face. Similarly, the height of forehead area for HR detection is set as $\frac{2}{5}$ of the forehead height defined previously because the area cannot include the eyebrows and

hair.

In this study, nostril is determined as the ROI for respiration rate estimation. Warm air from inside the lungs is released through respiratory system and it increases the temperature in the nasal region during exhalation, whereas cool air from the external environment is breathed in and it lowers the temperature in the nasal region during inhalation. Therefore, the respiration waveform can be obtained by using an infrared thermal camera to measure such nasal-region temperature changes associated with respiration. As a result, the nostrils area are chosen as ROI for respiratory signal detection. The nose tip landmark is chosen as the center-point of nostril area. So we assign the 90 percent of the distance between two mouth corners to the width of nostril area, and 60 percent of the distance between y coordinate of nose tip to the average y coordinates of mouth corners to the height of the nostril area. Although the human facial temperature distribution is internally controlled by blood vessel regulation, ambient temperature is an external factor that affects the thermal pattern in detection. We do not need to figure out the precise temperature values within the nostril ROI. Alternatively we focus more on the temperature variation that related to the each breath.

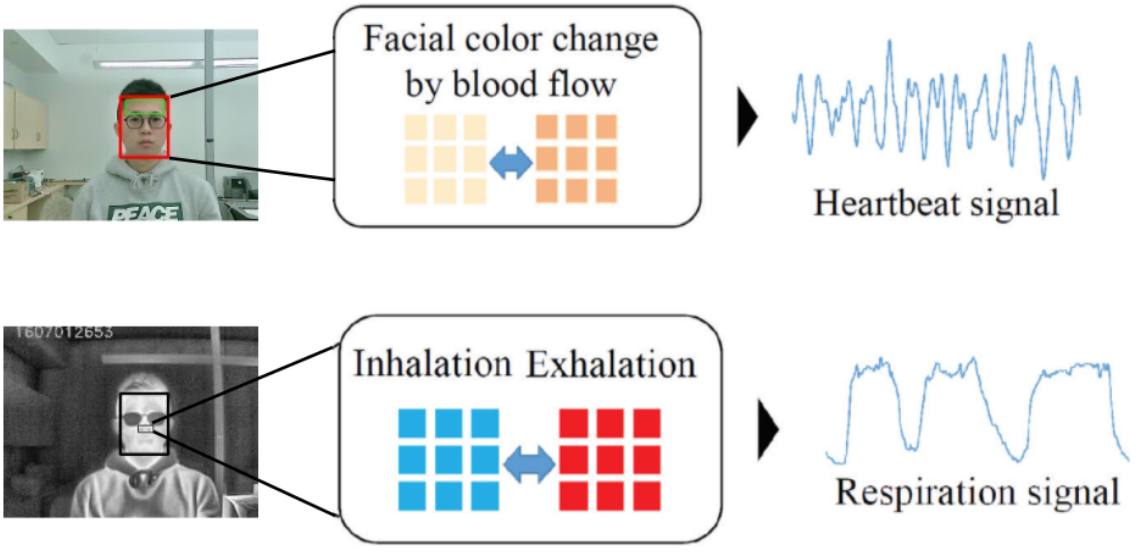


Figure 3.12: ROI in RGB frame and thermal frame.

3.2.3 Frame Registration

According to the specs of the RGB camera and thermal camera, we match the RGB frame to the thermal frame because the thermal camera has lower resolution compared to the RGB camera. In our system, two cameras are placed side by side in the same plane. Consequently

the alignment of frames from two cameras is simplified to an affine transformation problem. Affine transformation is a linear mapping method that preserves points, straight lines, and planes. Sets of parallel lines remain parallel after an affine transformation. We only consider translation and scaling in our two-camera system.

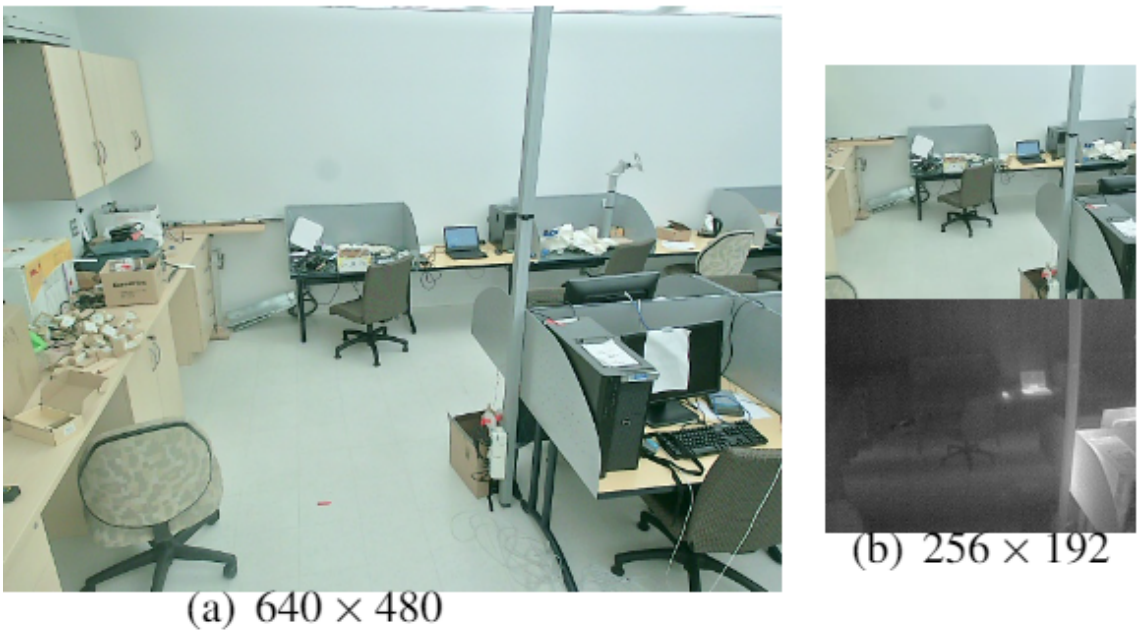


Figure 3.13: Registered synchronous frames from two cameras.

Transformation matrix T is defined as

$$T = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ t_x & t_y & 1 \end{bmatrix} \tag{3.6}$$

where t_x specifies the displacement along the x axis, t_y specifies the displacement along the y axis, s_x specifies the scale factor along the x axis, and s_y specifies the scale factor along the y axis.

Translation transformation is caused by the difference between the position of the two parallel cameras. Since the imaging planes of RGB camera and thermal camera are within the same plane, the displacements are along the x axis and y axis. Scale transformation is caused by the focal length. The lens of RGB camera has shorter focal length compared to the lens of thermal camera, and so RGB camera captures much wider field of view while smaller picture. The registered synchronous frame from RGB camera and thermal camera is illustrated in Figure 3.13.

3.2.4 Vital Signs Estimation

Body Temperature Measurement

Thermal camera can measure subject's surface skin temperature without being physically close to the person being evaluated. Forehead center point (T_x, T_y) in the thermal image has the corresponding temperature of which on the surface skin. Different from the pixel value of (T_x, T_y) in the thermal image that has been transformed and using the palette to be visible, the temperature values of each pixel point within the image are original and raw data calculated from the materials emmissivity table. Besides, all of 256×192 temperature values in one frame are stored in a list following the sequence of pixel array. Coordinate of forehead center point is functioning as key of the dictionary to extract forehead temperature. However, the thermal system measures surface skin temperature, which is usually lower than a temperature measured orally, and the experiment also shows that the environment could have impact on the measurement like the detection range and environmental temperature. So there are some difference between measured value and real body temperature that will be discussed in Chapter 4.

Respiration Rate Estimation

After locating the nasal region from thermal video, the breath-like signal is extracted by averaging the values of all pixels within $ROI_{nostril}$ frame by frame. However, the respiratory features extracted directly from the original thermal frames are weak when the respiration-induced thermal variance is weak, e.g. during shallow breathing. Besides, the low spatial resolution of the thermal imaging also leads to the weak signal. Consequently, we applied histogram equalization to solve the weak signal problem. Histogram equalization is an image processing technique that adjusts the contrast of an image by using its own histogram. Here we use the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm [93] to increase the contrast by spreading out the most frequent intensity values. As a result, we enhance the variation of RoI in thermal frames as illustrated in Figure 3.14 before we attempt to extract the respiratory feature.

Then, a 2nd order Butterworth bandpass filter with a lower cutoff frequency of 0.15 Hz and a higher cutoff frequency of 0.5 Hz (corresponding to 9-30 bpm) was applied for noise removal. The power spectrum of the extracted breath-like signal was obtained by applying Fast Fourier Transform (FFT) and the respiratory frequency is designated as the frequency that corresponds to the highest power of the spectrum. Therefore, the respiratory rate is obtained by multiplying respiratory frequency with 60.

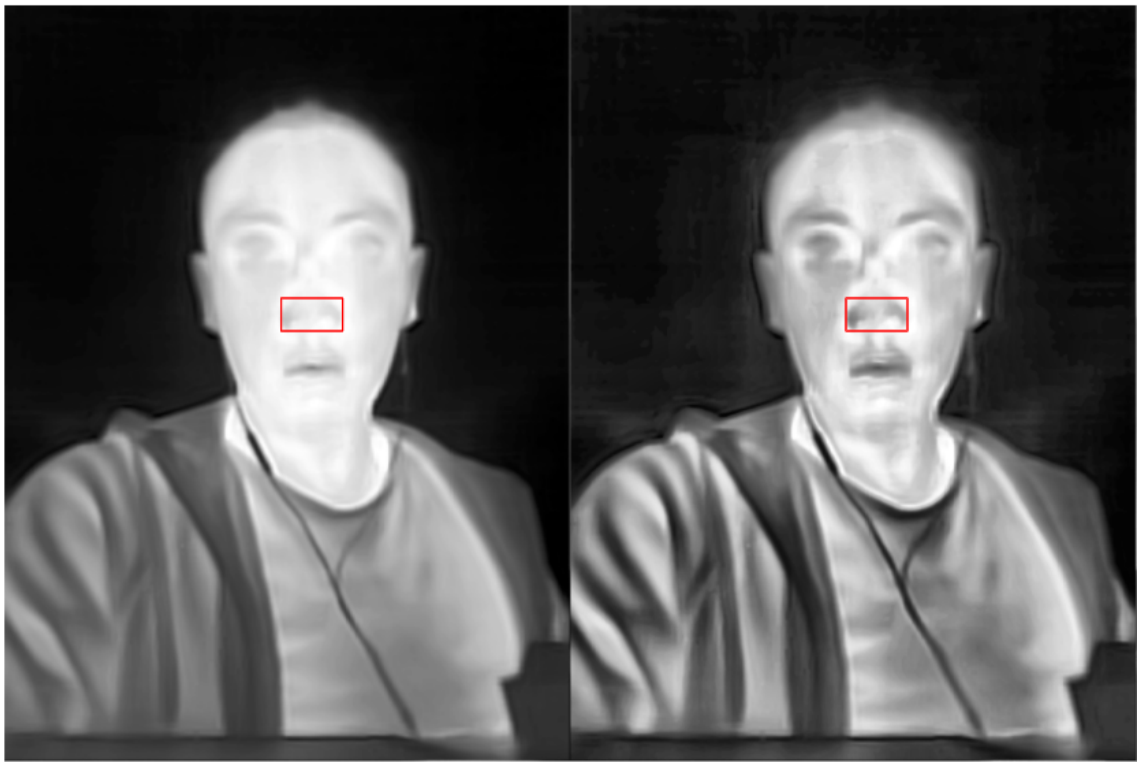


Figure 3.14: CLAHE is applied to the original thermal image (left) and enhance the contrast of detected RoI nostril area in the image (right).

Heart Rate Estimation

In this study, the independent component analysis (ICA) is applied for remote heart rate estimation which was originally employed in [78]. For $ROI_{forehead}$ extracted at time point t, the observed signals from red, green and blue channels are denoted as $x_1(t)$, $x_2(t)$ and $x_3(t)$ respectively, which are the averages of all pixels in the ROI region.Then, the raw RGB traces are normalized as follows:

$$x_i^{'}(t) = \frac{x_i(t) - \mu_i}{\sigma_i} \tag{3.7}$$

Where, μ_i and σ_i for $i = 1, 2, 3$ are the mean and standard deviation of $x_i(t)$ respectively.

The normalized raw traces are then decomposed into three independent source signals using ICA. The joint approximate diagonalization of eigenmatrices (JADE) algorithm developed by Cardoso [94] is applied, and the second component which always contains the strongest plethysmographic signal is selected as the desired source signal for heart rate estimation. One 3rd order Butterworth bandpass filter with a lower cutoff frequency of 0.8 Hz

and a higher cutoff frequency of 2 Hz (corresponding to 48-120 bpm) is applied for noise removal. Finally, the power spectrum of the selected source signal is obtained by applying fast Fourier transform (FFT) and the pulse frequency is designated as the frequency that corresponds to the highest power of the spectrum. The heart rate is obtained by multiplying pulse frequency with 60.

Chapter 4

Experiment and Results

This chapter documents and discuses the results of:

- Details of the depth image dataset
- Results of body tracking based on our own dataset
- Results of the face detector model's performance tested with the data collected in our lab
- Results for vital signs estimation and comparison with the reference dataset

4.1 Experiment Details

Both the depth camera based research and thermal camera based research have applied for the ethics approval of University of Ottawa. Besides the University of Ottawa Research Ethics Board has given certificate of ethics approval for both the research. Ethics approval is valid for the period indicated in the application form and is subject to the conditions listed in the section entitled "Special Conditions or Comments".

4.1.1 Depth Camera based Research

The purpose of this research is to develop a system that is capable of obtaining information pertaining to a human subject's vitals and state of activity using non-contact sensors in real time. The vital signs that this system attempt to monitor are the subject's breathing. The system was installed in the ward at hospital just like what is illustrated in Figure 3.1. For each collection, the system works for 48 hours to collect depth data, radar data and store them in the minicomputer under the requirements of the related ethical approval. We recruited 6

volunteer patients at the University of Ottawa Heart Institute, and the whole collections were made between July to December in 2019.

The system is developed specifically for use in nursing homes and hospitals. Vital signs are needed so that alarms can be sent to care takers in the event of a drastic change in health, such as if the subject stops breathing or some other unhealthy breath patterns like sleep apenea. In addition, information regarding the activity or mobility and posture of the subject is of interest to physicians so they can know how active their patients are on a daily basis. Two sensors are studied: an ultra wide-band (UWB) radar and a 3-D depth sensor. The purpose of using multiple sensors is to be able to combine the individual advantages together for solving the problem of contactless respiratory signal estimation. Software has been developed so that all sensors can be connected to one computer and they can begin recording data at the same instance in time. The research can ensure that the data from each modality is synchronized in time so that they can be compared properly during processing. Data fusion using the two sensors is also explored as a potential solution.

During analysis, only information regarding the test procedures are used. Personal information regarding the subject are not be needed or stored. This research will contribute to health care in Canada with respect to assurance of well being in nursing homes. The goal of this project is to develop a system that when applied in nursing homes can monitor vitals at a distance as well as keep track of the level of activity of occupants on a daily basis. This system is designed to preserve the privacy of individuals being monitored, be non-obtrusive, contact free and relatively affordable. Vital sign information from this system would allow for quick response to medical emergencies without requiring an alert issued by the person experiencing the emergency. This is especially important for emergencies that may leave the patient incapacitated and unable to issue and alarm, or for patients who may be reserved to issue an alarm when they need help. It will also help physicians gain information about how active their patients are on a daily basis so they can make informed recommendations for healthier lifestyles.

4.1.2 Thermal Camera based Research

The goal of this project is to implement a stand-alone system with a thermal imaging camera that would allow for real-time processing and estimating the following vital signs of people: temperature, heart rate and breathing rate. We will consider two different systems, one fixed at the entrance doors that determines temperature range of incoming people and another one fixed in the rooms of subjects that points towards the areas where they spend most of their time.

In the first subproject, the participants are expected to come in to the area of interest. The thermal camera is fixed at the entrance doors and it determines the temperature scope of incoming people. The task is tracking multiple people and their facial features using combination of RGB and thermal cameras and detecting fever of multiple people at once. The subject stands before our system for 15 to 20 seconds, and our system estimates the heart rate, respiratory rate of the subject by processing the data from cameras. Thermometer is also used to collect the temperature of the human body as baseline data. All the data collection take place in our laboratory at the University of Ottawa.

The goal of the second subproject is to monitor the subjects for a continuous longer period of time. Thermal camera is fixed in the rooms of subjects pointing towards the areas where they spend most of their time. The system allows for continuous monitoring of their breathing rate, heart rate and temperature. The system is stationary and will focus on monitoring only a single subject. Data are be recorded for 10 subjects in a lab located in SITE 5130. Four sensors are used during the recording procedure: thermal camera, RGB camera, thermometer and respiratory impedance plethysmography (RIP) belt. The RIP belt is be used for collecting baseline vital sign information so that the algorithms developed for processing the data from the non-contact sensors can be evaluated.

All the thermal data are collected in the format of series of frames while we process the offline data frame by frame. Temperature range could be directly extracted from the thermal image, and once the RoI has been selected, the temperature variation inside the selected area could be extracted. Temperature data collected from the thermometer are written in csv file and used as baseline standard to testify the thermal camera function. In the HR and RR estimation portion of analysis, the outputs of the algorithms are compared with the baseline data to estimate accuracy and confidence intervals. Baseline data are determined from the RIP by computing the average rate in each sample using simple spectrum peak detection. During the analysis, only information regarding the test procedures are used. Personal information regarding the subject are not be needed or stored.

The test protocol was developed to ensure proper controls and avoid any biasing for classification of activities and postures. The test protocol requires the subject to perform varying levels of activity while sitting, standing and lying down, as well as simulating stop breathing events by holding their breath for as long as possible while in each of the three postures. Tests are performed in different locations throughout the room so that the algorithms can be developed invariant to location and relative angle. Uncontrolled, or random tests are also be performed so that the developed algorithms can be tested on data that will closely represent real life data. In addition to the regular test protocol, 10 minutes of empty room data are recorded before each day of testing for use in classification of room occupancy.

4.2 Depth Image Dataset

All the raw depth images are converted to grayscale images visible to human eyes using the methods introduced in Section 3.1.2. Then we use VIA annotation tool to label the depth images just like the approach discussed in Section 3.1.3. Some examples of the depth dataset are as illustrated as Figure 4.1.

We annotated all the images by pointing out the contour of the human body that includes head, neck, shoulders, and the chest region. The exported file is in JSON format, and is consisted of the name of image, the mark type, and all the mask points' x and y coordinates. Currently, we only have 1000 labeled depth images for training and 200 labeled ones for testing so the scalability of our model remains to be further tested. Our training dataset includes as much as possible different postures of the subjects from what we collected from the hospital. However, there are still a lot postures and scenes we have not contained in both the training and test dataset.

We also used COCO Annotator [87] to build our human pose dataset. This dataset is just like the previously one annotated by VIA tool. The only difference is that the COCO Annotator also labels the keypoints on the depth images other than just mask segmentation as illustrated in Figure 4.2. However, this 4-keypoint human body dataset is relatively smaller because of the time-consuming annotation work, the training dataset has 100 images and testing dataset has 20 images.

In COCO format, each object instance annotation contains a series of fields, including the category id and segmentation mask of the object as illustrated in Figure 4.3. The segmentation format depends on whether the instance represents a single object (iscrowd=0 in which case polygons are used) or a collection of objects (iscrowd=1 in which case RLE is used). Note that a single object (iscrowd=0) may require multiple polygons, for example if occluded. Crowd annotations (iscrowd=1) are used to label large groups of objects (e.g. a crowd of people). In addition, an enclosing bounding box is provided for each object (box coordinates are measured from the top left image corner and are 0-indexed). Finally, the categories field of the annotation structure stores the mapping of category id to category and supercategory names. A keypoint annotation contains all the data of the object annotation (including id, bbox, etc.) and two additional fields. First, "keypoints" is a length $3k$ array where k is the total number of keypoints defined for the category. Each keypoint has a 0-indexed location x,y and a visibility flag v defined as $v = 0$: not labeled (in which case $x = y = 0$), $v = 1$: labeled but not visible, and $v = 2$: labeled and visible. A keypoint is considered visible if it falls inside the object segment. "num_keypoints" indicates the number of labeled keypoints ($v > 0$) for a given object (many objects, e.g. crowds and small ob-

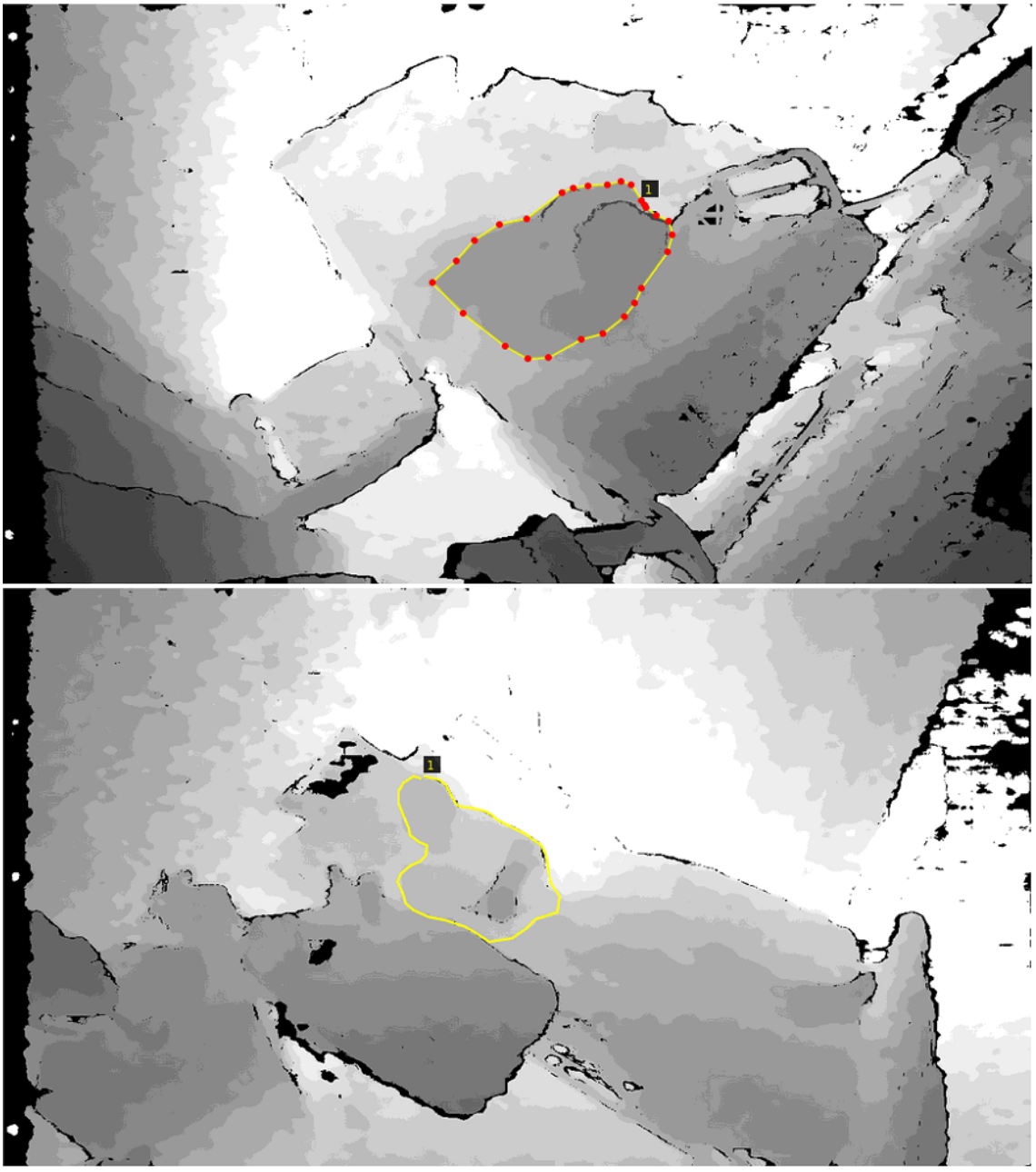


Figure 4.1: Examples from the training dataset.

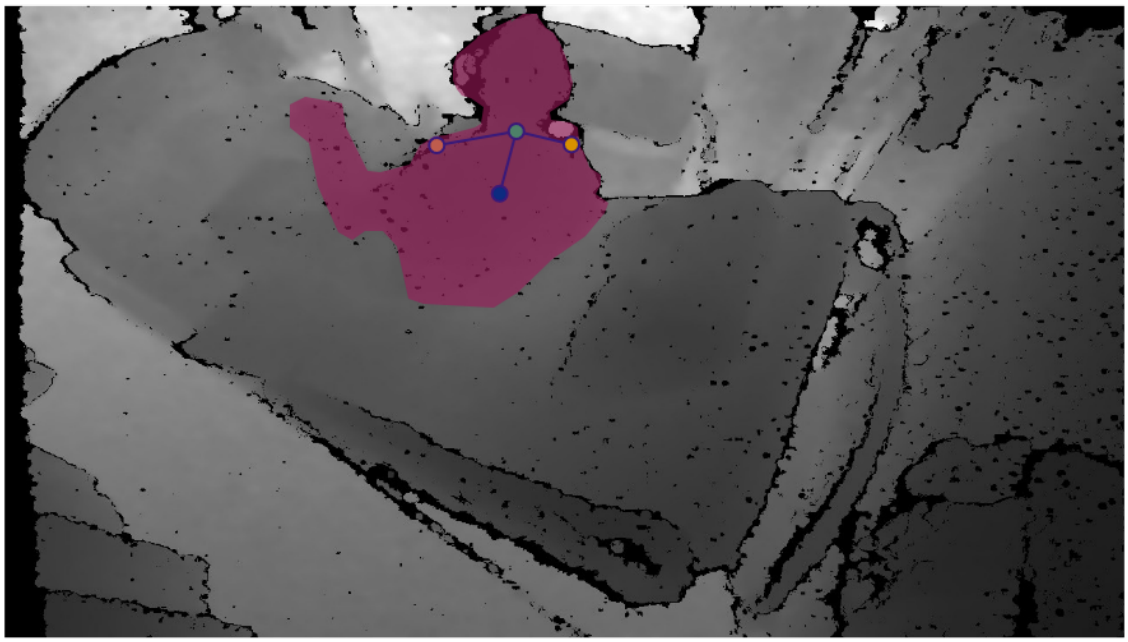


Figure 4.2: Examples from the 4-keypoint dataset.

jects, will have num_keypoints=0). Finally, for each category, the categories struct has two additional fields: ”keypoints,” which is a length k array of keypoint names, and ”skeleton”, which defines connectivity via a list of keypoint edge pairs and is used for visualization as illustrated in Figure 4.3.

```
annotation{
   "id"                : int,
   "image_id"          : int,
   "category_id"       : int,
   "segmentation"      : RLE or [polygon],
   "area"              : float,
   "bbox"              : [x,y,width,height],
   "iscrowd"           : 0 or 1,
}

categories[{
   "id"                : int,
   "name"              : str,
   "supercategory"     : str,
}]
```

```
annotation{
   "keypoints"         : [x1,y1,v1,...],
   "num_keypoints"     : int,
   "[cloned]"          : ...,
}

categories[{
   "keypoints"         : [str],
   "skeleton"          : [edge],
   "[cloned]"          : ...,
}]
```

"[cloned]": denotes fields copied from object detection annotations defined above.

Figure 4.3: COCO JSON file example.

4.3 Results of Body Tracking

4.3.1 Network Customization and Fine-tuning

Mask Detection

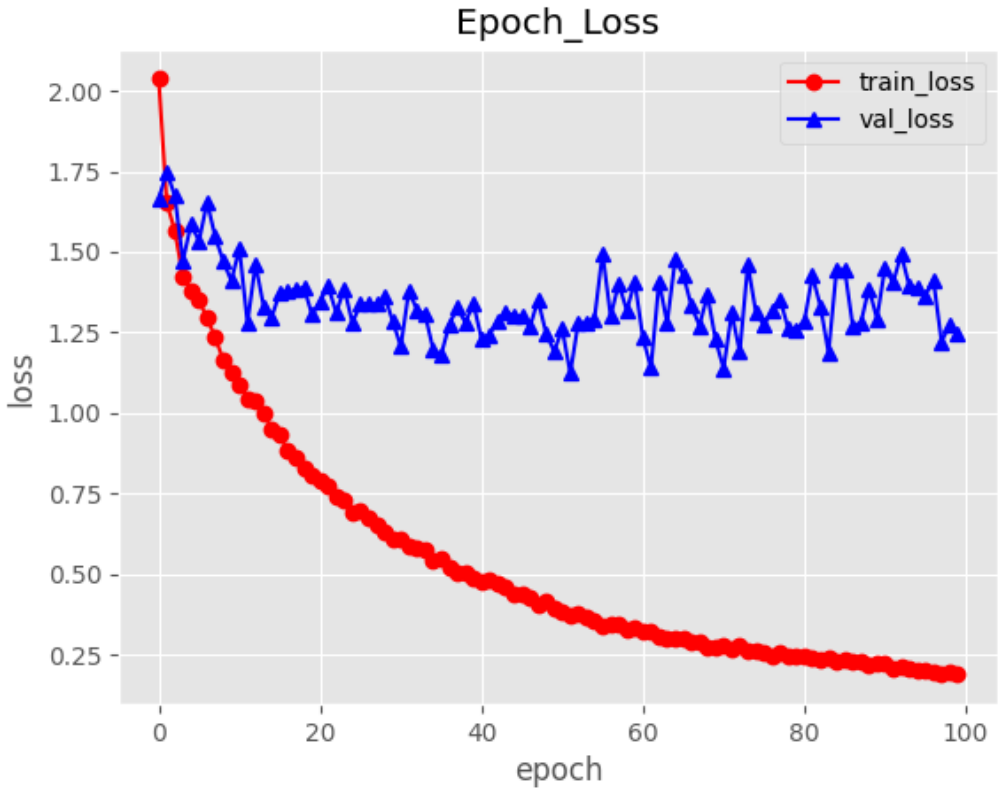


Figure 4.4: Model for mask detection epoch loss result based on 100 epochs training.

We present the results of our model fine-tuned based on the created training set. Mask R-CNN is implemented using Python 3.6 and Tensorflow 1. We trained the human body detection model based on our own labeled dataset mentioned above and tuned it with the weight previously trained on COCO dataset [95]. In detail, since we use a very small dataset compared to the huge COCO dataset and utilise the weight that is previously trained on the COCO dataset, we do not need to train too long starting from scratch. Also, there is no need to train all layers, just the fully-connected layers in the network should do it. Another basic modification is the input of the network because we use the depth images in our dataset. Since all the depth data are grayscale images which only have one color channel and they are totally different from the format of images in the COCO dataset. So we rebuild and fine-tune the one-channel depth images to the three-channel RGB images that have three identical

color channels.

The learning rate is set to 0.001 and the training process is performed on Nvidia Tesla P100 GPU. After 100 epochs training process, the aforementioned loss that is the sum of L_{cls}, L_{box}, and L_{mask} for the masks head on training dataset has dropped from 0.7498 to 0.1850 as illustrated in the Figure 4.4. In addition, the fine-tuned detection model runs at average 1 frame per second on an Nvidia GeForce GTX 1050 Ti. However, currently we only have 1000 labeled depth images for training and 200 labeled ones for validation so the scalability of our model remains to be tested.

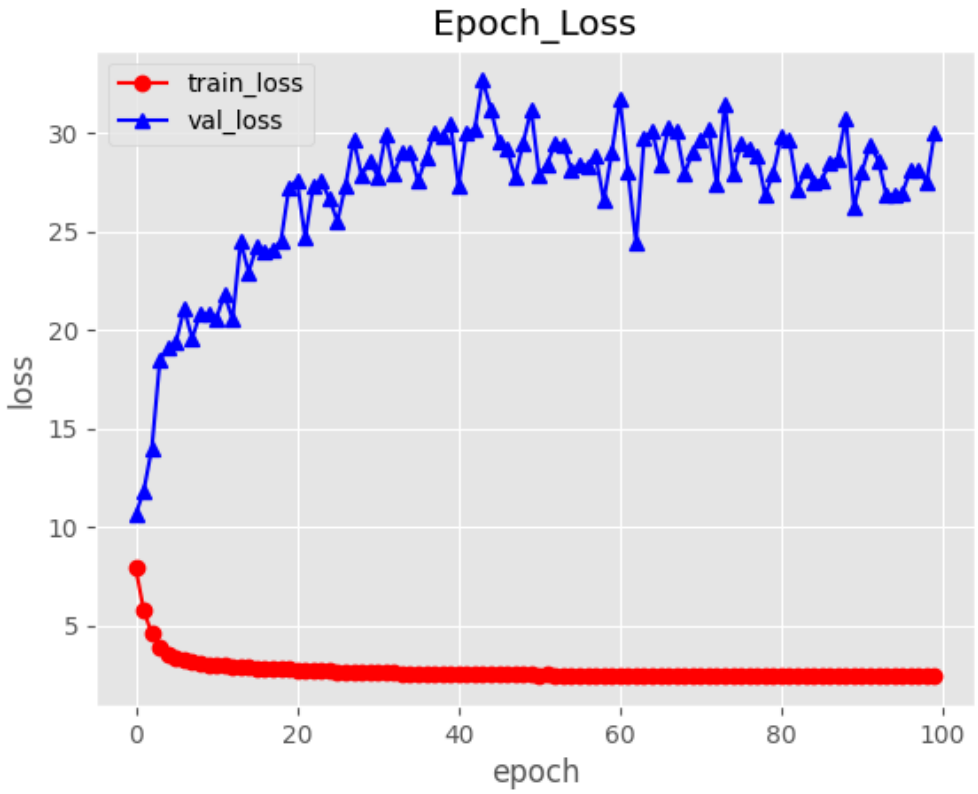


Figure 4.5: Model for keypoints detection epoch loss result based on 100 epochs training.

The reason why we use fine-tuning that is one trick of transfer learning is because it has the benefit of decreasing the training time for our neural network model and can result in lower generalization error. Besides, transfer learning can save the amount of labeled data for the training dataset when the first related problem has a lot more labeled data than the problem of our interest [96, 97].

Keypoints Detection

We also trained a Mask R-CNN based model to detect the 4 keypoints of the human body. We model a keypoint's location as a one-hot mask, and adopt Mask R-CNN to predict 4 mask, one for each of 4 keypoint types. These 4 keypoints include neck keypoint, left shoulder keypoint, right shoulder keypoint, and chest keypoint. We make minor modifications to the segmentation system when adapting it for keypoints. For each of the 4 keypoints of an instance, the training target is a one-hot binary mask where only a single pixel is labeled as foreground. During training, for each visible ground-truth keypoint, we minimize the cross-entropy loss over a softmax output (which encourages a single point to be detected). We note that as in instance segmentation, the 4 keypoints are still treated independently. Models are trained on all COCO format images that contain annotated keypoints.

However, as illustrated in Figure 4.5, the val_loss increases with the training which shows severe overfitting. Overfitting happens when a model begins to focus on the noise in the training data set and extracts features based on it. This helps the model to improve its performance on the training set but hurts its ability to generalize so the accuracy on the validation set decreases and val_loss keeps growing. Overfitting in our scenario is probably caused by the small size of training dataset.

4.3.2 Object Detection Analysis

As illustrated in the Figure 4.7, the output of body tracking algorithm includes the colored mask area and two important values: the "person" class probability and the center point distance. The example figures show both the original input depth grayscale frames and the results output from our body tracking method. The mask area is predicted by Mask R-CNN based model and the "person" class probability number indicates the probability of the class where the object belongs to. The estimated distance is extracted from the raw depth frame by using the pixel's coordinates that are located at the geometric center point of the mask area.

Mask Detection

After 100 epochs training, we obtain a Mask R-CNN based model that represents the certain number of neurons and weights of the network. We use this model to predict the ROI of each image. When the tested image is put into the neural network, the pre-trained model inspects and infers the input and then gives the prediction based on its own weights. Since there are more than one prediction with different shape and class probability generated by the inference model, just as illustrated in Figure 4.8, we set the threshold to filter out the

Table 4.1: Parameters and APs of the model.

	Parameters	**FLOPs**	**AP**	$AP^{IoU=0.50}$	$AP^{IoU=0.75}$
Mask RCNN	44,662,942	205,931,594,377	0.413	0.707	0.508

The table lists main features and APs of the Mask RCNN based methods for object detection trained on our own depth image dataset.

worse predictions. We threshold the confidence probability of the detected object, and put the bar to 95% which means that we only take the predictions with class probability more than that number. Our Mask R-CNN model has 44,662,942 parameters and 205,931,594,377 floating point operations per second (FLOPs). The mAP computed based on our testing set according to the evaluation metrics in Section 2.2.5 is 0.707 when we set the IoU threshold to 0.5 as listed in the Table 4.1. The performance of our model could also be evaluated by the precision-recall curve as illustrated in Figure 4.6. Besides, the inference result generated by our model as shown in Figure 4.7 gives out the red-colored mask region with a white-colored center point, as well as the ”person” class probability and distance extracted from the raw depth data.

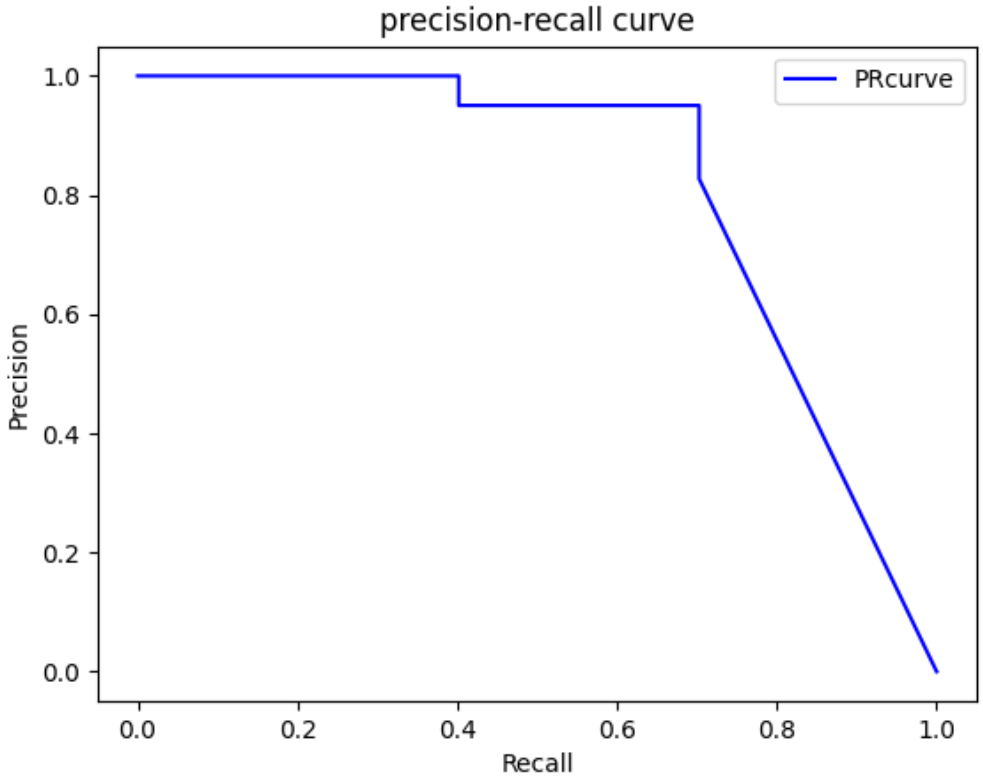


Figure 4.6: Precision-recall curve of our model.

However, the model is still not robust enough and sometimes will mistake the interested region and detect wrong objects. As shown in Figure 4.8, our model detects the object and gives prediction of different class probabilities. The three different colored regions such as blue, red, and green separately represent three different predictions with different probabilities. Besides, many of these predictions are incorrect estimations, and even with the high probability like the green one that is not even connected. Although higher the confidence probability, as the example illustrates, the closer the predicted mask area is to our ROI of the object detection task in the whole image, it is hard to pick up the right prediction by looking at the probabilities.

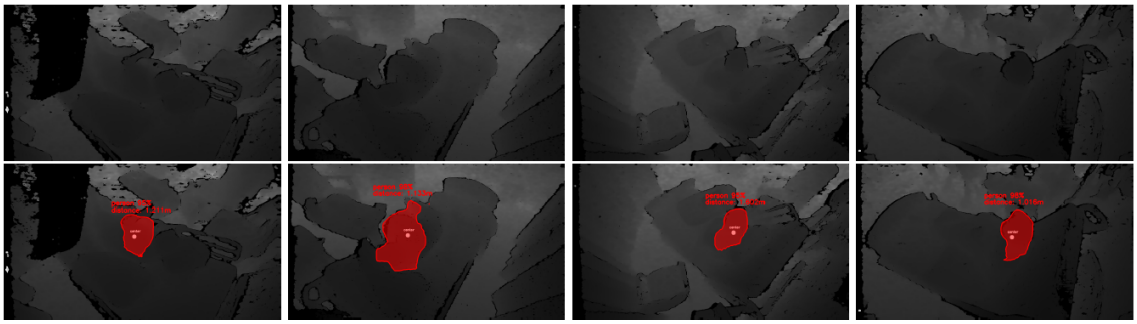


Figure 4.7: Body tracking result example.

We do not have enough resources to validate the accuracy of the distance measurement except the distance between the depth camera and the bed measured before the whole experiment. By taking into account time it takes to process the depth frames, it is feasible to realize the human tracking in real time based on depth camera and provide the radar system with a reference range.

Keypoint Detection

As illustrated in the Figure 4.9, our model has detected two objects, one in red and another in blue within the whole scene; however, there is only one subject (blue one) in the scene. The red colored mask, as well as the 4 keypoints, are false predictions given by the model. Similarly, the blue colored mask also includes a false area that does not belong to the subject even if the 4 keypoints predictions are accurate. In conclusion, the results generated by our keypoints detection model are not good as well due to the small size of training dataset. The 100 images training dataset generates an overfitted model that takes what is not the right mask as the prediction. These incorrect mask predictions definitely give out the incorrect predictions of keypoints because of the unsatisfactory precondition. As a result, this keypoint detection model is currently not applicable for the task of detecting keypoints based on depth frames at all.

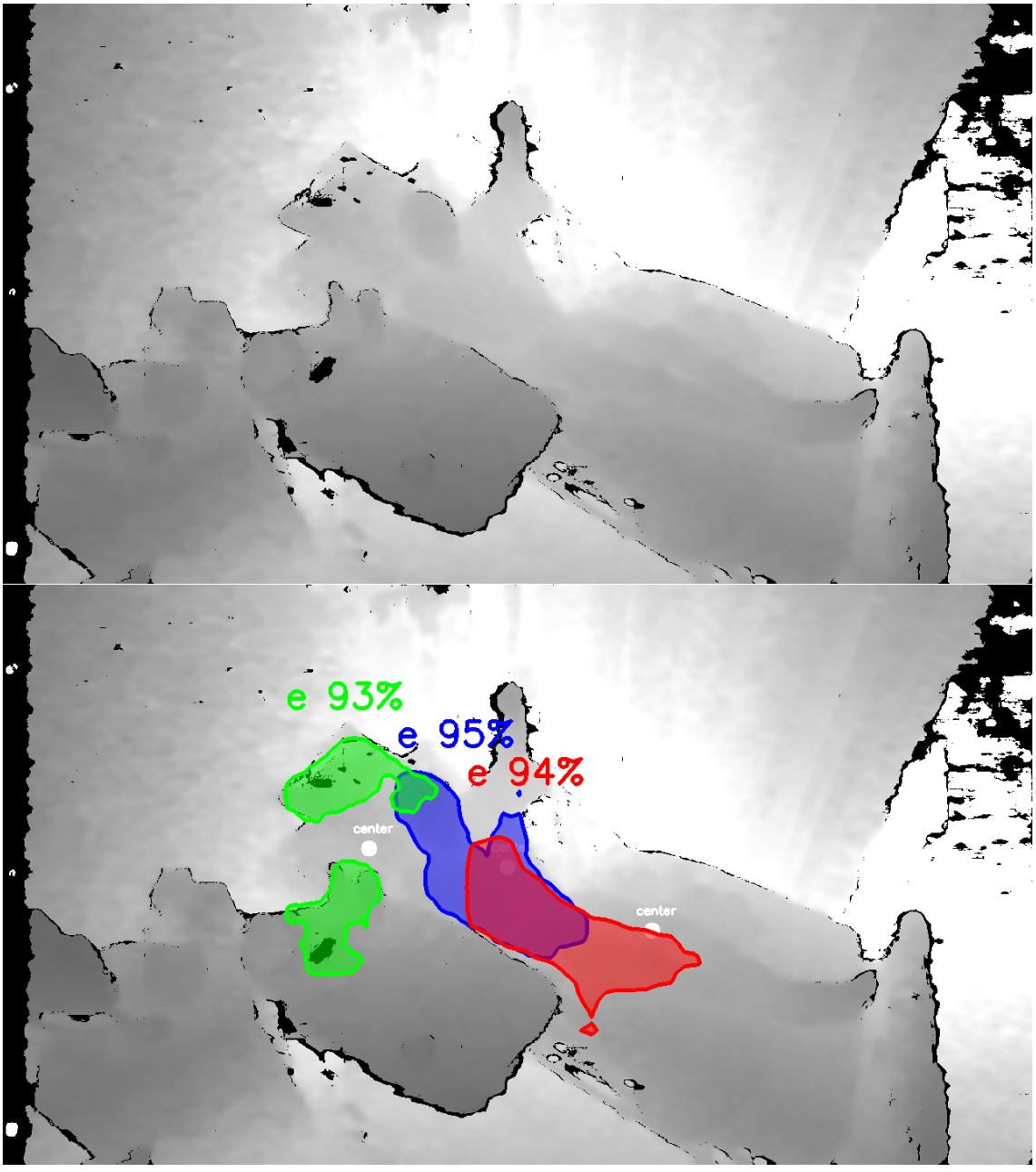


Figure 4.8: Unsatisfactory predictions generated by our body tracking model.

There exists a COCO dataset with labeled human keypoints. However, the keypoints are based on the RGB images rather than our depth images and each of the person is labeled with 17 keypoints instead of 4 keypoints like in our dataset. It is hard to perform transfer learning from the pre-trained keypoints model. The reason why we want to detect the keypoints like neck and chest is that the geometric center point estimated based on the mask mentioned in Section 3.1.4 is still not precise sometimes. The region confined by shoulders, neck and chest is exactly the area where respiratory signal is most obvious to be observed by the radar. Consequently we want to narrow down the interested region much more to just a few pixel area. In order to implement the accurate keypoint detection, we have to increase the size of training dataset and also try to include different scenarios in our dataset which will be further discussed in Chapter 5.

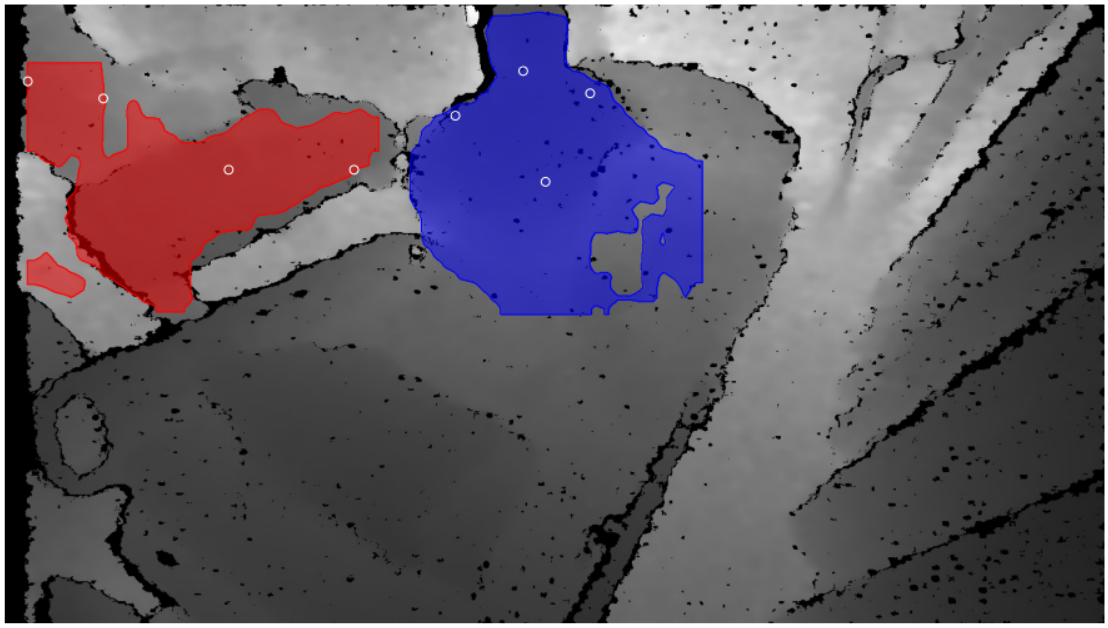


Figure 4.9: Keypoints detection result example.

Existed Limitations and The Reason

Figure 4.7, Figure 4.8, and Figure 4.9 show the detection results of Mask R-CNN based model training on our own dataset. We aim at designing a system that could detect our targeted subjects in the depth frames, and estimate the distance between the subject's chest and camera. The difficulty is mainly in detecting object in depth images since there is a lack of related datasets composed of well-labeled depth images. In addition, several possible reasons that limit the performance of our model are described next.

First of all, the quality of depth images remains to be improved. As shown in the figures above such as Figure 4.9, there is a lot noise in the depth image. Since depth camera takes the reflected infrared light as the pixels' value and environmental light interference always has impact on the reception, and so there are many black dots in the image. Some texture of the cloth will even absorb the infrared light emitted by the depth camera, which is also bad for depth camera's reception and makes the noise in the image. In addition, another reason for causing the bad image quality is the lossy compressing method that we applied to store the images. We convert depth images into 2 dimensional JPG format in order to reduce the hard drive space occupied by images, which however causes some loss to the original data.

Another big problem is the size of training dataset and validation dataset. Since the whole annotation work is done by myself, it consumes a lot of time to label thousands of the images. Even though Mask R-CNN can be applied to small-size dataset because there are existing weights pre-trained on the huge COCO dataset and we can perform transfer learning, our task is a bit different and our dataset is really small, especially the one labeled with keypoints. We cannot create a relatively larger dataset for training, and we cannot make sure the distribution of the data is fair and reasonable. Our dataset is definitely biased because we cannot include all the possible postures and scenarios in our dataset either.

In summary, there are many operations could be made to improve our body tracking methods which will be discussed in the Chapter 5 future work in details.

4.3.3 Respiratory Signal Extraction

We use the estimated centroid point distance from the body tracking framework to look into the corresponding bin of the radar data. We extract and process the radar signal from the estimated range bin and the radar signals from the range bins in its proximity. Linear filter like moving average filter is applied to draw the signal extracted from radar data, as illustrated in Figure 4.10. This work was done by another graduate students in the group. Radar breathing signal estimation relies on accurate detection and locating of the person which is done using algorithms presented in Section 3.1.4. We only show here the results of breathing range estimation from the radar from the correctly determined range bin collected while the subject was lying stationary.

However, even if we can estimate the right distance between the subject and the device and then detect and extract the signal, we cannot be certain that the collected signal is exactly the respiratory signal. Because we also need to qualify the signal weather is of satisfactory quality so that the breathing rate can be extracted. For example, Figure 4.11 illustrates the condition where the non-breath signal does not have the curve shape like the common respi-

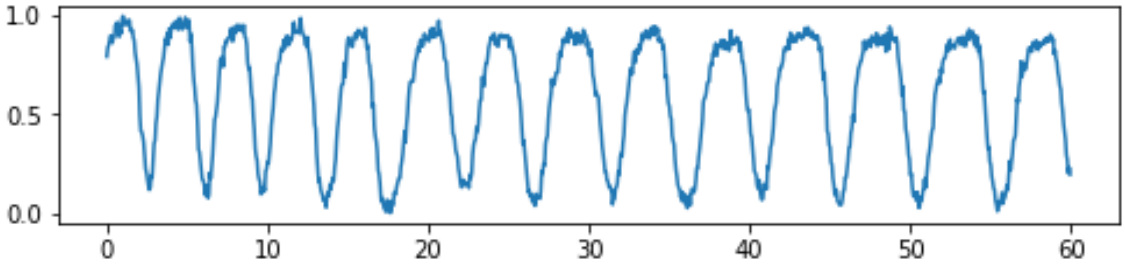


Figure 4.10: Breath-like signal extracted from radar data.

ratory signal. Besides, some of the signals have number of peaks that does not correspond to the common respiratory signal. In addition, to perform detection of the desired signal, we also need to propose some methods to classify signal as the respiratory signal or not which will be discussed in the future work.

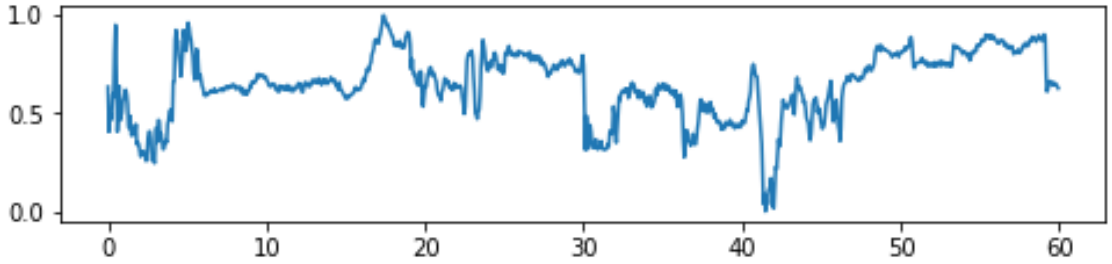


Figure 4.11: Non breath-like signal extracted from radar data.

The problem here is that we do not have the breathing belt data as the reference to compare because we did not have the approval to collect respiratory signal of the subjects in the hospital by wearing breathing belt. Besides we monitor the subjects in a continued time such as 48 hours, and it is impossible to have the reference signal for such a long time. We can only sample some part of the data and then analysis. As a result, one of the feasible approaches here to judge if our system works or not is designing a classifier to classify the breath-like signal and non breath-like data extracted from the radar, which will be briefly introduced in the future work in next chapter.

4.4 Results of Face Detection

4.4.1 Face Detectors Comparison

We have introduced several face detectors in Section 2.2.4, and we select some of them to compare under the condition of our use case where the person is wearing the mask, and the facial landmarks and ROI needs to be detected. We compare four face detectors including

MTCNN [47], S3FD [53], PyramidBox, and RetinaFace that we used in our method. The result is illustrated as Figure 4.12.

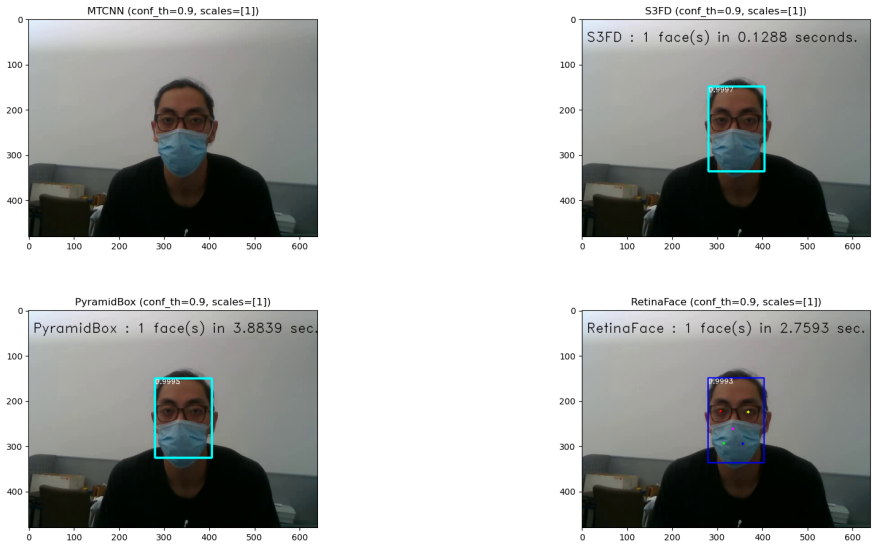


Figure 4.12: Face detection results using different face detectors.

We find that compared to the other three face detectors, MTCNN performs worst in our scenario because it cannot detect the face when the face is occluded with something like mask. So MTCNN cannot be used in the case of people wearing mask even though it is not complicated and is widely used by other researchers in many of the face detection applications now.

Although both of the S3FD and Pyramid Box face detectors can detect the face with the mask on as shown in Figure 4.12 (2,3), they are not our algorithms of choice because of their performance. These two face detectors have problem working on the facial landmarks detection while landmarks detection is needed significantly in our method.

4.4.2 Method Deployment

Our method, which applies RetinaFace detector, is deployed both on a local machine and an embedded device Nvidia Jetson Board for testing.

In terms of embedded device, we applied the whole RetinaFace network framework on the NVIDIA Jetson Tx2 board to do the experiment. Jetson TX2 is the fastest, most power-efficient embedded AI computing device. This 7.5-watt supercomputer on a module brings true AI computing at the edge. It is built around an NVIDIA Pascal™-family GPU and loaded with 8GB of memory and 59.7GB/s of memory bandwidth. It features a variety of

standard hardware interfaces that make it easy to integrate it into a wide range of products and form factors [98]. NVIDIA Jetson Tx2 board's graph processing unit has the 256-core NVIDIA Pascal™ GPU architecture with 256 NVIDIA CUDA cores. The RetinaFace detector takes average 5 seconds to load the whole network model. Besides it takes around 0.7 second to detect each frame, which means that the fps is more than 1. Due to the limited computing power, the embedded device is still reluctant to run this face detector especially in real time.

On the other side, RetinaFace detector has a better performance on local machine. We applied the whole network on a machine with a NVIDIA GeForce RTX 2080 Ti GPU that has total memory of 10.76 Gb. This graph processing unit has 7.5 computing capability and is much powerful than the one in embedded device mentioned above. RetinaFace detector takes average 0.025s to detect one frame and the running speed reaches around 40 fps which is far better than the real time detection. In addition, the detection time will grow with the increase of number of objects in the image. For example, when there is only one subject in the image, the detector takes around 0.02s to detect one frame, sometimes even less than 0.02s.

4.4.3 Face Detection Result

We collected data according to the experiment plan set in the ethics application introduced in Section 4.1.2. The subject is required to stand at different distances in front of our system, so there are different sizes of the face for the same person in our experiment. Though the face detector enables us to detect tiny faces, the moving range of the subject is limited because small faces with very low resolution usually lack a lot details information both in RGB frame and thermal frame. So the moving range of the subject in our experiment is limited to 0.4m to 3m in the direction that is parallel to the view plane of the camera. All the face detection results that we are going to display in this section is based on the RGB frames.

Figure 4.13 illustrates the performance of RetinaFace detecting multiple faces with the mask on. The image is one sample of the data collected from the laboratory where one subject is standing closer to the camera, while another one is a further away from the camera. This experiment imitates the scenario that pedestrians entering the stores or some special rooms. Before they get the access to the room, our system will perform the vital signs estimation automatically so that the potential subjects with symptoms like high body temperature, shortness of breath or difficulty breathing are identified. When the closest subject is standing right before the system for 15 to 20 seconds, other people that are waiting in the line could also be captured by the camera and their vital signs could be processed. Our approach has

the capability of detecting multiple faces in one frame enabling the system to save time on detecting vital signs like body temperature. In addition, the function of multi-face detection provides the potential for detection in crowded and complicated scenes like classroom and shopping mall.

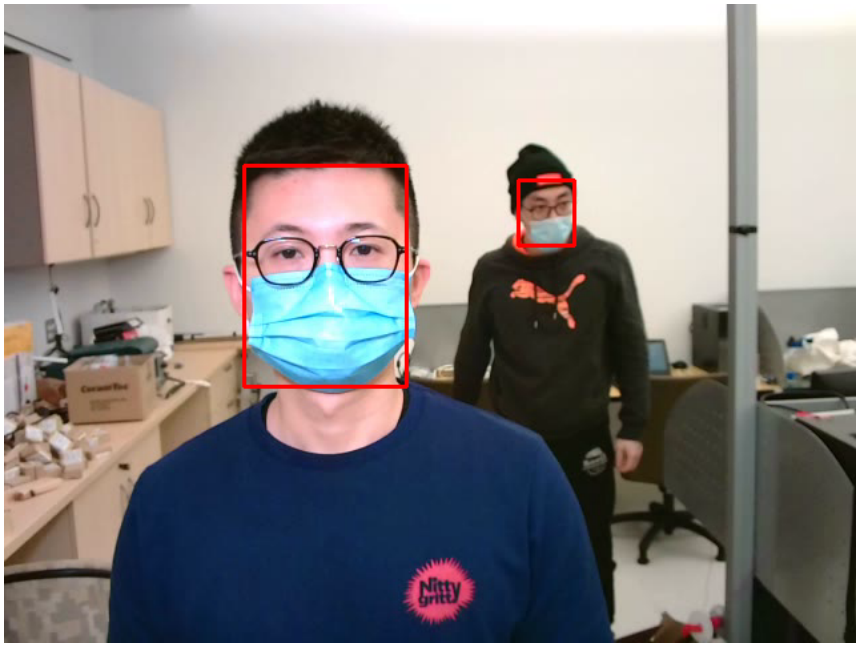


Figure 4.13: Demonstration of the functionality of RetinaFace detector in cases where multiple faces are detected.

Figure 4.14 illustrates the performance of RetinaFace detecting faces with different sizes caused by the different ranges between the subject and the camera. The subject in the first one of the Figure 4.14 is 3.9 meters far away from our system. While in the seconds figure, the subject is standing 1 meter from the camera. RetinaFace detector has the power to detect really tiny faces, however, the tiny faces are useless in our use case as mentioned above. As it is illustrated, the detected face at the distance of 3.9m only contains a small area of the image, and the area of interested region like $ROI_{forehead}$ and $ROI_{nostril}$ contain only several pixels which is not enough to extract signals of satisfactory quality.

The classification scores showing probability of the detection of the human face tested on all four subjects at different positions with or without mask reach 99.95% or even higher using Retinaface detector. Besides, there are no difference between the subject wearing the mask and the one without mask when we evaluate the face classification probabilities. In addition, to evaluate the performance of facial landmarks detection which is the base of our facial ROI detection, we use the normalised mean errors (NME) metric:

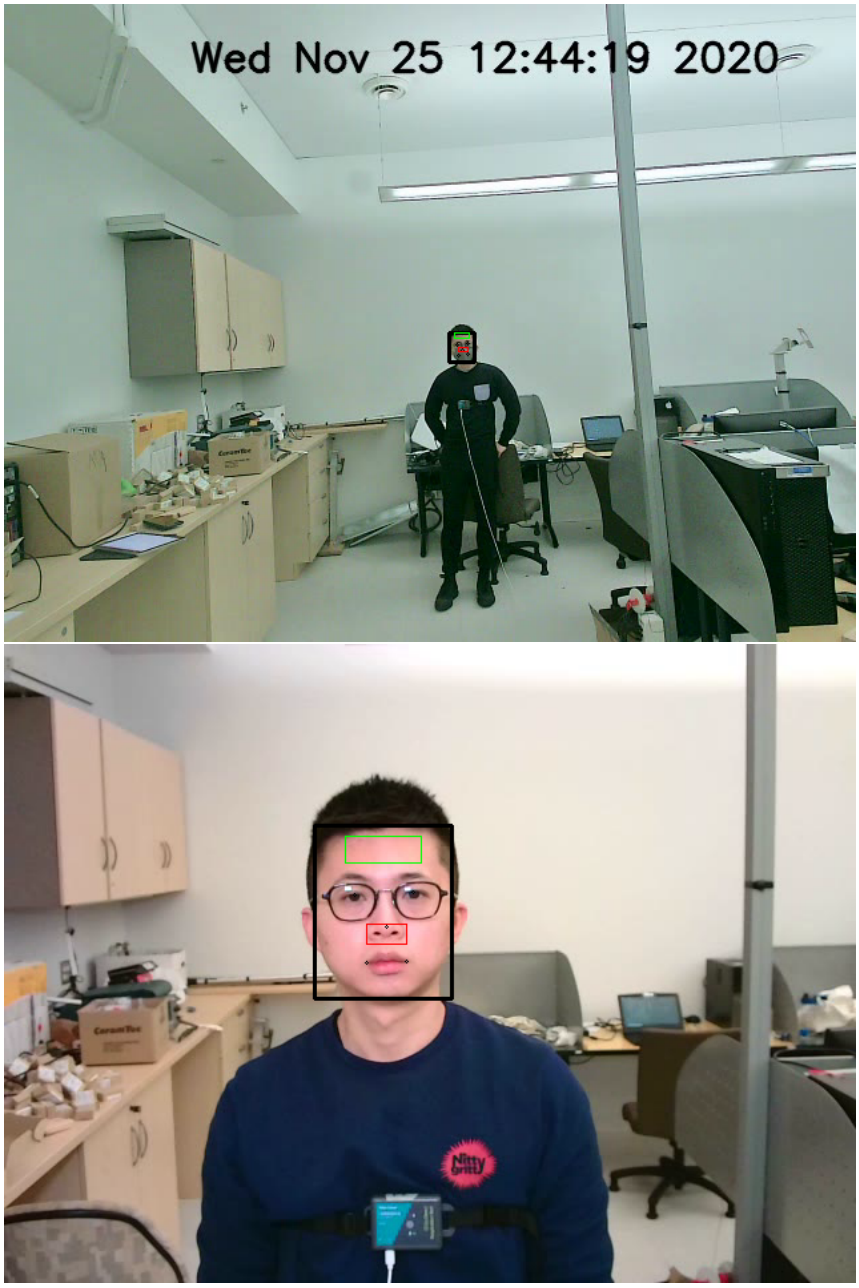


Figure 4.14: Demonstration of the functionality of RetinaFace detector in cases the subject is far from (upper image) and close to (bottom image) the camera.

$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{\Delta x_i^2 + \Delta y_i^2}}{d} \tag{4.1}$$

where N denotes the number of facial landmarks which is five in our situation and d denotes the normalized distance. Δx_i^2 and Δy_i^2 are deviations between the i^{th} predicted landmark and the ground truth in x axis and y axis. We employ the face box size ($\sqrt{W \times H}$) as d in the evaluation. We found that the NME is on average 1.5% when the subject is not wearing the mask, while the NME reaches 2.4% when the subject is wearing the mask. As a result, the face detector is qualified to detect the ROIs that we need in the extraction of vital signs because of the low NME for facial landmarks detection.

4.5 Estimations of Vital Signs

The analysis of face detection by RetinaFace detector shows that tiny faces and faces occluded with masks can be detected by our approach. Besides, our method's execution time is short and therefore the method is suitable for realtime implementation in the future. We are next going to analyze the result of vital signs estimation from the processed frames. We mainly focus on the respiratory signal and heart rate. We acknowledge the contribution in the signal processing work for estimating RR and HR of Shan He.

4.5.1 Respiratory Signal Estimation

Respiratory signal is estimated from processed thermal frames. Our face detection method detects the face and related RoI where the signal is directly extracted. Although face and RoI detection almost have no limitations in our scenario, the vital sign estimation is limited by many factors like light source and distance. Several different positions were used to test the limitations. The performance of the respiration rate estimation was also compared with reference breathing belt data illustrated in Figure 4.15. Go Direct Respiration Belt uses a force sensor and an adjustable nylon strap around the chest to measure respiration effort and respiration rate [99].

For each of the subject we collect 8 different groups (4 distances * 2 mask/no mask) of data including the RGB videos, thermal videos, baseline BR signal and baseline HR data. The subject is required to stand at 4 different distances away from our system in each collection from near to far, and each collection takes 2 to 3 minutes. In detail, the subjects face the cameras directly at the distances of 1 m, 1.5 m, 2 m and 2.5 m. Besides, we collected

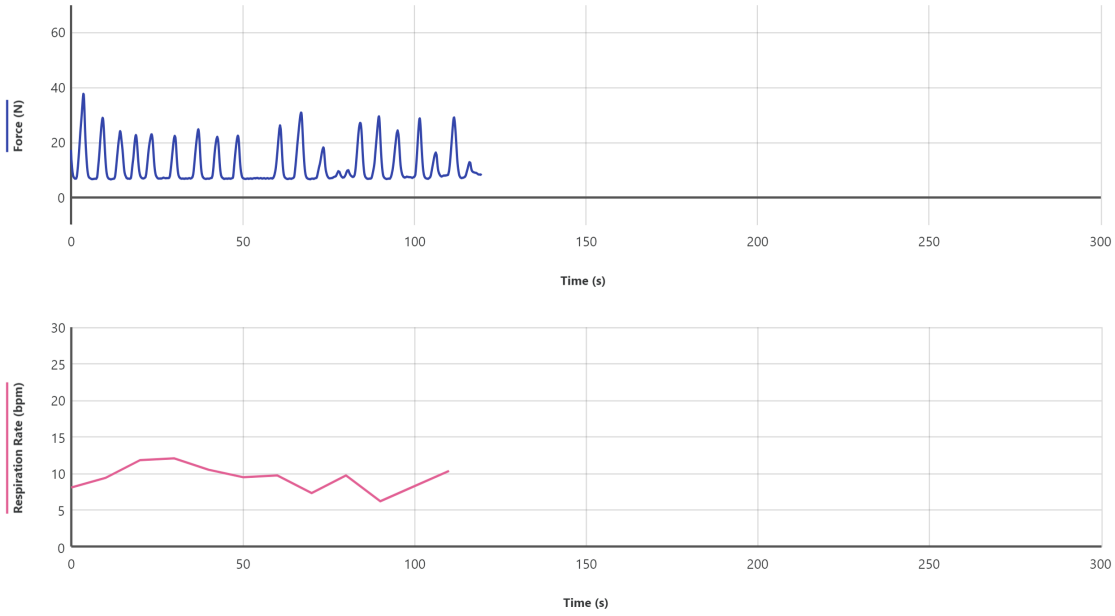


Figure 4.15: Panel view of the Vernier Graphical Analysis™ software displaying one sample of the subject's breathing signal (upper image) measured with Go Direct® Respiration Belt. The bottom image shows the respiration rate which updates every 10 seconds. The sample window for the RR calculation is 30 seconds.

data when the subjects are not directly facing the camera but they are moved perpendicularly from 0.9 m to 1.9 m depending on the distance. The moving range of the subject is limited by the FoV of the thermal camera that we must assure the full face is in the picture. In addition, the whole experiment has two sets of the different setups, where one requires the subject to wear the mask, while another does not, so there are totally 8 experiments for each subject. We call them the "mask group" and "no-mask group" in our experiments. With increasing the distance between the camera and the subject, the face becomes smaller in the image and so the interested region shrinks. However less pixels of the image makes it harder to extract the related signal. Our method works without any problem in estimating the signal from the distances that are less than 1 meter. What we want to explore is the potential whether our system can estimate the respiratory signal at a longer distances.

Through the experiment, we found that "mask group" can go further distance than the "no-mask group" while keeping the same level of accuracy. The former one can detect the respiratory signal when the subject is even at 3 meters from the camera. However, "no-mask group" performs badly when the distance between the subject and camera is over 1 meter. Figure 4.16 shows the waveform of estimated BR signal and baseline BR signal of one subject not wearing the mask and standing at 1m distance away from the system. Similarly, Figure 4.17 shows the waveform of both estimated and baseline BR signal of the same sub-

ject wearing mask and standing at 2.5m distance away from our system. In both Figure 4.16 and Figure 4.17, the first row represents the waveforms of the signal extracted from the detected nostril RoI in the thermal frames, and the second row shows the synchronous baseline respiratory signal collected from breathing belt with the same timestamp. Besides, the red marked waveform in the first column is zoomed in and displayed in the second column for the inspection. It is obviously to see that the estimated BR signal with "mask group" has the more similar waveform compared to the synchronous baseline BR signal even at further distance. All the signal waveforms were analyzed every 15 seconds with 50 percent overlapping. This work is done by another graduate students in the group where BR estimation relies on accurate detection and tracking of the nostril RoI which is done using algorithms presented in Section 3.2. We only show here the results of BR estimation from the thermal frames. The average mean absolute error (MAE) of "no-mask group" at 1 meter is less than 2 breath per minute (bpm) compared to the reference data from breathing belt.

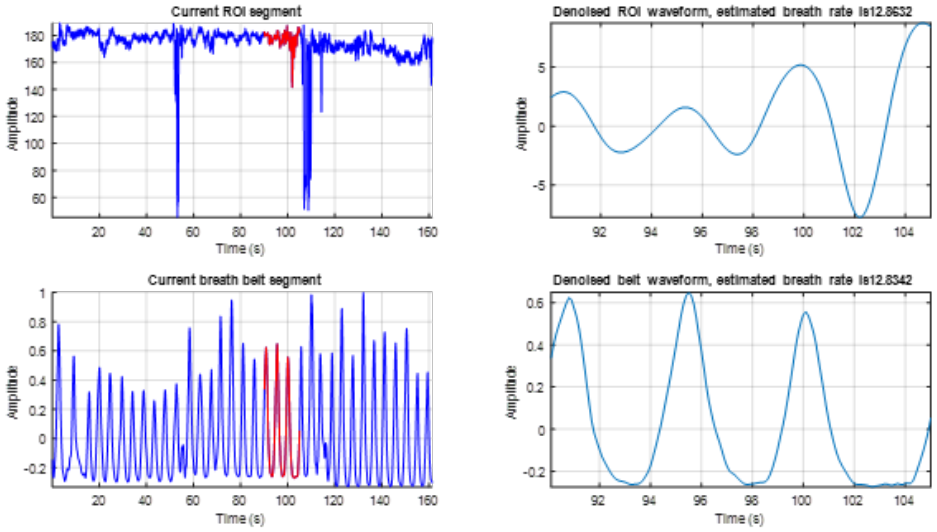


Figure 4.16: BR signal without mask at 1m.

Another big factor that affects the performance of respiratory signal estimation is the distance. With the distance between the subject and camera increasing, the mean absolute error (MAE) between the estimated BR and reference BR increases from less than 1 bpm to even more than 2 bpm. We also found that a smaller MAE of the BR and the RR result estimated by our algorithm for both the "mask group" and "no-mask group" at no more than 1.5 meters. Under this condition, we can make sure a MAE around 1 bpm compared to the breathing belt reference BR, which allows the system to be applied in real world RR measurement. The MAE between the estimated RR and the reference RR for all subjects at 4 different distances and under 2 different conditions is 1.55 ± 0.78 breaths per minute

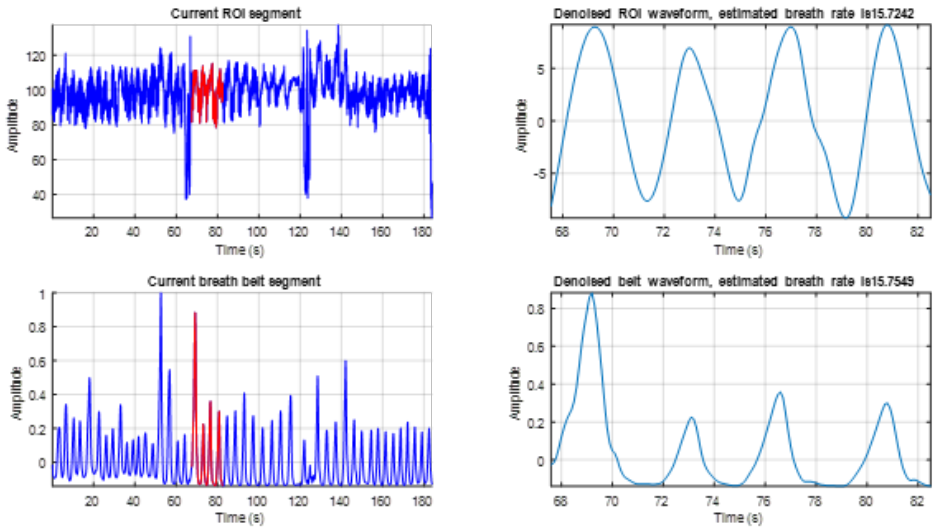


Figure 4.17: BR signal with mask at 2.5m.

Table 4.2: Mean and standard deviation of the absolute error between the estimated RR and the reference RR (unit:bpm, bpm: breaths per minute)

	Subject1		Subject2		Subject3		Subject4	
	M_{eRR}	SD_{eRR}	M_{eRR}	SD_{eRR}	M_{eRR}	SD_{eRR}	M_{eRR}	SD_{eRR}
60cm	2.72	2.57	1.45	1.47	0.78	0.77	2.64	1.51
60cm mask	1.90	1.45	1.64	1.87	0.50	0.68	1.92	1.71
80cm	1.28	1.88	2.56	1.57	1.78	1.38	1.61	2.12
80cm mask	1.91	1.07	0.86	0.82	0.79	0.40	0.38	0.56
100cm	1.79	2.38	3.38	2.87	1.87	2.09	2.44	1.59
100cm mask	0.88	1.39	1.31	1.00	0.73	1.03	0.59	0.66
120cm	0.62	0.75	1.40	1.86	1.87	1.78	2.97	2.64
120cm mask	0.74	1.26	1.20	1.72	1.17	0.88	1.87	1.32

as listed in Table 4.2. However, the waveform of the thermal frame extracted signal is not perfectly matching the reference data, showing some delay at timeline and some noises. In our future work, we will discuss the approach to filter out the noise of extracted signal and so that we could have a better analysis of the respiratory signal from contactless measurement.

Besides, the feasibility of estimating multiple subjects' respiration rates at the same time using our proposed methods was verified. Two subjects with face masks were required to stand in front of the cameras for two minutes at a distance of 80cm and 100cm respectively as illustrated in Figure 4.18(a). The breathing belts were used to collect reference respiratory waveform from the chests. An example of the extracted respiratory waveform using the aforementioned methods were shown in Figure 4.18(b) where blue color and green color describe the respiratory waveform from two subjects at the same time. The mean absolute error (MAE) between the estimated respiration rate and the reference respiration rate for Subject1 is 0.97 ± 0.714 bpm, and the MAE between the estimated respiration rate and the reference respiration rate for Subject2 is 0.35 ± 0.293 bpm.

4.5.2 Heart Rate Estimation

Different form extracting pixel value variation from the thermal frames, heart rate estimation is based on the subtle color changes of the face and thus has a higher requirements on the RoI selection. One of the most significant factors that affect the HR estimation is the image resolution. We compared 640×480 resolution RGB frames and 1080×720 resolution RGB frames, and found that the higher image resolution allow for more accurate estimation of HR than the lower one compared to the reference HR. The MAE between the estimated HR using our proposed method and the reference HR from PPG sensors for all subjects at different distances is 4.24 ± 2.47 beats per minute. For application scenarios like detecting the vital signs of an emergency situation, HR measurement with error less than 5 bpm is likely to be acceptable [78].

As illustrated in Figure 4.19, we assess our HR estimation methods by comparing the ICA signal extracted from detected RoI of the subject's face and the PPG baseline signal collected at the same time. Besides, we calculate the HR value from both the signals by peak finding algorithms described in Section 3.2.4. Our methods have an average 4.24 ± 2.47 beats per minute (bpm) of the MAE compared to the baseline HR when the subject is at less than 1 meter far away from the camera. When the person is wearing the mask and at the same distance as before, the accuracy goes a bit down and the MAE is around 5 bpm compared to the baseline data because the occluded face causes the less area for the signal extraction as listed in Table 4.3. Compared to previous work [81], we improve the distance between the

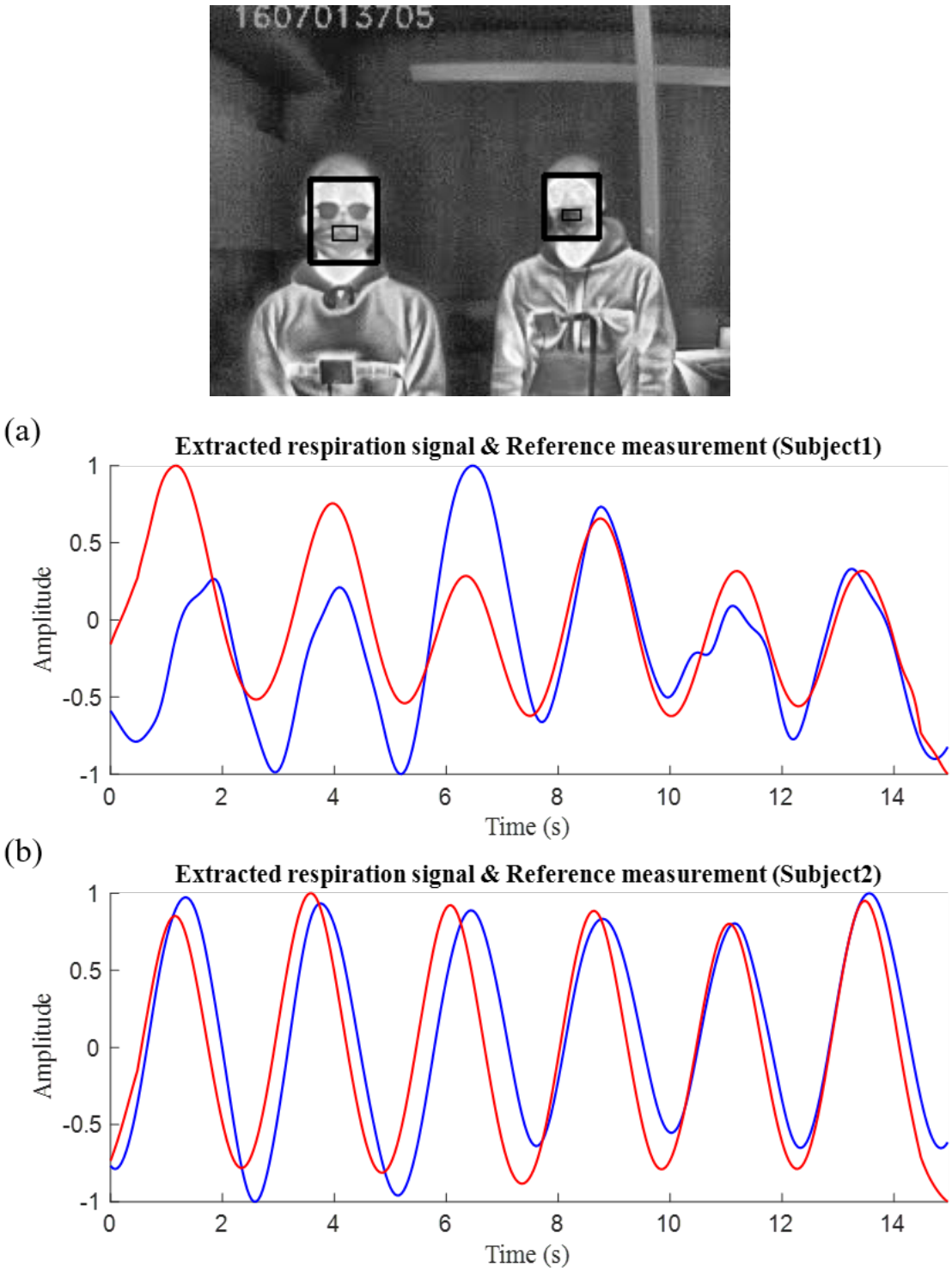


Figure 4.18: Two subjects are standing in front of the thermal camera during the measurement of RR, the two faces and the respective $ROI_{nostril}$ are detected by our method at the same time.(a) Extracted respiration waveform from thermal video and simultaneous reference respiration waveform (Subject1, blue: extracted waveform, red: reference waveform. (b) Extracted respiration waveform from thermal video and simultaneous reference respiration waveform (Subject2, blue: extracted waveform, red: reference waveform)

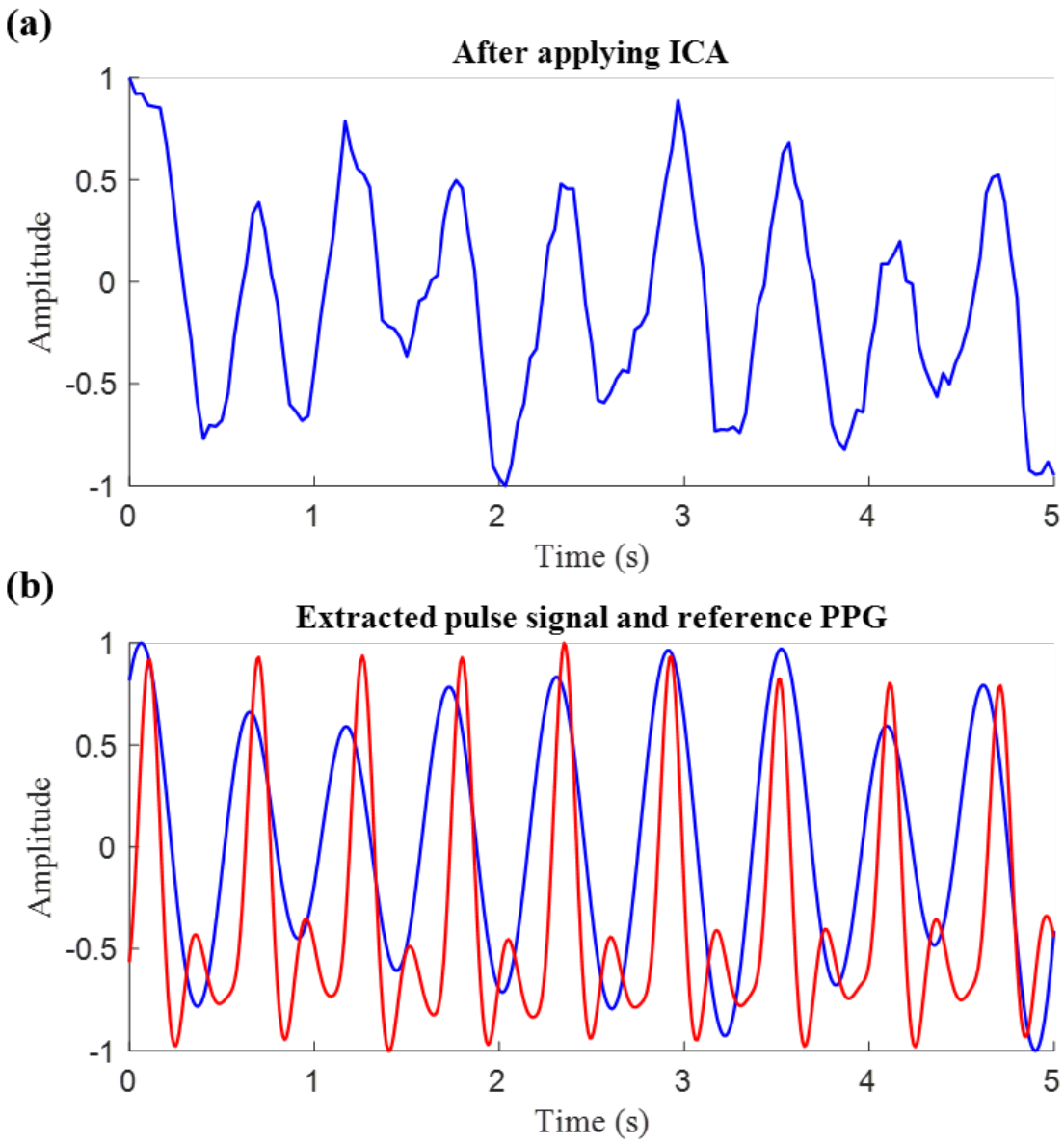


Figure 4.19: (a) Pulse signal extracted from video using ICA (b) Denoised pulse-like signal (resampled to 100Hz) and simultaneous reference PPG signal, blue: video pulse signal; red: reference PPG.

Table 4.3: Mean and standard deviation of the absolute error between the estimated HR and the reference HR (unit:bpm, bpm: beats per minute)

	Subject1		Subject2		Subject3		Subject4	
	M_{eHR}	SD_{eHR}	M_{eHR}	SD_{eHR}	M_{eHR}	SD_{eHR}	M_{eHR}	SD_{eHR}
60cm	3.05	1.76	1.17	0.68	1.39	0.45	0.50	0.43
60cm mask	5.15	4.33	4.15	2.15	3.82	4.02	8.70	5.90
80cm	5.34	4.42	3.06	1.21	2.46	3.36	0.89	1.00
80cm mask	5.13	4.30	5.73	2.34	3.56	2.56	10.37	5.78
100cm	2.92	2.99	4.91	3.18	1.98	1.97	1.02	1.19
100cm mask	5.47	5.61	2.88	2.54	4.05	3.71	4.62	3.19
120cm	4.42	4.17	4.11	2.25	3.79	3.02	9.13	4.92
120cm mask	4.60	5.44	2.80	1.70	4.91	3.61	9.66	6.88

device and the subject from 35cm up to 120cm and keep the error under 5 bpm as well for the near distance. Our methods also address the challenge where the subject is wearing the mask. However, the camera resolution and illumination have the impact on the estimation of HR from videos which will be also discussed in next chapter.

Chapter 5

Final Remarks

5.1 Conclusion

Through the research of both the depth camera based system and thermal camera based system, we explore the CNN based object detection algorithms. Our overall camera-based systems show the possibility of contactless and remote monitoring of vital signs with acceptable accuracy. In terms of depth camera based system, we modify the CNN based object detection algorithm Mask R-CNN to adapt to our special depth images. Transfer learning is applied here for the training of our body tracking model based on our own dataset. The well-trained model shows the ability to detect the location of the person lying in bed. In addition, our signal processing algorithms prove the accurate estimation of the subject's respiratory signal based on the position detected by our object detection model. Similarly, we also utilize another CNN based object detection framework in the scenario of contactless vital signs estimation. Our thermal camera based system applies the CNN based RetinaFace detector to detect the RoI of subjects' faces. Our approach shows the possibility to be applied in real world applications today especially following the requirements of Covid-19 pandemic policies.

Our manually annotated dataset carries depth images with labeled class and annotated mask of the people's upper body. As much as our dataset helps the model to learn the mask distributions, it carries noisy annotations as our comparisons show, but increasing size of such noisy data can compensate for its noise as shown by our results. Though Mask R-CNN framework allows for the small size of dataset due to the transfer learning like our first mask-oriented dataset, our dataset composed of four-keypoints annotation is not applicable to train a robust model and needs more data.

With increasing concerns on privacy protection and mass hysteria of public in making sensitive data available for research, the creation of dataset containing personal information

needs more examination. Our data are collected according to the ethics approval and completely meets the requirements of both the school laboratory and cooperated hospital and retirement home. In order to protect the privacy of subjects from the hospital and retirement home, we only use depth camera and radar device to collect their health data without collecting any identity information including RGB face images of the subjects. Similarly, the data collected in the laboratory also satisfies the related requirements, and we keep the whole data in the encrypted drives.

Our experiments show the feasibility of our approaches to address the challenges and problems in human vital signs remote monitoring. The experiments prove the ability of our methods in detecting people's occluded body in the depth images and occluded faces in the RGB frames. Besides, within the combination of radar sensor and thermal camera, additional information such as respiratory signal and body temperature are extracted. In addition by comparing with the standard reference data collected from the wearable devices at the same time, our contactless vital signs estimation approaches have respectable results under the limited conditions and provide us with the potential to be applied in real world vital signs monitoring.

5.2 Summary of Contributions

We designed a system composed of depth camera and radar device to collect data from patients at a hospital. We used annotation tools to create a dataset that contains manually labeled depth images with part of the human body mask and some keypoints. For our objective of extracting vital signs of the subject, we utilized a state-of-the-art (SOTA) CNN based model with pre-trained weights. We fine-tuned the model with transfer learning for our task of object detection and RoI detection. We also compared different CNN based object detection frameworks that are designed to detect faces in different scenes. Our minor contribution also includes the integration of the SOTA CNN face detector to our contactless vital signs estimation methods offering accurate RoI detection for the accurate estimation of subject's vital signs. We improved the existing contactless vitals signs monitoring methods by allowing the subjects to wear the mask during the detection and multiple subjects to be detected at the same time if they are in the field of view. We also enhances the operation range in the remote detection of respiratory signal compared to the existed methods by applying the SOTA face detector.

5.3 Boundaries and Limitations

Our system and methods still have many boundaries and limitations according to the goals discussed in the Section 1.2, even though we have addressed some of the challenges and resolved the existed problems by the contribution we make in this thesis. We are going to discuss these limitations in this section and also the future work based on the current boundaries in the next section.

One of the biggest limitations of our methods is the dataset. Since we created all of the dataset by ourselves, the scope and extent of our dataset is small compared to other public datasets. The previously designed experiments of both the depth camera based research and thermal camera based research have limitations regarding the amount of data that we could collect. Not only each experiment done in the university laboratory but also every collection made in the hospital have limitations in the duration time, the number of the subjects an so on. Additionally, we were not able to recruit more volunteers for the experiment because of the Covid-19 pandemic policies. So the set cannot contain all of the possible scenarios in our database. The people's postures and locations in the depth images have certain pattern which means that our dataset is basically biased. Similarly, the collected baseline respiratory signal and heart rate only include limited amounts of possible samples in the real world without any unhealthy subjects. Another reason that limits the scale of our dataset is due to the annotation work. We have a small dataset composed of less than 2,000 depth images that needs weeks of annotation work. This is also a very common issue in many of the medical image researches. However, the current deep learning models like CNN based frameworks require as much as possible data for the training process of supervised learning. Therefore, the lack of well labeled datasets exists almost in every field of the related research. Because our problem is different from the standard task of object detection, we have specific subjects needed to be detected in our task. In addition, sometimes the source image data have different format compared to the normal RGB format, like the depth images and thermal images.

Another big limitation is the experiment device, and especially the cameras used in the experiment which have several limitations. The image quality of our collected depth data is influenced by many factors like environmental noises and the way we store the data. For example, the florescent light in the hospital ward interfere with the infrared reception of the depth camera thus generating black dots as shown in Figure 3.3. Besides, some special material and fabrics of the clothes still absorb the infrared radiation and make noises on the depth data as discussed in Section 4.3.2. In addition, we have trouble in capturing the subjects in the depth image at a good angle because the installation of our system has been limited to specific places around the patient's bed so that the devices do not disturb the patient

and do not cause potential fall of the patient if they are on his/her way. The device cannot be installed above the head of the bed where patients are resting so that the view of our depth images are narrow and a little leaned. The lossy compression method we used to compress the depth frames also decreases the image quality when we store the collected the data into the local drive. In terms of the thermal camera, the illumination is still one of the factor that has impact on the image quality of the collected data. Besides, the frames per second (fps) of the thermal camera in our system varies with the time, and the dropped frame rate of the camera causes the missing frames when we synchronise the RGB video and thermal video. For now, we drop the few frames where the frame rate is much lower than the threshold of 25 frames and interpolate the median value into the processed RR signal and HR signal.

5.4 Future Works

We have already discussed the boundaries and limitations of our system and methods proposed in this thesis in the last section, and so that we can make some modifications and improvements to enhance the performance of our approaches in terms of both the hardware and the software in the future.

In order to collect more data using our system each time, the storage space of the mini-computer needs to be increased. And with the high-volume of the storage, we could also modify the restoring methods of the image data to lossless strategy so that the quality of the images remains unchanged. Meanwhile, the big storage would also allow us to increase the fps of the camera so that we could collect depth images at over 30 fps. Besides, using this kind of depth videos, the respiratory signals can be even directly extracted from the frames once the subject is still and the chest region is detected like the methods introduced in [100].

In terms of dataset, we can also make some adjustments and improvement based on the current version. First of all, the scales of dataset needs a big increment. We can take some time to label more depth images not only with the mask annotation but also the keypoints annotation in the future. Our present dataset is insufficient due to our time limitation in creating the dataset. Secondly, the contents of the dataset need a screening. We should make our dataset as little biased as possible which means that the collected data from the subjects both at the hospital and the laboratory need to be categorized and selected. The dataset should contain different human postures at different locations.

With the availability of resources, we were limited to to train our models on 11GBs of GPU memory with which we could safely experiment with a model of size of over 20 GBs or even more. In future, by utilizing more resources, pre-trained models of the similar architecture with a higher number of parameters can be trained in these settings after appropriate

modifications.

We observed the validation loss increases when the Mask R-CNN model is trained for detecting the four-keypoints of the people's upper body in the depth images. This phenomenon is the reflection of overfitting during the training, and it might be caused by many reasons. One of them is the scale of training dataset and validation dataset is too small. In addition to increase the the size of the dataset, we can also use transfer learning to train based on other related datasets. For example, COCO has keypoints dataset for the human pose detection, however, each person is labeled with 17 keypoints of the whole body. We can modify it by keeping only four of the keypoints to first obtain a network model, and then apply this pre-trained weights to our depth images keypoints detection.

Finally, our thermal camera based system can be applied in real-time in the near future. Our methods are proven to be fast enough to run in real-time like which is discussed in Section 4.4.2. The further work here is about the hardware selection. We should consider both the computing performance and the cost when we attempt to integrate our algorithms to the hardware and make our system used in the real world applications. Once the algorithms for face detection together with RoI detection are executed fast enough to support the frame rate of more than 25 fps, the overall vital signs estimation methods could run in real-time and the estimated results could be also shown immediately. There are also some works about software optimization could be done in the future.

Appendix A

Ethic Approvals

A.1 Depth Camera Research

This section lists the ethics approvals of depth camera based research.

January 31, 2019

Dr. Peter Liu

The University of Ottawa Heart Institute 40 Ruskin Avenue, Room H-2238 Ottawa, ON K1Y 4W7

Re: UOHI Institutional Approval for Ottawa Health Science Network Research Ethics Board (OHSN-REB) Submission

20180637-01H;

Respiration pattern monitoring of heart failure patients using non-contact sensors: a sub-study of Biomarker candidates to guide discharge of patients admitted to hospital with heart failure (RADAR)

Dear Dr. Peter Liu,

This letter serves as **University of Ottawa Heart Institute (UOHI)** Institutional Approval for the above-referenced study. Please maintain this documentation in your investigator study file.

Based on the information you provided about this study through the Clinical Research Registration Form, you have satisfied the requirements for institutional (UOHI) approval. This includes initial research ethics approval by OHSN-REB, appropriate departmental/service area notifications and execution (fully signed versions) of all agreement(s) required to begin the study locally. Please note there may be additional agreement(s) pending execution that are required to send funds, samples, or data to external sites, but are not required for you to begin your study locally.

Changes and/or additions to your study that may require additional agreement(s) or revisions to existing agreement(s) must be communicated to the UOHI Legal Affairs. This should be undertaken simultaneously with any related OHSN-REB amendment submission.

Changes and/or additions to your study that affect various hospital/institution departments (e.g., pharmacy, Department of Medical Imaging, EORLA, EEG, etc.) must be communicated to the relevant departments.

As mentioned in the 'Response' tab of the Ethics application, you have 3 months from the date of initial OHSN-REB approval to submit French documents including the translation certificate to OHSN-REB through the Translated Documents section of the ethics application (if applicable).

Should you have any questions, please contact REBadministration@ohri.ca or 613-798-5555 extension 16719.

Figure A.1: Ethics Approval of depth camera research

A.2 Thermal Camera Research

This section lists the ethics approvals of thermal camera based research.

17/12/2020

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE | CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-09-20-6144
Titre du projet / Project Title	THERMAL IMAGING FOR EFFICIENT DETECTION OF VITAL SIGNS DURING COVID-19 PANDEMIC
Type de projet / Project Type	Recherche de professeur / Professor's research project
Statut du projet / Project Status	Approuvé / Approved
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	17/12/2020
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	16/12/2021

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Miodrag BOLIC	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Chercheur Principal / Principal Investigator
Shan HE	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Co-chercheur / Co-investigator
Fan YANG	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Co-chercheur / Co-investigator

Conditions spéciales ou commentaires / Special conditions or comments

Figure A.2: Ethics Approval of thermal camera research

Appendix B

Dataset Format

```
{"0028.jpg111097":{"fileref":"","size":111097,"filename":"0028.jpg","base64_img_data":"","file_attributes":{},"regions":{"0":
{"shape_attributes":{"name":"polygon","all_points_x":
[362,363,371,371,371,380,391,393,397,394,373,363,358,359,349,352,353,353,364,395,397,417,431,474,522,540,586,599,589,583,575,569,562,557,55
4,545,535,522,517,511,500,490,477,475,477,475,475,476,483,493,492,491,491,362],"all_points_y":
[2,14,22,31,40,53,58,69,75,84,102,122,131,147,165,187,210,214,234,251,261,298,302,315,293,287,269,190,173,158,140,125,114,92,79,67,59,51,48
,48,48,48,47,40,35,29,24,23,20,20,13,4,0,2]},"region_attributes":{}}}},"0035.jpg88321":
{"fileref":"","size":88321,"filename":"0035.jpg","base64_img_data":"","file_attributes":{},"regions":{"0":{"shape_attributes":
{"name":"polygon","all_points_x":
[465,448,441,429,420,409,398,387,375,361,346,345,347,347,345,341,332,315,297,292,290,290,288,291,295,295,301,305,335,341,345,351,395,412,42
8,455,474,501,527,553,587,618,628,641,646,652,661,675,690,704,717,721,723,720,715,713,707,699,698,710,720,742,804,809,809,821,835,857,885,8
83,914,926,939,958,982,988,986,987,985,972,962,946,929,910,901,881,855,813,789,781,783,788,797,830,881,883,880,884,885,880,876,869,863,857,
849,847,846,846,848,849,849,843,835,813,791,773,757,746,734,720,705,693,674,655,643,627,613,596,581,567,546,540,540,500,484,475,465],"all_p
oints_y":
[366,363,355,349,344,343,336,333,333,339,341,345,360,372,375,383,392,394,397,398,409,423,431,442,450,461,468,471,496,495,510,516,504,510,53
1,556,571,571,571,579,595,624,645,661,676,687,694,697,699,700,691,674,651,630,607,592,578,559,542,534,539,535,455,430,423,413,389,350,300,2
76,238,230,224,216,206,208,196,182,179,175,173,181,188,194,198,216,227,274,295,291,277,264,249,218,190,172,166,151,144,132,124,118,117,117,
124,130,138,147,156,164,172,182,190,200,206,216,231,239,249,261,275,277,282,288,292,300,306,313,320,332,345,349,359,382,382,376,366]},"regi
on_attributes":{}},"1":{"shape_attributes":{"name":"polygon","all_points_x":
[474,465,460,468,477,467,476,479,482,498,505,502,492,477,511,516,549,615,622,628,629,636,646,658,673,723,729,726,717,726,730,733,737,750,75
0,743,741,727,721,704,689,693,684,672,662,653,618,618,620,623,478,474],"all_points_y":
[11,20,25,31,32,45,53,63,78,105,114,122,133,141,188,190,199,210,214,236,243,247,245,235,223,197,188,182,185,157,146,134,118,90,82,76,75,75,
75,74,74,69,60,55,51,45,13,11,6,0,0,11]},"region_attributes":{}}}},"0042.jpg89403":
{"fileref":"","size":89403,"filename":"0042.jpg","base64_img_data":"","file_attributes":{},"regions":{"0":{"shape_attributes":
{"name":"polygon","all_points_x":
[487,485,483,483,476,483,490,490,501,506,507,507,548,569,585,606,608,615,614,630,634,631,631,646,689,706,716,731,735,724,739,750,750,758,75
8,758,755,745,737,729,715,704,698,692,686,677,666,654,647,636,633,629,628,629,637,649,651,492,492,487],"all_points_y":
[12,21,30,38,46,51,60,66,89,105,116,128,150,154,165,194,201,211,220,220,229,245,254,244,211,200,195,195,188,184,141,114,107,90,85,82,75,74,
73,73,73,72,70,62,57,56,54,43,37,30,24,21,14,10,7,5,2,2,6,12]},"region_attributes":{}},"1":{"shape_attributes":
{"name":"polygon","all_points_x":
[602,618,645,660,679,770,787,802,843,847,848,849,847,846,846,850,855,861,868,870,881,881,881,882,881,873,850,843,825,782,779,782,784,799,81
4,834,883,910,932,957,975,992,993,991,979,917,839,831,779,804,803,805,709,694,702,709,713,718,723,723,724,730,725,721,719,721,717,694,666,6
55,623,614,467,430,400,370,364,337,311,295,293,286,286,286,291,294,303,315,323,336,350,349,346,363,385,405,423,485,502,544,545,556,570,582,
595,602],"all_points_y":
[311,302,291,285,281,229,222,215,189,180,167,156,144,135,128,127,123,121,121,119,127,134,152,163,174,199,227,232,246,275,284,287,291,285,27
4,254,222,207,190,180,178,187,198,205,208,264,323,320,388,415,425,455,551,554,566,582,594,607,620,629,638,639,650,663,672,679,687,702,695,6
95,678,656,599,592,582,550,533,523,509,475,460,440,430,412,402,398,395,395,394,392,380,368,349,336,335,334,343,383,389,350,344,340,334,327,
313,311]},"region_attributes":{}}}},"0050.jpg85001":
```


Figure B.1: JSON file example of people's mask dataset.


```
{"id":125,"dataset_id":7,"category_ids":
[1,2],"path":"/datasets/val/13.jpg","width":1280,"height":720,"file_name":"13.jpg","annotated":true,"annotating":
[],"num_annotations":2,"metadata":{},"milliseconds":28825,"events":
[{"_cls":"SessionEvent","user":"fyang064","milliseconds":28825,"tools_used":[]}],"regenerate_thumbnail":false,"is_modified":false},
{"id":126,"dataset_id":7,"category_ids":
[1,2],"path":"/datasets/val/16.jpg","width":1280,"height":720,"file_name":"16.jpg","annotated":true,"annotating":
[],"num_annotations":2,"metadata":{},"milliseconds":50659,"events":
[{"_cls":"SessionEvent","user":"fyang064","milliseconds":50659,"tools_used":
[]}],"regenerate_thumbnail":false,"is_modified":false}],"categories":
[{"id":1,"name":"4keypoints","supercategory":"","color":"#7a4bd9","metadata":{},"creator":"fyang064","keypoint_colors":
["#bf5c4d","#d99100","#4d8068","#0d2b80"],"keypoints":["left_shoulder","right_shoulder","neck","chest"],"skeleton":[[1,3],[2,3],[3,4]]},
{"id":2,"name":"mask","supercategory":"","color":"#9d195b","metadata":{},"creator":"fyang064","keypoint_colors":[]}],"annotations":
[{"id":253,"image_id":102,"category_id":2,"dataset_id":7,"segmentation":
[[317.3,353.7,320.4,361.8,328.4,369.9,337.6,384.1,395.2,398.2,419.5,406.3,429.6,397.2,433.6,389.1,444.8,374,457.9,338.6,467,323.4,475.1,314
.3,484.2,307.2,490.3,297.1,521.7,265.7,551,262.7,578.3,264.7,617.7,272.8,642,295.1,667.3,352.7,688.6,355.8,697.7,360.8,689.6,421.5,679.4,45
3.9,659.2,479.2,645.1,490.3,614.7,506.5,578.3,531.8,554,565.1,521.7,600.5,481.2,613.7,427.6,611.7,399.3,604.6,387.1,598.5,348.7,587.4,313.3
,554,297.1,524.7,290,517.6,275.8,495.3,274.8,456.9,262.7,432.6,260.7,394.2,271.8,372.9,292,345.6,316.3,350.7]],"area":96111,"bbox":
[261,263,437,351],"iscrowd":false,"isbbox":false,"creator":"fyang064","width":1280,"height":720,"color":"#9dbf23","metadata":
{},"milliseconds":18777,"events":[{"_cls":"SessionEvent","created_at":
{"$date":1595186374593},"user":"fyang064","milliseconds":18777,"tools used":["Polygon"]}]},
```


Figure B.2: JSON file example of people's keypoint dataset.

Reference

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In: *Neural Information Processing Systems* 25 (Jan. 2012). DOI: 10.1145/3065386 (*see pp. 1, 20, 21, 26*).

[2] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A Deeper Look at Dataset Bias. In: *CoRR* abs/1505.01257 (2015). arXiv: 1505.01257. URL: http://arxiv.org/abs/1505.01257 (*see p. 1*).

[3] EU. 2016. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). Retrieved December 26, 2018 from. In: 2018 (*see p. 1*).

[4] Dengsheng Lu. A Survey of Image Classification Methods and Techniques for Improving Classification Performance. In: *International Journal of Remote Sensing* 28 (Mar. 2007), pp. 823–870. DOI: 10.1080/01431160600746456 (*see p. 1*).

[5] Bharath Hariharan, Pablo Arbelaez, Ross B. Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. In: *CoRR* abs/1407.1808 (2014). arXiv: 1407.1808. URL: http://arxiv.org/abs/1407.1808 (*see p. 1*).

[6] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware Semantic Segmentation via Multi-task Network Cascades. In: *CoRR* abs/1512.04412 (2015). arXiv: 1512.04412. URL: http://arxiv.org/abs/1512.04412 (*see p. 1*).

[7] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-Based Object Tracking. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 25.5 (May 2003), pp. 564–575. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2003.1195991. URL: https://doi.org/10.1109/TPAMI.2003.1195991 (*see p. 1*).

[8] Thorsten Behrens, Karl Rohr, and H. Stiehl. Robust segmentation of tubular structures in 3-D medical images by parametric object detection and tracking. In: *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of*

the IEEE Systems, Man, and Cybernetics Society 33 (Feb. 2003), pp. 554–61. DOI: 10.1109/TSMCB.2003.814305 (*see p. 2*).

[9] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. In: *International Journal of Computer Vision* 88 (June 2010), pp. 303–338. DOI: 10.1007/s11263-009-0275-4 (*see p. 2*).

[10] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In: May 2011, pp. 1817–1824. DOI: 10.1109/ICRA.2011.5980382 (*see p. 2*).

[11] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. In: vol. 56. June 2011, pp. 1297–1304. ISBN: 978-3-642-28660-5. DOI: 10.1109/CVPR.2011.5995316 (*see pp. 3, 8*).

[12] Christoph Bernhard Hoog Antink, Simon Lyra, Michael Paul, Xinchi Yu, and Steffen Leonhardt. A Broader Look: Camera-Based Vital Sign Estimation across the Spectrum. In: *Yearbook of medical informatics* 28.01 (2019), pp. 102–114. ISSN: 2364-0502. DOI: 10.1055/s-0039-1677914. URL: https://publications.rwth-aachen.de/record/768483 (*see p. 6*).

[13] David Evans, Brent Hodgkinson, and Judith Berry. Vital signs in hospital patients: A systematic review. In: *International journal of nursing studies* 38 (Jan. 2002), pp. 643–50. DOI: 10.1016/S0020-7489(00)00119-X (*see p. 6*).

[14] Nuitrack. *NUITRACK SDK*. 2020. URL: https://nuitrack.com/ (*see p. 8*).

[15] Pushmeet Kohli Nathan Silberman Derek Hoiem and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In: *ECCV*. 2012 (*see p. 9*).

[16] Angel Martínez-González, Michael Villamizar, Olivier Canévet, and Jean-Marc Odobez. Real-time Convolutional Networks for Depth-based Human Pose Estimation. In: (Oct. 2019) (*see p. 9*).

[17] Intel. *Beginner's guide to depth (Updated)*. 2019. URL: https://www.intelrealsense.com/beginners-guide-to-depth/ (*see pp. 13, 14, 15*).

[18] Fraser Macdonald. *Thermal camera smartphone clip-on answers age-old question*. 2014. URL: https://www.stuff.tv/news/thermal-camera-smartphone-clip-answers-age-old-question (*see p. 17*).

[19] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. *Object Detection in 20 Years: A Survey*. 2019. arXiv: 1905.05055 [cs.CV] (*see p. 18*).

[20] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In: vol. 1. Feb. 2001, pp. I–511. ISBN: 0-7695-1272-0. DOI: 10.1109/CVPR.2001.990517 (*see pp. 18, 26*).

[21] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In: *Computational Learning Theory*. Ed. by Paul Vitányi. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 23–37. ISBN: 978-3-540-49195-8 (*see p. 19*).

[22] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)* 2 (June 2005) (*see p. 19*).

[23] Pedro Felzenszwalb, Ross Girshick, and David Mcallester. Visual Object Detection with Deformable Part Models. In: vol. 56. June 2010, pp. 2241–2248. DOI: 10.1109/CVPR.2010.5539906 (*see p. 19*).

[24] Pedro Felzenszwalb, Ross Girshick, David Mcallester, and Deva Ramanan. Object Detection with Discriminatively Trained Part-Based Models. In: *IEEE transactions on pattern analysis and machine intelligence* 32 (Sept. 2010), pp. 1627–45. DOI: 10.1109/TPAMI.2009.167 (*see p. 19*).

[25] D.H. Hubel and T.N. Wiesel. Receptive Fields and Functional Architecture of Monkey Striate Cortex. In: *The Journal of physiology* 195 (Apr. 1968), pp. 215–43. DOI: 10.1113/jphysiol.1968.sp008455 (*see p. 20*).

[26] Kunihiko Fukushima and Sei Miyake. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In: *Competition and Cooperation in Neural Nets*. Ed. by Shun-ichi Amari and Michael A. Arbib. Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 267–285. ISBN: 978-3-642-46466-9 (*see p. 20*).

[27] Vadim Romanuke. Appropriate Number and Allocation of RELUS in Convolutional Neural Networks. In: *Research Bulletin of the National Technical University of Ukraine Kyiv Politechnic Institute* (Mar. 2017), pp. 69–78. DOI: 10.20535/1810-0546.2017.1.88156 (*see p. 20*).

[28] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. 2014. arXiv: 1311.2524 [cs.CV] (*see p. 20*).

[29] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (Dec. 2015), pp. 1–1. DOI: 10.1109/TPAMI.2015.2437384 (*see pp. 20, 21*).

[30] Jasper Uijlings, K. Sande, T. Gevers, and A.W.M. Smeulders. Selective Search for Object Recognition. In: *International Journal of Computer Vision* 104 (Sept. 2013), pp. 154–171. DOI: 10.1007/s11263-013-0620-5 (*see p. 21*).

[31] Pedro Felzenszwalb, David Mcallester, and Deva Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In: vol. 8: June 2008. DOI: 10.1109/CVPR.2008.4587597 (*see p. 21*).

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In: *Lecture Notes in Computer Science* (2014), pp. 346–361. ISSN: 1611-3349. DOI: 10.1007/978-3-319-10578-9_23. URL: http://dx.doi.org/10.1007/978-3-319-10578-9_23 (*see pp. 22, 23*).

[33] Ross Girshick. *Fast R-CNN*. 2015. arXiv: 1504.08083 [cs.CV] (*see p. 22*).

[34] Olivier Fercoq, Zheng Qu, Peter Richtárik, and Martin Takáč. Fast distributed coordinate descent for non-strongly convex losses. In: *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2014, pp. 1–6 (*see p. 23*).

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV] (*see pp. 23, 25, 43, 45, 48*).

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: 1612.03144 [cs.CV] (*see pp. 23, 24, 47*).

[37] Petru Soviany and Radu Tudor Ionescu. *Optimizing the Trade-off between Single-Stage and Two-Stage Object Detectors using Image Difficulty Prediction*. 2018. arXiv: 1803.08707 [cs.CV] (*see p. 24*).

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV] (*see p. 24*).

[39] Joseph Redmon and Ali Farhadi. *YOLO9000: Better, Faster, Stronger*. 2016. arXiv: 1612.08242 [cs.CV] (*see p. 25*).

[40] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV] (*see p. 25*).

[41] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In: *Lecture*

Notes in Computer Science (2016), pp. 21–37. ISSN: 1611-3349. DOI: 10.1007/978-3-319-46448-0_2. URL: http://dx.doi.org/10.1007/978-3-319-46448-0_2 (*see pp. 25, 27*).

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV] (*see pp. 26, 28*).

[43] Haoxiang Li, Zhe Lin, Xiaohui Shen, and Jonathan Brandt. A convolutional neural network cascade for face detection. In: June 2015, pp. 5325–5334. DOI: 10.1109/CVPR.2015.7299170 (*see p. 26*).

[44] Sachin Sudhakar Farfade, Mohammad Saberian, and Li-Jia Li. *Multi-view Face Detection Using Deep Convolutional Neural Networks*. 2015. arXiv: 1502.02766 [cs.CV] (*see p. 27*).

[45] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV] (*see p. 27*).

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV] (*see p. 27*).

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. In: *IEEE Signal Processing Letters* 23.10 (Oct. 2016), pp. 1499–1503. ISSN: 1558-2361. DOI: 10.1109/lsp.2016.2603342. URL: http://dx.doi.org/10.1109/LSP.2016.2603342 (*see pp. 28, 47, 71*).

[48] Vidit Jain and Erik Learned-Miller. *FDDB: A Benchmark for Face Detection in Unconstrained Settings*. Tech. rep. UM-CS-2010-009. University of Massachusetts, Amherst, 2010 (*see p. 28*).

[49] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (*see p. 28*).

[50] Peter M. Roth Martin Koestinger Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In: *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*. 2011 (*see p. 28*).

[51] Jianfeng Wang, Ye Yuan, and Gang Yu. *Face Attention Network: An Effective Face Detector for the Occluded Faces*. 2017. arXiv: 1711.07246 [cs.CV] (*see p. 28*).

[52] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry Davis. *SSH: Single Stage Headless Face Detector*. 2017. arXiv: 1708.03979 [cs.CV] (*see p. 28*).

[53] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. *S^3FD: Single Shot Scale-invariant Face Detector*. 2017. arXiv: 1708.05237 [cs.CV] (*see pp. 29, 47, 71*).

[54] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. In: *International Journal of Computer Vision* 88 (June 2010), pp. 303–338. DOI: 10.1007/s11263-009-0275-4 (*see p. 30*).

[55] Dan Wu, Ping Yang, Guan-Zheng Liu, and Yuan-Ting Zhang. Automatic estimation of respiratory rate from pulse transit time in normal subjects at rest. In: Jan. 2012. DOI: 10.1109/BHI.2012.6211699 (*see p. 31*).

[56] E.A.F. Simoes, R Roark, S Berman, L Esler, and J Murphy. Respiratory rate: Measurement of variability over time and accuracy at different counting periods. In: *Archives of disease in childhood* 66 (Nov. 1991), pp. 1199–203. DOI: 10.1136/adc.66.10.1199 (*see p. 31*).

[57] Yee Siong Lee, Pubudu Pathirana, Christopher Steinfort, and Terry Caelli. Monitoring and Analysis of Respiratory Patterns Using Microwave Doppler Radar. In: *IEEE journal of Translational Engineering in Health and Medicine* 2 (Oct. 2014). DOI: 10.1109/JTEHM.2014.2365776 (*see p. 31*).

[58] Pubudu Pathirana, Sanvidha Herath, and Andrey Savkin. Multitarget Tracking via Space Transformations Using a Single Frequency Continuous Wave Radar. In: *Signal Processing, IEEE Transactions on* 60 (Oct. 2012), pp. 5217–5229. DOI: 10.1109/TSP.2012.2206588 (*see p. 31*).

[59] A.B. Hertzman and C.R. Spealman. Observation on the finger volume pulse recorded photoelectrically. In: *Am. J. Physiol.* 119 (Jan. 1937), pp. 334–335 (*see p. 32*).

[60] Yu Sun and Nitish Thakor. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. In: *IEEE Transactions on Biomedical Engineering* 63.3 (2016), pp. 463–477. DOI: 10.1109/TBME.2015.2476337 (*see p. 32*).

[61] Fokko Wieringa, Frits Mastik, and Antonius van der Steen. Contactless Multiple Wavelength Photoplethysmographic Imaging: A First Step Toward "SpO2 Camera" Technology. In: *Annals of biomedical engineering* 33 (Sept. 2005), pp. 1034–41. DOI: 10.1007/s10439-005-5763-2 (*see p. 32*).

[62] Ming-Zher Poh, Daniel McDuff, and Rosalind Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. In: *IEEE transactions on bio-medical engineering* 58 (Oct. 2010), pp. 7–11. DOI: 10.1109/TBME.2010.2086456 (*see p. 32*).

[63] Fang Zhao, Meng Li, Yi Qian, and Joe Tsien. Remote Measurements of Heart and Respiration Rates for Telemedicine. In: *PloS one* 8 (Oct. 2013), e71384. DOI: 10.1371/journal.pone.0071384 (*see pp. 32, 33*).

[64] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian Video Magnification for Revealing Subtle Changes in the World. In: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31 (July 2012). DOI: 10.1145/2185520.2185561 (*see pp. 32, 34*).

[65] Stephanie Bennett, Rafik Goubran, and Frank Knoefel. Comparison of motion-based analysis to thermal-based analysis of thermal video in the extraction of respiration patterns. In: vol. 2017. July 2017, pp. 3835–3839. DOI: 10.1109/EMBC.2017.8037693 (*see p. 32*).

[66] Xiaochuan He, Rafik Goubran, and Frank Knoefel. IR night vision video-based estimation of heart and respiration rates. In: Jan. 2017, pp. 1–5. DOI: 10.1109/SAS.2017.7894087 (*see p. 32*).

[67] Carina Barbosa Pereira, Xinchi Yu, Michael Czaplik, Rolf Rossaint, Vladimir Blazek, and Steffen Leonhardt. Remote monitoring of breathing dynamics using infrared thermography. In: *Biomed. Opt. Express* 6.11 (Nov. 2015), pp. 4378–4394. DOI: 10.1364/BOE.6.004378. URL: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-6-11-4378 (*see p. 33*).

[68] Ramya Murthy and Ioannis Pavlidis. Noncontact measurement of breathing function. In: *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine and Biology Society* 25 (June 2006), pp. 57–67. DOI: 10.1109/MEMB.2006.1636352 (*see p. 33*).

[69] Stephanie Bennett, Rafik Goubran, and Frank Knoefel. The detection of breathing behavior using Eulerian-enhanced thermal video. In: vol. 2015. Aug. 2015, pp. 7474–7477. DOI: 10.1109/EMBC.2015.7320120 (*see p. 33*).

[70] Youngjun Cho, Simon J. Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. In: *Biomed. Opt. Express* 8.10 (Oct. 2017), pp. 4480–4503. DOI: 10.1364/BOE.8.004480. URL: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-8-10-4480 (*see p. 33*).

[71] Zheng Jiang, Menghan Hu, Lei Fan, Yaling Pan, Wei Tang, Guangtao Zhai, and Yong Lu. *Combining Visible Light and Infrared Imaging for Efficient Detection of Respiratory Infections such as COVID-19 on Portable Device*. 2020. arXiv: 2004.06912 [cs.CV] (*see p. 33*).

[72] Peter Rolfe. In Vivo Near-Infrared Spectroscopy. In: *Annual review of biomedical engineering* 2 (Feb. 2000), pp. 715–54. DOI: 10.1146/annurev.bioeng.2.1.715 (*see p. 34*).

[73] Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable Photoplethysmographic Sensors—Past and Present. In: *Electronics* 3.2 (2014), pp. 282–302. ISSN: 2079-9292. DOI: 10.3390/electronics3020282. URL: https://www.mdpi.com/2079-9292/3/2/282 (*see p. 34*).

[74] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart Rate Variability. In: *Circulation* 93.5 (1996), pp. 1043–1065. DOI: 10.1161/01.CIR.93.5.1043 (*see p. 34*).

[75] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. In: *Opt. Express* 16.26 (Dec. 2008), pp. 21434–21445. DOI: 10.1364/OE.16.021434. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-26-21434 (*see p. 34*).

[76] Yu Sun and Nitish Thakor. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. In: *IEEE transactions on bio-medical engineering* 63 (Sept. 2015). DOI: 10.1109/TBME.2015.2476337 (*see p. 34*).

[77] Ting Wu, Vladimir Blazek, and Hans Schmitt. Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes. In: *Proceedings of SPIE - The International Society for Optical Engineering* (Nov. 2000). DOI: 10.1117/12.407646 (*see p. 34*).

[78] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. In: *Opt. Express* 18.10 (May 2010), pp. 10762–10774. DOI: 10.1364/OE.18.010762. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-10-10762 (*see pp. 34, 53, 79*).

[79] Xiaochuan He, Rafik A. Goubran, and Xiaoping P. Liu. Using Eulerian video magnification framework to measure pulse transit time. In: *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2014, pp. 1–4. DOI: 10.1109/MeMeA.2014.6860029 (*see p. 34*).

[80] Stephanie L. Bennett, Rafik Goubran, and Frank Knoefel. Adaptive eulerian video magnification methods to extract heart rate from thermal video. In: *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2016, pp. 1–5. DOI: 10.1109/MeMeA.2016.7533818 (*see p. 34*).

[81] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote Heart Rate Measurement from Face Videos under Realistic Situations. In: June 2014, pp. 4264–4271. DOI: 10.1109/CVPR.2014.543 (*see pp. 34, 79*).

[82] Xiaochuan He, Rafik Goubran, Stephanie Bennett, Stephen Robinovitch, Bobbi Symes, Bryan Lo, Andreas Ejupi, and Frank Knoefel. Video-Based Analysis of Heart Rate Applied to Falls. In: *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2018, pp. 1–5. DOI: 10.1109/MeMeA.2018.8438773 (*see p. 35*).

[83] Jelena Nikolic-Popovic and Rafik Goubran. Impact of motion artifacts on video-based non-intrusive heart rate measurement. In: *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2016, pp. 1–6. DOI: 10.1109/MeMeA.2016.7533740 (*see p. 35*).

[84] Intel. *D400-Series-Datasheet*. https://www.intelrealsense.com/wp-content/uploads/2019/10/Intel-RealSense-D400-Series-Datasheet-Oct-2019.pdf. Accessed Jan 1, 2020 (*see pp. 38, 39*).

[85] Gary Bradski. The openCV library. In: *Dr. Dobb's Journal of Software Tools* 25 (Jan. 2000) (*see p. 39*).

[86] Abhishek Dutta and Andrew Zisserman. The VIA Annotation Software for Images, Audio and Video. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: ACM, 2019. ISBN: 978-1-4503-6889-6/19/10. DOI: 10.1145/3343031.3350535. URL: https://doi.org/10.1145/3343031.3350535 (*see p. 41*).

[87] Justin Brooks. *COCO Annotator*. https://github.com/jsbroks/coco-annotator/. 2019 (*see pp. 42, 58*).

[88] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV] (*see pp. 43, 44*).

[89] Jonathan Long, Evan Shelhamer, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2015. arXiv: 1411.4038 [cs.CV] (*see p. 44*).

[90] Xu Tang, Daniel K. Du, Zeqiang He, and Jingtuo Liu. *PyramidBox: A Context-assisted Single Shot Face Detector*. 2018. arXiv: 1803.07737 [cs.CV] (*see p. 47*).

[91] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. *RetinaFace: Single-stage Dense Face Localisation in the Wild*. 2019. arXiv: 1905.00641 [cs.CV] (*see pp. 47, 48*).

[92] Peiyun Hu and Deva Ramanan. *Finding Tiny Faces*. 2017. arXiv: 1612.04402 [cs.CV] (*see p. 47*).

[93] Ali Reza. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. In: *VLSI Signal Processing* 38 (Aug. 2004), pp. 35–44. DOI: 10.1023/B:VLSI.0000028532.53893.82 (*see p. 52*).

[94] J.F. Cardoso and Antoine Souloumiac. Blind Beamforming for non Gaussian Signals. In: *Radar and Signal Processing, IEE Proceedings F* 140 (Jan. 1994), pp. 362–370. DOI: 10.1049/ip-f-2.1993.0054 (*see p. 53*).

[95] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV] (*see p. 62*).

[96] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016 (*see p. 63*).

[97] Jason Yosinski, Jeff Clune, Y. Bengio, and Hod Lipson. How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems (NIPS)* 27 (Nov. 2014) (*see p. 63*).

[98] Nvidia. *Jetson Modules*. 2021. URL: https://developer.nvidia.com/embedded/jetson-modules (*see p. 72*).

[99] Vernier. *go-direct-respiration-belt*. 2021. URL: https://www.vernier.com/product/go-direct-respiration-belt/ (*see p. 75*).

[100] Yunlu Wang, Menghan Hu, Qingli Li, Xiao-Ping Zhang, Guangtao Zhai, and Nan Yao. *Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner*. 2020. arXiv: 2002.05534 [cs.LG] (*see p. 86*).