COMPUTER VISION BASED RECOGNITION OF DOCTOR'S ACTIONS DURING MEDICAL CONSULTATIONS

by

Laura Schiphorst

B.S., Applied Mathematics, Utrecht University, 2017

Submitted to the Institute for Graduate Studies in
Natural Science in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in MSc Artificial Intelligence
Utrecht University

2019-2020

COMPUTER VISION BASED RECOGNITION OF DOCTOR'S ACTIONS DURING MEDICAL CONSULTATIONS

APPROVED BY:

Prof. dr. A.A. Salah .....
(Thesis Supervisor)

Prof. dr. S. Brinkkemper .....

DATE OF APPROVAL: 15.01.2020

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. dr. Albert Ali Salah, for guiding me throughout the thesis work and for taking the time to help me in the process. I've learned many things from him for which I'm truly grateful. I am very thankful for having had the chance of working with him.

Additionally, I would like to thank Prof. dr. Sjaak Brinkkemper for his enthusiasm and continuous encouragements throughout the work. Under his supervision, Care2Report will achieve some great results, and I am thankful that I was allowed to do pioneering work in the action recognition field of the project.

Furthermore, I would also like to thank MSc. Metehan Doyran for taking the time and patience to guide me in the machine learning aspects of my thesis and MSc. Lientje Maas and MSc. Sabine Molenaar for helping me in the Care2Report aspects.

Moreover, I wish to express my gratitude to the medical student for providing important information and for the participation in creating our dataset. I would also like to thank MSc. Lientje Maas and Brandon Koffijberg for their acting work in the database.

Lastly, I am truly moved by the loving support of my family and friends and would like to thank their encouragements on my work and their never failing stimulating words and comfort when I needed it.

ABSTRACT

COMPUTER VISION BASED RECOGNITION OF DOCTOR'S ACTIONS DURING MEDICAL CONSULTATIONS

In the Netherlands, general practitioners have to prepare a report for each consultation and store this in the electronic medical record. This is time-consuming and automating the reporting procedure could solve this. However, recognising medical actions for the support of automatically storing patients information in the electronic medical record is limited, since there are no publicly available medical databases. Therefore, we present Video2Report, a database consisting of one-on-one medical consultations between a general practitioner and a patient. We construct a method that consists of selecting the most important medical actions and carefully recording and annotating the sessions. From the videos, we extract the skeleton positions by utilizing OpenPose. These skeleton positions are used to calculate useful mathematical information and use this to create feature sets. With these feature sets we will train and test three basic classifiers, i.e. a decision tree, random forest, and k-nearest neighbor classifiers. Our database consists of 192 sessions recorded with up to three cameras, accounting for a total of 451 videos, of which 332 consists of single actions and 119 consists of multiple action sequences. While Video2Report is too small for end-to-end deep learning, the results on the basic classifiers show promising results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbf{A}_{Z_{i1}}$	Activation output
$E(w)$	Error function
I_i	Input i
P_p	Predicted probability
W_i	Weight i
W_{ij}	Weights from input/neuron i to neuron j
X	Input X
Y_i	Output for neuron i
$\hat{Y}_i$	Predicted output for neuron i
Z_{i1}	Neuron 1
Δ	Gradient descent
η	Learning factor
$\nabla_w E$	Gradient Descent Vector
∂	Partial derivative

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two dimensional
3D	Three dimensional
C2R	Care2Report
CM	Confusion matrix
CNN	Convolutional neural network
CP	Care provider
DT	Decision trees
ECG	electrical cardiogram monitor
EMR	Electronic Medical Record
fps	Frames per second
GP	General practitioner
HAR	Human action recognition
k-nn	k-nearest neighbors
LSTM	Long Short Term Memory
MRQ	Main research question
MSE	Mean squared error
NHG	'Nederlandse Huisartsen Genootschap'
NN	Neural network
PDD	Product delivery diagram
RF	Random forest
SVM	Support vector machine
V2R	Video2Report

1. INTRODUCTION

In healthcare, the care providers (CPs) are obligated to accurately report on the encounters and treatments with their patients in their electronic medical record (EMR). These EMRs are designed for improved communication between CPs and capture previous diseases, treatments, and observations (1; 2). Moreover, they serve to comply with guidelines and can support medical decisions (3). Even though the EMRs support the medical care for patients, accurately documenting all aspects of healthcare is time consuming, since it is done manually by the CPs. A more efficient and less time-consuming way of reporting medical consultations is necessary. Automatically constructing and storing medical reports in the EMR may be a solution. Recognising medical actions from videos could aid in automatically constructing these reports and recent developments in human action recognition (HAR) provide promising results.

The proposed research is performed within the broader context of the Care2Report (C2R) project. This project aims to automatically report and document the medical documents in the Electronic Medical Record (EMR) by combining speech with action recognition to recognise the relevant (medical) actions that are performed during medical encounters. The C2R project is described more elaborately in Chapter 2.

This study aims for recognising medical actions from videos of medical consultations using Computer Vision. In order to do so, a suitable dataset on medical actions is required. Since this is currently not publicly available, we design and collect one ourselves. We conduct research to find the most relevant and occurring medical actions, such that we can record and annotate those. Then we retrieve the skeleton joints from the persons in the videos. With these skeleton joints, we can calculate distances between limbs and persons. These mathematical representations can be used to train our classifiers, those will then be able to recognise medical actions from new and unseen videos.

1.1. Human action recognition

In the field of computer vision, recognising human actions from videos has been extensively studied in the last years. Human actions can be detected, for instance, by using object detection, pose detection, or action spotting. Recognising human actions can be seen as finding a representation of the video and then classifying these into the right actions (4). Online processing means that the videos are processed in real time and this is useful in human-machine interaction, such as automatic surveillance, support in smart homes for elderly people, human-robot interaction, and recognising medical actions.

Recognising human actions started by recognising single actions from trimmed videos, i.e. videos with single actions. These actions were mostly performed by one actor per video, e.g. running, walking, and boxing. This focus has shifted towards analysis of multiple people simultaneously (5). Moreover, the use of cameras has also evolved. Starting with the use of a single fixed camera in 2000, this was expanded to using multiple cameras simultaneously (2003) and to using smart cameras (2004) (6).

Activities can be detected on several levels of abstraction. In (7), the authors defined different types of activity, namely action primitives, action, and activity, an approach we adopt here. An action primitive is a singular movement that, combined with other action primitives, results in an action. An activity is a combination of actions and depends on the environment, used objects and (human) interaction. As an example, in playing handball, an action primitive could be "running", "catching the ball", "jumping", and "throwing the ball". Combining these into an action would result in "jump shot". Combining several actions like "jump shot", "break out", "stopping the ball" etc., would create the activity "playing handball".

Many applications with regard to HAR exist, such as surveillance, elderly care, and healthcare. For instance, in healthcare, HAR can be used to recognise which medical actions are performed by the medical staff. This can be used to assist CPs in diagnosis of diseases or it can help in reporting medical information in the EMR, and

we focus on the latter.

While detecting human activities, we can focus on recognising variances within a movement to spot individuals, or on generalising over variances and recognise the action (4). The aim of this research is not to implement a system that identifies individuals in a patient-doctor scenario, rather our primary aim is to identify the actions of the GP, to recognize which action is being performed, selected from a closed set of pre-defined actions. The actions in a consultation session will be listed, and this will be the output of the proposed system. Since the analysis is primarily targeting post-consultation reporting, real time assessment is not required.

1.2. Automizing reporting in healthcare

In healthcare, CPs are obligated to report the encounters and treatments with their patients in the EMR. Even though the EMRs support the medical care for patients, accurately documenting all aspects of healthcare is time consuming. Using speech recognition to reduce the workload has been studied extensively. These vary from dictation for reporting to automatically subtracting clinical meaning directly (8; 9). Automatically subtracting clinical meaning consists of extracting relevant medical information from the conversation between the CP and their patient. This relevant information can be used for automatic reporting to the EMR. The developments of HAR technology provide opportunities for improved reporting to EMR. However, combining speech with medical action recognition has not yet been done.

Research in the domain of action recognition in healthcare has been limited to detecting fine-grained movements during eye surgery (10) and assisting in medical aid for elderly (6). However, it has not focused on recognising the medical actions as performed during consultations. Applying HAR on consultations can serve as a reinforcement to speech recognition.

Since the difficulty of recognising actions increases with the number of persons in the videos, we focus on human-human interactions between general practitioners (GPs)

and their patients, rather than a team of specialists operating simultaneously. More precisely, we aim to recognise the medical actions performed by the GP during a medical consultation, e.g. blood pressure measurement and auscultation of the heart and lungs. This provides a proof of concept for future research. This information will be used to automatically report all relevant information in the EMR. An example of different abstractions in the medical field could be as follows. A medical action primitive could be "get stethoscope", "listen to heart at point x_1", "listen to heart at point x_2", and "listen to heart at point x_3". Combining these medical action primitives would create "auscultation of the heart". Combining this medical action with "auscultation of the lungs" and "informing the patient" would create the medical activity "medical consultation".

In order to recognise medical actions, a dataset that contains the correct subset of medical actions is needed. Many datasets on human activities are currently available. The activities found in these datasets range from running, jogging, and robbing to checking watch, playing golf and getting out of the car (11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21). We discuss this in detail in Section 3.1.

However, to our knowledge, no datasets consisting of one-on-one interactions between GPs and their patients are publicly available. Therefore, in this work, we collect and annotate a database of medical actions, with conditions similar to real medical consultation scenarios. Challenges that arise when collecting a dataset are maintaining variation in the videos, selecting the right recording settings, and eventually annotating the data correctly. We go into more detail on these challenges in Chapter 3.3.

1.3. Research questions

For this work, we investigate within the healthcare domain, more specifically, we focus on one-on-one interactions between GPs and their patients. We would like to extract the relevant medical actions as performed by the GP. This information can then be used in the C2R project to automatically store reports in the EMR. The main research question (MRQ) that we want to address in this work is:

'Can we recognise the health practitioners' actions during a medical consultation from video recordings, using state-of-the-art human action recognition technology?'

To answer this question, we investigate what research has been done already in healthcare settings, with regard to action recognition. Moreover, we investigate which actions are relevant for reporting and decide which of these are most important to focus on.

Furthermore, we conduct research in what techniques can be utilized to recognise actions from videos and how we can recognise the different (medical) actions. We have to find out what training and testing data we need, as well as annotations that are needed in the dataset. Lastly, we have to find a decent validation method.

To make it more precise, the subquestions that we want to answer are:

- RQ1: *What research has already been done within healthcare, in combination with HAR?*
- RQ2: *Which actions during a treatment are relevant for reporting?*
- RQ3: *What kind of algorithms will be usable?*
- RQ4: *What basic techniques can be utilized to recognise medical actions?*
- RQ5: *How can we recognise different actions using HAR?*
- RQ6: *What will be the validation method?*

1.4. Outline of the thesis

In Chapter 2, we give more information on the C2R project. In Chapter 3 we first go into more detail on existing databases, then discuss how we designed our experimental setup and collected Video2Report (V2R), our dataset. In Chapter 4, we describe the machine learning preliminaries, such as neural networks (NNs), gradient descent, backpropagation, classification, and convolutional neural networks (CNNs). Then, in Chapter 5 we describe our method and approach to collecting V2R, and training and testing our classifiers. In Chapter 6, we go into more detail on how we

collected V2R and on the dataset itself. In Chapter 7, we discuss our machine learning approach with regards to training and testing our classifiers. Subsequently, we describe our experiments and results in Chapter 8. In Chapter 9 we discuss the limitations of V2R and our machine learning approach. Finally, in Chapter 10 we conclude our work by answering our research questions.

2. Care2Report

In this Section, we provide more information on the C2R project, which forms the context for the work performed in this thesis. This project[1] is conducted by a research team of Utrecht University, under the supervision of prof. dr. S. Brinkkemper[2] .

Administration tasks in healthcare require over 100,000 full-time positions in long-term care in the Netherlands. The total cost exceeds 5 billion euros per year[3] . In the United Stated, it sums up to about 13.5% of their time, for a total value of around 15.5 billion dollars (22).

The administrative costs have increased rather than decreased with the introduction of the EMR (1). This problem is experienced in most healthcare disciplines, e.g. general practice, trauma surgery, medical specialty, and home care (1; 23; 24; 25). This includes both recording and maintaining the patient medical information in the EMR. This administrative load is experienced as a burden by the CPs and causes their job satisfaction to drop. Two out of three CPs indicate that the administrative burden is too high (24). Issues that arise are data inaccessibility, a difficult user interface, and an overload of information (26).

Automated medical reporting via an innovative integration of state-of-the-art speech and action recognition is the goal of the C2R research program. In order to make the C2R device a solution, rather than more work, C2R aims for the following eight goals, which are also listed in Table 2.1.

Firstly, using the C2R device should not interfere with any of the current working procedures of the CPs (G_1). In other words, it must be an easy to use device, that does not require extra steps by the CPs. Secondly, the input of all medical devices (these are called the 'modalities') should be easy (G_2). So all modalities should be

[1] See the project website http://www.care2report.nl/
[2] http://www.cs.uu.nl/staff/sjaak.html
[3] https://www.berenschot.nl/actueel/2016/juli/administratieve-taken/

able to be easily connected to the device. Thirdly, the report will be made in real time (G_3) and should give a complete and concise summary of the consultation (G_4). After the report has been made, the CP must be able to check it and, if necessary, edit the report (G_5) before it is uploaded to the EMR. The system will be able to learn from the adjustments made by the CP, such that the need for these adjustments will decrease overtime (G_6). Furthermore, the device must be applicable as widely as possible, such that it can be used in multiple domains within healthcare, e.g. general practice, home care, and specialists in hospitals (G_7). Lastly, privacy has a crucial role in the project (G_8). The rights and responsibilities as laid out in the General Data Protection Regulation (27) are taken into account in the entire C2R project.

G_1	No interference with current working procedures.
G_2	Simple input control of all modalities.
G_3	Report generation in real time.
G_4	Complete and concise summary of consultations.
G_5	Care provider must check and possibly edit report.
G_6	System learns from edits by care provider.
G_7	Applicable for multiple healthcare disciplines and languages.
G_8	Compliant with privacy regulations.

Table 2.1: Design principles of the C2R system (28).

Many research challenges arise in the scope of the C2R program. Our research focuses on automatically recognizing the medical actions that need to be reported and that are performed by the GP in medical consultations. Therefore, when discussing the architecture of C2R, we only discuss the parts that are required for action recognition, rather than the entire architecture.

2.1. Functional architecture

Figure 2.1, shows the functional architecture of the C2R system. The input to the system are the audio and video of the consultation, as well as information of the medical devices that are used. These medical devices are also called the domotics, i.e. the medical Bluetooth instruments, as found in the MySignals kit (29). These three

inputs are processed individually, and serve as a support for each other as well. The videos are preprocessed and then given to the action recognition pipeline. The output of our algorithm will be the medical actions that have been performed by the GP. The output of speech recognition can be used as an extra reinforcement that a certain action (most likely) will be performed. The medical action as recognised by our action recognition pipeline will serve as a reinforcement that certain medical instruments (domotics) are used by the GP, represented in the Measurement Aggregator.

The information in the Measurement Aggregator combined with the textual dialogue will then serve as the Consultation Interpreter. This will then be transformed into the correct information and format for the EMR. More information on the entire functional architecture can be found in (28).

2.2. Technical architecture

In figure 2.2, the technical architecture of the system is shown. It contains the linguistic software components of the system. On the left, the domotics are represented. Once used, these instruments contain valuable medical information, that are stored in the Client. The Client also stores information from the Audio and Video interface. Here, the patient data for the EMR are stored temporally, such that reports can be read back and adjusted until it is stored in the EMR.

The audio and video information is controlled in the Server Cluster, or more precisely, in the Server Controller that interacts with the Client. The Microanalyzer Controller in the Server Controller controls all analysis processes and invokes the needed microanalyzer. In the Microanalyzers we find the audio and video preprocessing steps and the information gained from the domotics, of which the dependencies are shown with blue arrows. The audio and video input are preprocessed, such that the valuable information in the form of triples can be extracted. Combined with the output of the domotics, we can select the valuable information and then the reports can be generated, which can be send back to the Client via the Server Controller. Details of the entire technical architecture can be found in (28).

Our research falls entirely within the Microanalyzer part. The videos are pre-processed such that they all have the same resolution and frame rate in frames per second (fps). Then we extract the skeleton joints with the use of OpenPose, of which the working discussed in Section 4.3. This representation is further processed into a meaningful mathematical representation, as be discussed in Section 4.4, and then fed to a classifier, which will be trained to recognise the medical actions that are performed by the GP. We go into more detail of this last step in Section 4.1.

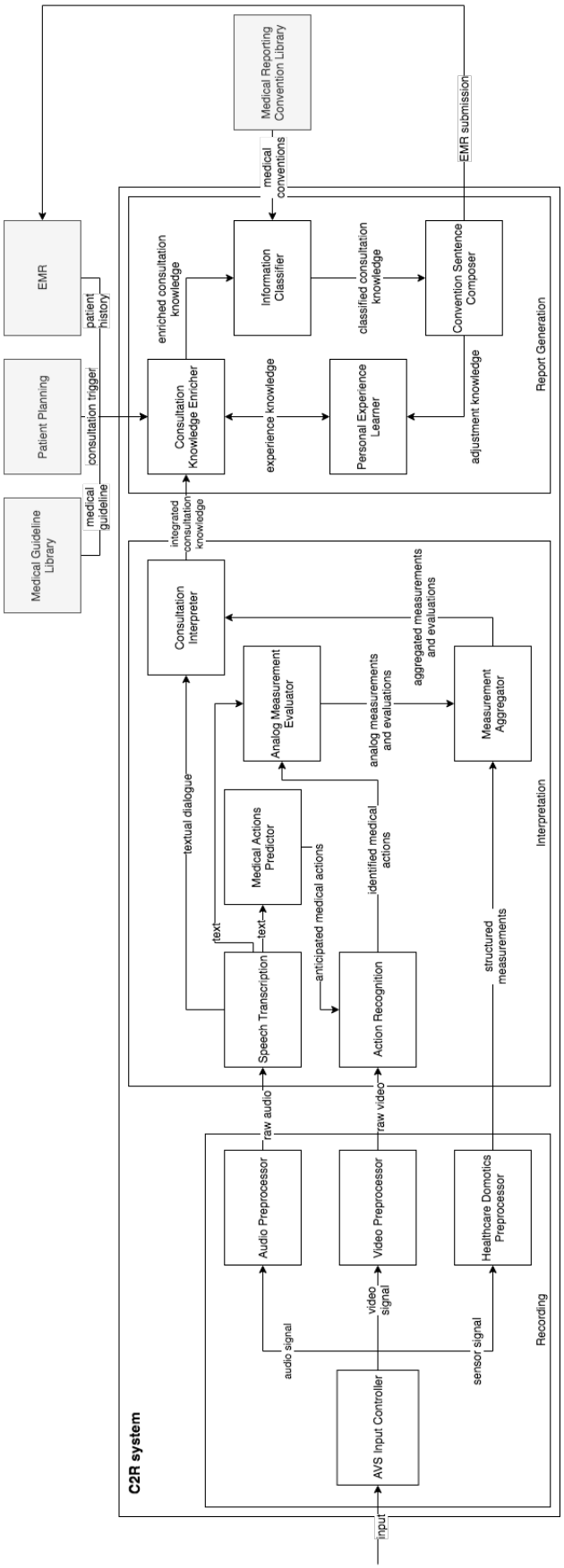


Figure 2.1: Functional architecture of the C2R system with components based on microservices, from (28)

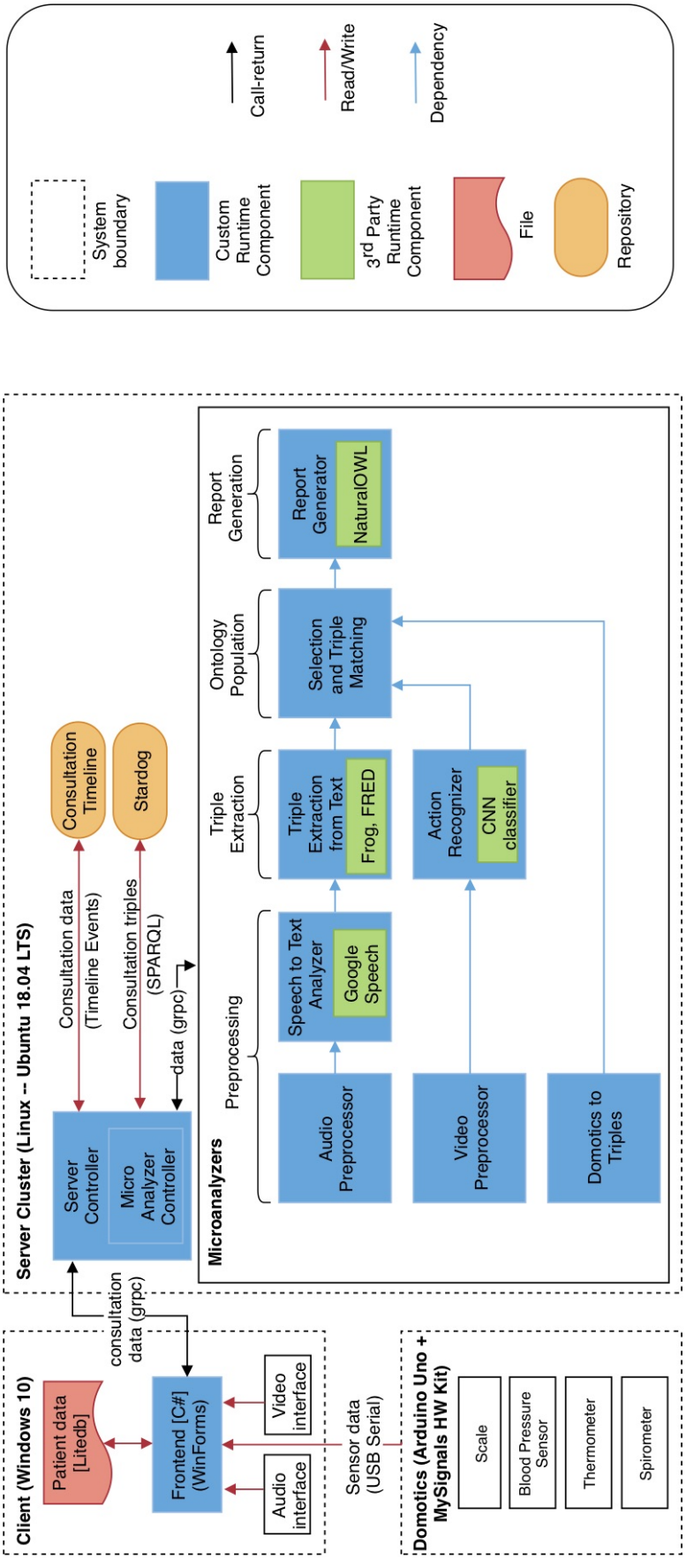


Figure 2.2: Technical architecture of the C2R system, from (28).

3. Activity Recognition datasets

In this Chapter, we first give an overview of publicly available dataset. The aim is to get insight in how they are structured, rather than providing a complete overview of all available datasets. Subsequently, we provide more insights in the challenges that arise when we want to recognise actions from videos. Lastly, we provide details on our own dataset.

3.1. Background in datasets

Over the years, the publicly available datasets have become more diverse. The first datasets consisted of simple actions like walking and running, whereas nowadays datasets consist of over 400 different human actions. Moreover, these datasets are less controlled and recording settings are more realistic (4). Even though many datasets are publicly available, to the extent of our knowledge, no such datasets are available on medical consultations between general health practitioners and their patients. Subsequently, we collect an acted doctor-patient interaction dataset in a way that the videos can be shared with researchers.

In order to do this, we first provide an overview of publicly available and relevant datasets. We do not aim to create a complete coverage of all available datasets, but rather to provide an overview and an inspiration to collect our own dataset. Important factors are the number of actors in the videos and the number of actions in the datasets. Furthermore, the viewpoint and number of cameras being used for recording is important for creating our own dataset.

The first publicly available datasets consisted of only a few actions. These actions were simple single-person actions, like walking, running, and clapping hands (30; 31; 32; 33; 34; 35), and simple two-person actions, like fighting and meeting (36; 37; 38; 39) or a combination of single- and multiple person action (40; 41; 42; 43; 44; 45). These actions became increasingly harder, involving multiple persons in an action, like playing

sports (46; 47).

In addition, the amount of classes increased. Where the former datasets consisted of 3-20 classes, this increased to more than hundreds of classes (48; 49; 50; 51; 52; 53; 54; 55; 56; 57), ranging from cooking, to sporting and householding activities. Naturally, these datasets consisted of both single and multiple person action classes.

The first datasets had static backgrounds and were recorded with a single static camera, since these actions consisted of actions that could be recorded while the actors stayed in place, like jumping and clapping (30; 31; 36; 37). While over time, these actions were recorded with multiple (calibrated) cameras (32; 40; 41; 42) and ((34; 47)) or recorded with a dynamic and cluttered background (33). Moreover, the source of the videos in the datasets also evolved over time. From recording own datasets, this evolved in retrieving videos from other sources, like movies or TV shows (38; 44; 45; 54) or from YouTube and Google (48; 49; 51; 53; 55; 56; 57; 58), in which the cameras viewpoints can both be static and dynamic. Naturally, these last datasets, as well as the datasets recorded with multiple cameras, consists of multiple viewpoints, rather than a single viewpoint. An overview of publicly available datasets can be found in Table 3.1

	Dataset name, [Ref]	D.O.P.	#subj	#act	Viewpoint, (#cam)	#videos	Fps	Resolution	Duration (s)
1	CAVIAR (36; 59)	May, '04	NA	9	Static (1 and 2)	60	25	384 x 288	20-56
2	KTH (11; 30)	Aug., '04	25	6	Static (1)	600	25	160 x 120	8-59
3	BEHAVE (37; 60)	Oct., '04	125[4]	10	Static (1)	4	25	640 x 480	768
4	ETISEO (40; 61; 62)	May, '05	NA	15	Static (4)	86	NA	Variable	NA
5	Weizmann (20; 31)	Oct., '05	9	10	Static (1)	90	50	180 x 144	1-3
6	IXMAS (41)	Nov.-Dec., '06	11	13	Static (5)	1,800	23	390 x 291	1-5
7	CASIA (42)	Jun., '07	24	7	Static (3)	1,446	25	320 x 240	5-30
8	Hollywood-1(HOHA) (16; 44)	Jun., '08	NA	8	Dynamic (NA)	700	25	640 x 480	30-240
9	UCF-Sports (15; 46)	Jun. 2008	NA	10	Static (1)	150	10	720 x 480	2-14
10a	UIUC-1 (43)	Oct., '08	8	14	Static (1)	532	15	400px	NA
10b	UIUC-2 (43)	Oct., '08	5	5	Static (1)	3	15	60X80	109, 204 & 262
11	Hollywood-2 (17; 45)	Jun. '09	-	12	Dynamic (NA)	3,669	25	640 x 480	30-240
12	MSR Action (33; 63)	Jun. '09	10	3	Static (1)	63	15	320 x 240	32-76
13	i3D Post Multi-View (34; 64)	Nov. '09	8	13	Static (8)	104	25	1920 x 1080	NA
14	UCF-ARG (32)	Jul. '10	12	10	Static (3)	1,440	60	1920 x 1080	7200-10800
15	UT Interaction (39)	Aug. '10	NA	6	Static (1)	20	30	720 x 480	˜ 60
16	UT Tower (35)	Aug. '10	6	9	Static (1)	108	10	360 x 240	NA

[4]A total of 125 persons were marked with bounding boxes in the dataset. The authors do not specify wether these are 125 individuals, or that a single individual occurs multiple times in the videos (60)

Table 3.1 continued from previous page

17	MuHAVi (14; 47)	Aug.-Sept., '10	14	17	Static (8)	1904	25	720 x 576[5]	NA
18	TV Human Interaction (38)	Aug.-Sept., '10	NA	4	Dynamic (NA)	300	-	Variable	30-600 frames
19	HMDB51 (58; 65)	Nov., '11	NA	51	Dynamic (NA)	6,766	30	Variable[6]	>1
20	MPII Cooking (50; 66)	Jun., '12	12	65	Static	44	29.4	1624 x 1224	Total: 28800
21	UCF50 (48; 67)	Sept., '12	NA	50	Static and Dynamic (NA)	6,676	Variable	Variable	NA
22	UCF 101 (49; 68)	Nov., '12	NA	101	Dynamic (NA)	13,320	Variable	Variable	NA
23	Sports 1M (51; 69)	Jun., '14	NA	487	Dynamic (NA)	1,133,158	Variable	Variable	avg. 336
24	ActivityNet (52; 70)	Jun., '15	NA	203	Dynamic (NA)	~ 27,800	Variable	Variable	NA
25	THUMOS'15 (53; 71)	Jun., '15	NA	101	Dynamic (NA)	>23,700	Variable	Variable	1-250
26	Hollywood-Charades (54)	Oct., '16	267	157	Dynamic (NA)	9848	Variable	Variable	avg. 30
27	Kinetics 400 (55; 72; 73)	May, '17	NA	400	Dynamic (NA)	~ 300,000	Variable	Variable	NA
28	Kinetics 600 (56; 74)	Aug., '18	NA	600	Dynamic (NA)	~ 500,000	Variable	Variable	NA
29	Kinetics 700 (57; 75)	Jul., '19	NA	700	Dynamic (NA)	~ 650,000	Variable	Variable	NA

Table 3.1: An overview of datasets and their statistics.

[5] 7x: 720 x 576, 1x: 704 x 576
[6] 240 x unknown

3.2. Collecting dataset

To the extent of our knowledge, there is no publicly available (annotated) database on medical encounters between general health practitioners and their patients. Therefore, we present our own dataset. The sessions consist of one-on-one encounters between GPs and their patients. In order to best represent a real consultation, we use existing clinical guidelines for Dutch health practitioners. These are available online at the website of the Dutch Health Practitioners Society ('Nederlandse Huisartsen Genootschap')[7] .

From these guidelines, we can find the typical medical actions and treatments for ninety syndromes. A selection of these actions and treatments are used to record one-on-one interactions. We utilise a publicly available video annotation tool, ELAN, to annotate the data (76). We annotate the medical actions, as well as the posture of the patient, contact between GP and their patient, and the area in which the GP performs a medical action, i.e. the area of investigation.

3.3. Challenges in collecting a database

To extract the correct and meaningful features from the input videos, we have to understand what the dataset consists of and what challenges we face. We describe these in the following Sections.

3.3.1. Inter- and intra-class variation

Within a class of actions, variances occur in the performance of an action. For instance, a GP may start palpating the abdomen on the left side of the body, as well as on the right side. And she/he may move clockwise, or counterclockwise, or address the area of the abdomen in a random order. These differences occur for the movements of a GP compared to movements of other GPs. This is referred to as intra-class variation. Conversely, inter-class variation refers to differences between actions of different types

[7] https://www.nhg.org/nhg-standaarden

(i.e. different action classes). Notably, the classes palpating abdomen and auscultating lungs are easier to distinguish than distinguishing auscultating lungs from auscultating heart. A good action recognition algorithm should be able to account for these intra- and inter-class variations. This becomes a bigger challenge as the total number of classes increases (4).

3.3.2. Environment and recording settings

While filming the videos, the environment has an influence on the variation in the recordings. It is harder to classify actions when the environment is dynamic rather than static. Moreover, lighting conditions also play an important role in difficulty of the dataset. Daylight has a different effect on the videos than indoor lighting and indoor lighting can be influenced by outdoor lighting if there are windows in the room. Furthermore, occlusions, both partly or a person/object as a whole, can make a dataset more challenging to correctly classify.

The use of different angles for the cameras for the same action class, might cause a different representation and therefore the classification algorithm has to be flexible. On the other hand, by using multiple cameras simultaneously and from different angles, occlusions can be alleviated. However, in order to do so, the exact location, height, and distance of the cameras must be known (4).

In figure 3.1, three images of the same recording, at the same moment in time, are shown. Even though the images are from the exact same moment, there are differences in lighting conditions, camera angle, and zooming options. The left figure shows the recordings of a camera that is slightly higher in the air, creating a bird's eye view. The lighting is warm, compared to the lighting created by the GoPro, as shown in the middle. The GoPro has an 170°angle camera and is zoomed out, creating an overview of the entire setting. Lastly, the iPad, shown on the left, is more zoomed in and also has a warmer lighting conditions. The zooming causes for occlusions in the videos, as the GP may walk around the patient and disappear from sight.

Figure 3.1: Images from the same session at the same moment in time, that were captured by three different cameras. Left: camera, middle: GoPro, right: iPad

3.3.3. Temporal variations

It is important to note that variations may occur with regard to the length of an action, as well as the rate at which it has been recorded. These play an important role with regard to the temporal extent of an action (4). If the duration of an action is 10 seconds and it has been recorded at a frame rate of 30 fps, then the action is represented by 300 frames, whereas if it were recorded at 20 fps, this same action would be represented by 200 frames.

In V2R, the duration of the medical actions differ. The average duration of the singular medical actions are listed in Table 6.4, e.g. blood pressure measurements takes on average 1 min 37 seconds, whereas palpation of the abdomen takes 34 seconds. The movements that are performed by the GP during the medical treatments are discussed in detail in Section 6.3.

3.3.4. Obtaining and labeling training data

An annotation task is the labeling of segments of the data. It gives information about occurrences in the frames of the videos, and provides it with the correct information, e.g. a label on 'Posture Patient' can be either laying down or sitting. We represent this as a binary label.

Annotating the videos can be done either automatically or manually. When annotating the set manually, it is important to note that actions might be perceived differently across the annotators. Therefore, it is important to make sound and measurable

agreements on the annotations, such that that they have high inter-rater reliability.. Agreements on annotations are discussed in Section 6.6.

We use the online ELAN tool (76) for the annotations. It allows us to annotate multiple videos simultaneously and to annotate several occurrences in the same file, as shown in Figure 3.2. In a single file, the posture of the patient, the area of investigation, and the corresponding medical action are annotated. After annotating the videos, we can represent the annotations mathematically. We describe this more elaborately in Section 6.6. By annotating these occurrences, we can ensure that we can start with recognizing simple movements, like the posture of the patient, as well as recognizing more complex actions like a medical action.

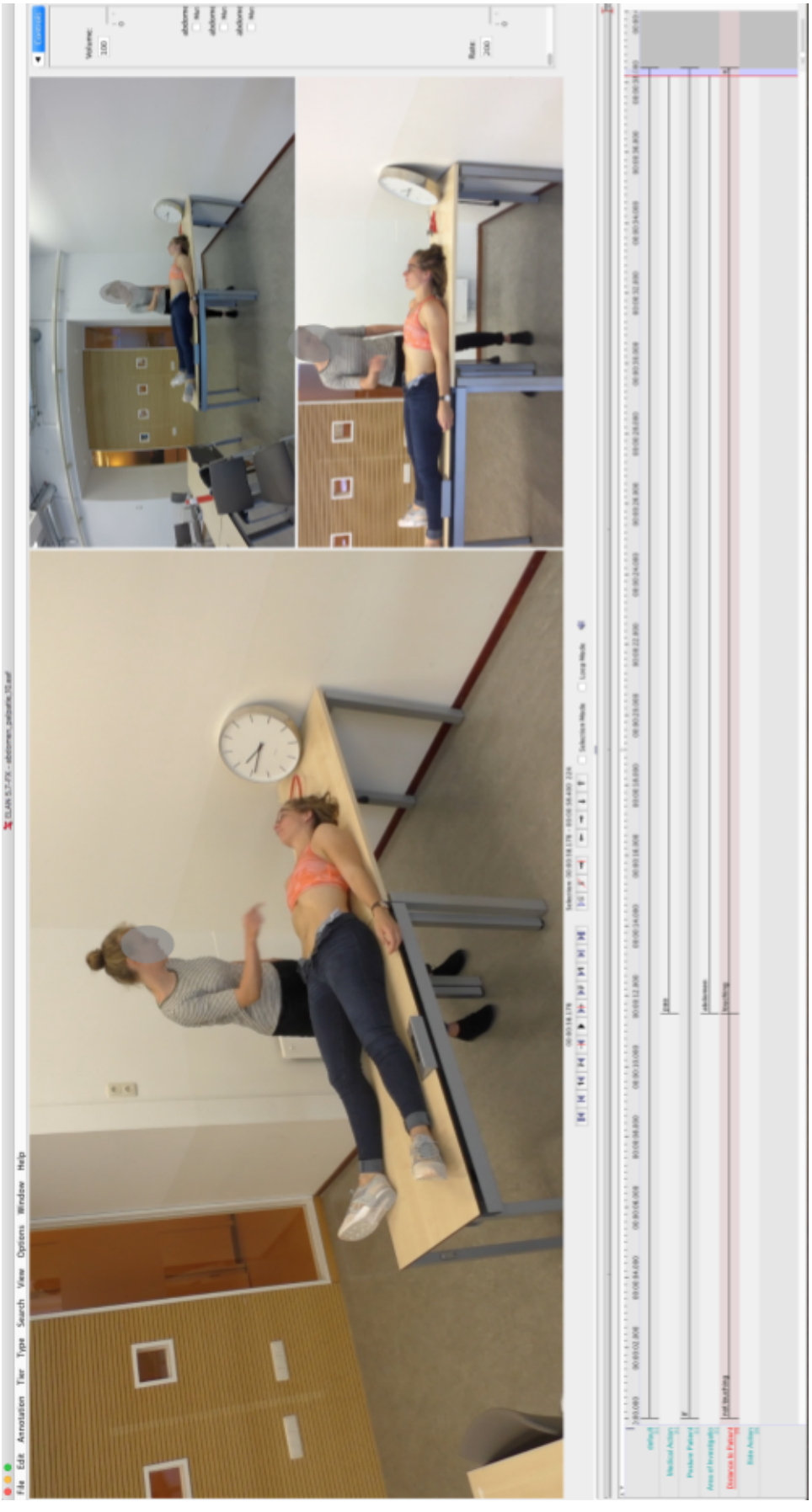


Figure 3.2: Annotation mode of the ELAN tool

3.3.5. Segmenting videos

Many of the algorithms that have been developed over the years, have been trained on segmented videos, i.e. videos with action boundary annotations. Thereby, the action detection part could be ignored, since it was done manually. However, for online applications such as recognising and identifying medical actions on the spot as found in this thesis, action detection is important. The aim of action detection, also called action spotting, is to detect when and where an action begins in the video.

3.3.6. Distinguishing doctor from patient

In order to recognise medical actions being performed in the sessions, it is important to know which person is the GP and which person is the patient. However, in the Netherlands, GPs wear regular clothes, rather than white coats, therefore, GPs can not be distinguished by their clothing.

In most situations, the GP enters his/her office in the morning, without patients. From that moment, we can detect the GP and she/he can then be tracked throughout the day. In (77), the authors propose a method to locate and identify persons in videos, by using multiple videos and by matching faces with the colors of their clothing. Moreover, the position of the persons in the room might be an indicator as well. The GP can be sitting behind the computer, whereas the patient will not, and the patient may be laying or sitting on the bed, whereas the GP will not.

3.3.7. Privacy

Privacy is of great importance for the C2R project. While certain actions, such as undressing, mostly happen behind a closed curtain, the intimacy of some medical actions may also be inappropriate to film. Moreover, patients may feel uncomfortable being recorded during their consultation, since they might be discussing private issues. A solution to this might be to have a device with two buttons. A red button for turning the camera off and a green button for recording with the camera.

Note that the storage of the videos lasts until the report is stored in the EMR. Since the aim is to save the reports in real time, the videos will be stored for a short time only and deleted afterwards. It is important that this is known by the patients as well, for this may influence their sense of feeling safe and secure.

3.4. Details of Video2Report

Our collected dataset consists of one-on-one interactions between GPs and their patients. A total of four subjects acted as patient, of which three also acted as GP. The distribution of GP/Patient is shown in Figure 3.3. We recorded V2R with at most three different camera's. In Figure 3.4, an overview of the three different cameras is shown.

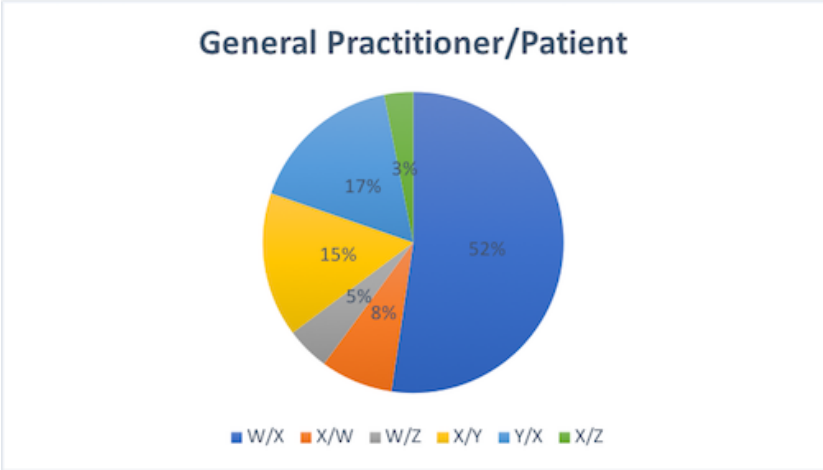


Figure 3.3: The distribution of GP/Patient for subjects W, X, Y, and Z. Subjects W, X, and Z are female, while subject Y is male.

Figure 3.4: Setup of the cameras while recording.

A total of six medical actions were recorded in V2R. These are blood pressure measurement, auscultation of the heart, the lungs, and the abdomen, percussion of the abdomen, and palpation of the abdomen. A detailed description of the medical actions can be found in Section 6.3. We annotated the posture of the patient, the distance from either the hands of the GP or the medical instrument to the patient, the area of the patients body where the GP performs a medical action, and lastly of the medical actions that is being performed. More details can be found in Section 6.6. Lastly, we divide V2R into a training, a validation, and a test set. Details can be found in Section 6.9.

4. Machine learning preliminaries

In this Chapter we describe more details on the machine learning aspects. First, we discuss the classification step and several classifiers in Section 4.1. Second, we describe neural networks (NNs) in Section 4.2, and while doing so we describe the workings of gradient descent and backpropagation as well. Third, we discuss the workings of convolutional neural networks (CNNs) in Section 4.3. Lastly, we describe how we extract useful features to represent our videos in Section 4.4.

4.1. Classification

A classifier is a mathematical function that maps data to a certain class. We do this by means of machine learning, as we do not want to manually fine tune the parameters of the classifier. The classifier is first trained on a dataset and then it is tested for the accuracy on a set that it has not been trained on. The algorithm, i.e. the classifier, is trained to identify to which classes the data belongs, e.g. auscultation of the lungs or auscultation of the heart.

Common algorithms for classification are decision trees (DTs), k-nearest neighbors (k-nns), random forests (RFs), and NNs (78). Because we have a limited amount of samples, we did not train end-to-end deep NN classifiers, but used simpler machine learning algorithms, i.e. a k-nn, a DT, and an RF classifier.

K-Nearest Neighbor. A nonparametric approach is used when no assumptions can be made with regard to the input density and it can be used to estimate the class-conditional densities, $p(\boldsymbol{x}|C_i)$. The kernel density estimator of the class-conditional density $\hat{p}$ for N_i instances belonging to class i, and $N_i = \sum_t r_i^t$, is given in Eq. 4.1, where r_i^t is 1 if $\boldsymbol{x}^t \in C_i$ and 0 otherwise.

$$\hat{p}(\boldsymbol{x}|C_i) = \frac{1}{N_i h^d} \sum_{t=1}^{N} K(\frac{\boldsymbol{x} - \boldsymbol{x}^t}{h}) r_i^t \tag{4.1}$$

The maximum likelihood estimation of the prior density can be written as $\hat{P}(C_i) = \frac{N_i}{N}$ and the discrimimant, g_i, is represented in Eq. 4.2:

$$g_i(\boldsymbol{x}) = \hat{p}(\boldsymbol{x}|C_i)\hat{P}(C_i) = \frac{1}{Nh^d} \sum_{t=1}^{N} K(\frac{\boldsymbol{x} - \boldsymbol{x}^t}{h}) r_i^t \tag{4.2}$$

and $\boldsymbol{x}$ will be assigned the class label for which $g_i(\boldsymbol{x})$ takes the maximum value. For the k-nn classifier, the following equation holds, Eq. 4.3:

$$\hat{P}(C_i|\boldsymbol{x}) = \frac{\hat{p}(\boldsymbol{x}|C_i)\hat{P}(C_i)}{\hat{p}(\boldsymbol{x})} = \frac{k_i}{k} \tag{4.3}$$

A k-nn classifier assigns class labels to its input, by checking the examples surrounding the input value and assigning it the label of the most occurring label of its neighbors. The algorithm checks the labels of the k nearest neighbors of the input value, and assigns the label of the most occurring label amongst these k neighbors. To avoid ties, k is normally set as an odd number. The k-nn algorithm is intuitively and simple and therefore easy to interpret. However, since it needs to calculate all distances for every iteration, it is time consuming for more complex input data.

Decision Tree. A DT classifier consists of a nodes, branches, and leaf nodes. In the nodes, the input value is tested with some test function $f_n(x)$, which has a discrete outcome, corresponding to one of its branches. At each node n in the DT, starting at the root, the data is tested with its test function $f_n(x)$, until a leaf node is reached. The algorithm is greedy, and in every node, it searches for the best split. The leaf node corresponds to an output value, i.e. a class. Figure 4.1 shows a simplified example of a dataset and a DT corresponding to it.

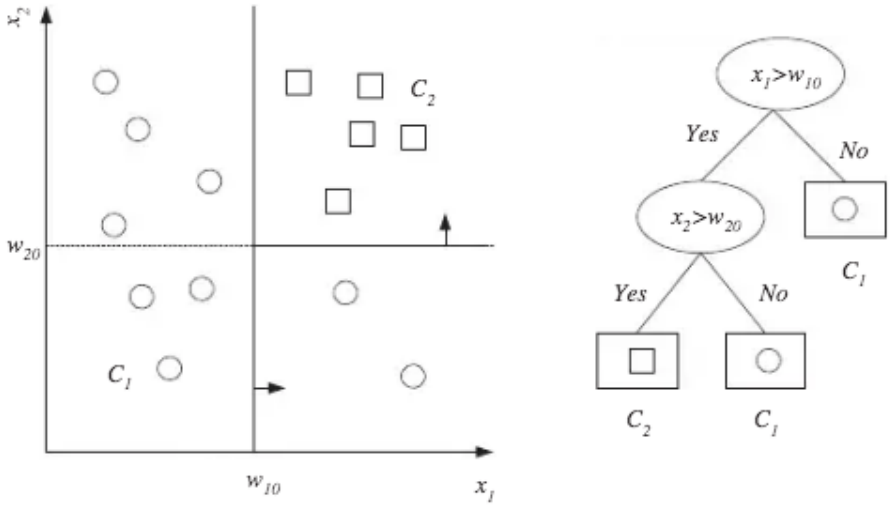


Figure 4.1: Simple scheme of a decision tree, image from (78).

Each $f_n(x)$ in the nodes divides the input space into smaller regions and represents a discriminant in the d-dimensional input space. The boundaries of the regions are defined in the discriminants.

The goodness of the split can be calculated by an impurity measure. A split is considered to be pure, if all branches after the split, consists of instances belonging to the same class. Consider node n, and the amount of training instances A_n reaching it. A_n^i of A_n belong to class C_i, with $\sum_i A_n^i = A_n$. If an instance reaches node n, we can calculate the probability $\hat{P}(C_i)$ of that instance belonging to class C_i with Eq. 4.4:

$$\hat{P}(C_i|\boldsymbol{x}, n) \equiv p_n^i = \frac{N_n^i}{N_n} \tag{4.4}$$

Node n is pure if $\forall i, p_n^i = 0, 1$, i.e. it is 0 if none of the instances in the branch belongs to class C_i and 1 is they all belong to it. If it is pure, we have reached the leaf node, and can assign it a class label. A common function to measure the purity is entropy (79), Eq. 4.5:

$$I_m = -\sum_{i=1}^{K} p_m^i \log_2 p_m^i \tag{4.5}$$

with $0 \log 0 \equiv 0$.

One of the advantages of DTs is that due to the hierarchical structure, the DT allows for a fast search of regions. However, DTs tend to overfit on the input data if there are many classes and relatively little training data.

Random Forest. An RF is an ensemble of multiple DTs. While training an RF, one is training multiple DTs simultaneously, each on a different subset of the input set. The predictions of the individual trees are all combined and the average of the predictions are calculated. By doing so, the overall accuracy of the algorithm can be increased, compared to DTs. By taking the average of multiple DTs combined, the tendency of overfitting of the DT is reduced.

4.2. Neural Networks

Convolutional neural networks (NNs) can be used to extract information from videos to correctly classify them into one of finitely many classes. A CNN is a special form of neural network (NN), in which the nodes in a layer are all connected to the nodes in the consecutive layer, creating a fully connected NN. The output is a numerical representation of the video, also known as a vector. Every layer in the network performs their own set of calculations, by which corners and circular objects can be recognised

and also object shapes. All nodes in the layers have certain weights. This weight determines how much that specific node contributes in that layer and this can be trained using back propagation. A NN is a multi-layered network of neurons that can be used to correctly classify and predict certain classes. Figure 4.2 shows a NN with an input layer, with four input features, one hidden layer, with five neurons, and an output layer with one neuron.

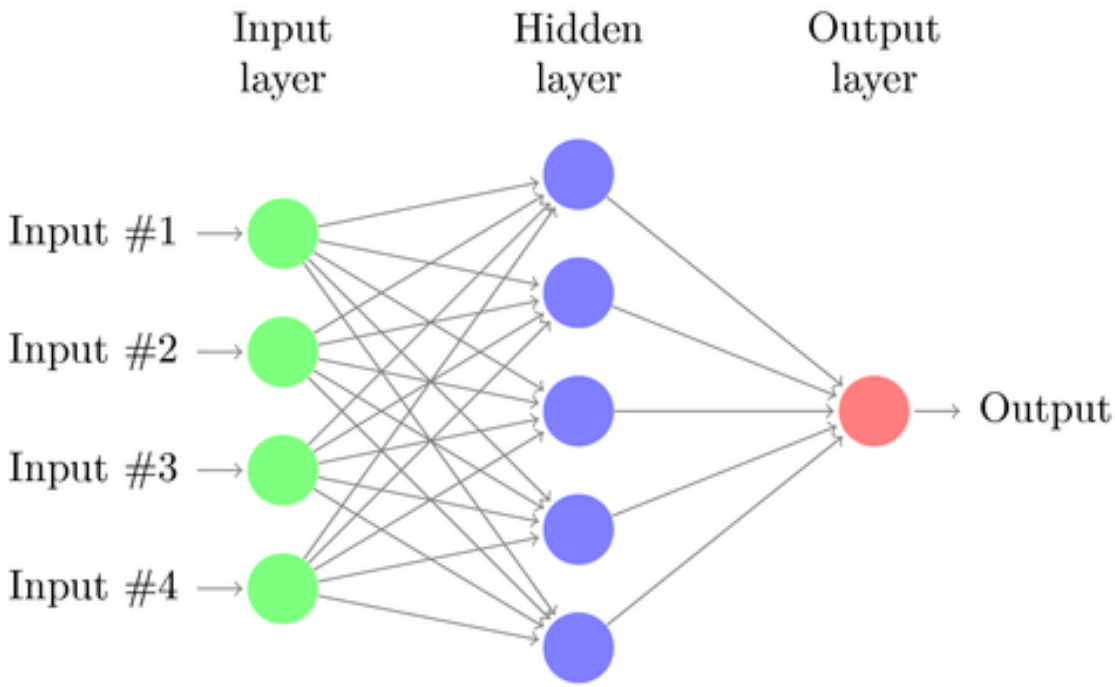


Figure 4.2: A display of a NN with four input features, one hidden layer, and an output layer, from (80).

The neurons j in hidden layer i get their input x_k from its penultimate layer $i-1$. The neuron performs a calculation and the output serves as the input for the neurons in the consecutive layer that it is connected to. The connection between neuron k in hidden layer $i-1$ to neuron j in hidden layer i contains a weight w_{ikj}. This weight is used by neuron j to multiply with the input x_k from neuron k. The output y_{ij} of neuron j in hidden layer i is some activation function ϕ over the summation of the multiplication between the input from neuron k and its corresponding weight w_{ikj}, as represented in Eq. 4.6. The bias of a hidden layer can be added to the summation by means of an extra neuron $x_0 = 1$ and the bias $w_{ik0} = bias$ for hidden layer i .

$$y_{ij} = \phi(\sum_{k=0}^{n} w_{ikj} x_k) \tag{4.6}$$

There exist several activation functions ϕ, whose main goal is to map the resulting values for the neurons within a certain range, e.g. between 0 to 1, or -1 to 1, or ≥ 0. The activation function adds non-linearity to the output, such that the performance of the classifier can be improved. The most commonly used activation functions for NNs are the Sigmoid (81), the ReLU (Rectivied Linear Unit) (82), and the Tanh function. The Sigmoid function maps its input to a value between 0 and 1, as shown in Eq. 4.7:

$$\omega(x) = \frac{1}{1 + e^{-x}} \tag{4.7}$$

The Sigmoid activation function has clear bounds, making it easy to use in the neurons. Moreover, the Sigmoid function tends to map the inputs towards either 0 or 1, making the neuron more discriminatory. The Tanh activation function is similar to the Sigmoid function, however maps its input to a value between -1 and 1, making the neuron even more discriminative towards its input, as shown in Eq. 4.8:

$$\omega(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4.8}$$

Lastly, the ReLU function maps the output as the input value, if the input value is positive, otherwise it maps the input to 0, as shown in Eq. 4.9:

$$\omega(x) = \left\{ \begin{array}{r} if x \geq 0 : x \\ else : 0 \end{array} \right\} \tag{4.9}$$

Once the amount of hidden layers, neurons in these hidden layers, and activation functions are chosen for the NN, the network needs to be trained to predict the output as close to the ground truth as possible. This is done by a process called backpropagation.

To start the backpropagation process, random weights and biases are assigned to the network. With these settings, the network makes a prediction for the given input, which is compared to the expected result. The error, i.e. the loss, is calculated with the loss function, and backpropagation aims to minimize the loss by adjusting the weights of the connections between the neurons.

Several loss functions can be used in the process, and a commonly used loss function is the Mean Square Error (MSE), in which the predicted output θ is subtracted from the exact output $\hat{\theta}$ and the error is squared, resulting in $Loss(\hat{\theta}, \theta) = [\hat{\theta} - \theta]^2$. Since we need to calculate the derivative during backpropagation, we want to make it easier to calculate the derivative. Therefore, this error is usually multiplied by 0.5, resulting in $Loss(\hat{\theta}, \theta) = 0.5(\hat{\theta} - \theta)^2$. By squaring the difference, outliers have a great influence in the loss function, which might influence the backpropagation algorithm strongly. Another loss function may therefore be to calculate the absolute difference between the predicted output θ and exact output $\hat{\theta}$, resulting in $Loss(\hat{\theta}, \theta) = [\hat{\theta} - \theta]$. Even though this reduces the influence of outliers, this function is non-differentiable, making it harder to calculate the derivative. The error is averaged over all n examples in the training batch, resulting in the error (E): $E = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2$ or $E = \frac{1}{n}\sum_{i=1}^{n}[\hat{\theta}_i - \theta_i]$, depending on the chosen error function. Then the backpropagtion algorithm calculates which weights and biases it needs to adjust, to minimize the error. It does so by means of gradient descent, in which the direction of the minimal loss is calculated.

Gradient descent calculates the derivative of the loss function, in order to find the impact of the changes in the network. Backpropagation starts at the end of the network, i.e. the output, and calculates how changes in the previous layers influence the output. Therefore, we calculate the derivative of the error function E with regards to the changed weights w in hidden layer i from neuron k in layer $i-1$ to neuron j in layer i, as shown in Eq. 4.10:

$$\frac{\partial E}{\partial w_{ikj}} = \frac{\partial E}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial net_{ij}} \frac{\partial net_{ij}}{\partial w_{ikj}} \tag{4.10}$$

In which $\frac{\partial E}{\partial y_{ij}}$ represents the impact of neuron j in layer i, so for predicted output y_{ij} this is $\frac{\partial E}{\partial y_{ij}} = Loss(\hat{\theta}, y_{ij})$, and this loss is back propagated to the network. Therefore, we need the activation function for layer i, neuron j: $\frac{\partial y_{ij}}{\partial net_{ij}}$. Since all neurons are connected and therefore influence each other, we need to propagate to all inputs of the neuron with $\frac{\partial net_{ij}}{\partial w_{ikj}}$. We repeat this process for all output neurons, which creates a list of impact for each weight in the network. Since we calculated the derivatives, this list, i.e. a gradient vector $\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, ..., \frac{\partial E}{\partial w_d}\right]^T$, gives us the direction to change the weights, for the best result on the loss function. The gradient descent procedure updates the weights in the opposite direction of the gradient, as shown in Eq. 4.11:

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}, \forall i \tag{4.11}$$

$$w_i = w_i + \Delta w_i \tag{4.12}$$

In which η is the learning rate of the backpropagation algorithm, usually $0 \leq \eta \leq 1$. Once the derivative reaches 0, the procedure is terminated. Since the learning rate influences the speed of convergence, it is important to find a good value η, as for a large η the algorithms may miss the trends of the input, while for a small η, the convergence may be too slow (78), which results in an increase of training time. A solution may be to start with a high learning rate, which decreases over time. Thereby, the results improve rapidly in the beginning, and gets more fine tuned towards the end of the algorithm.

4.3. Convolutional Neural Network based action recognition

A convolutional neural network (CNN) is a type of NN, with specific special layers. One such specific layer is the convolutional layer, in which a rectangular filter is placed over the input array and slided over it. This layer has a certain activation function, as depicted in Figure 4.3. With these simple calculations, lower level features, e.g. lines, can be detected. These features can then serve as input for the next convolutional layer, which can then detect higher level features, e.g. curves and edges.

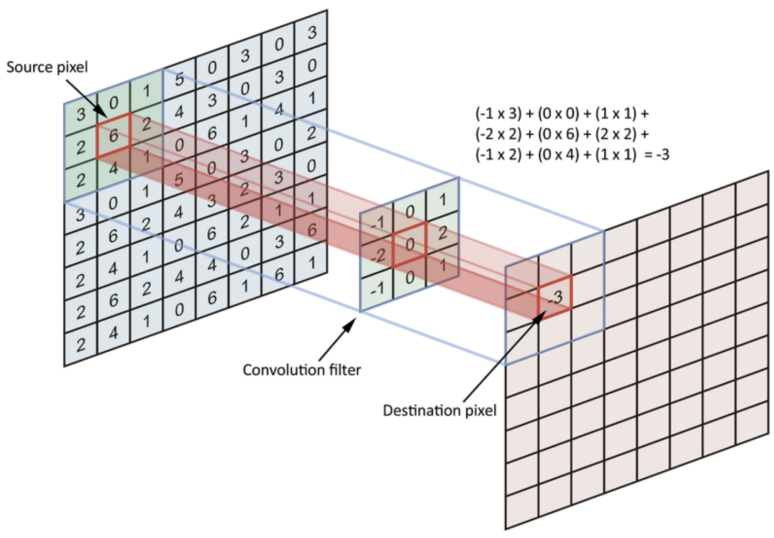


Figure 4.3: The activation operation within the convolutional layer, image from (80).

Every filter in the convolutional layer adds a layer to the 2D input array, creating a 3D filter space. Combined with a convolutional layer, we can add another special layer, namely a pooling layer. This is used to reduce the features created in the

convolutional layer. A commonly used pooling layer is the max pooling layer, in which the maximum value in a certain area is saved as the value representing that specific area. Other pooling options are minimum and average pooling, which take the minimum or average value over a specific area respectively. The pooling layer assist in locating features that are invariant to rotation (83). A CNN is a special type of NN with convolutional and pooling layers, as depicted in Figure 4.4. The final layer is a fully connected classification layer, whose goal is to provide a prediction with regard to the input. It provides us with an n-dimensional array of predictions, in which n is the amount of possible labels, indicating whether a certain object, e.g. a person, is in the input.

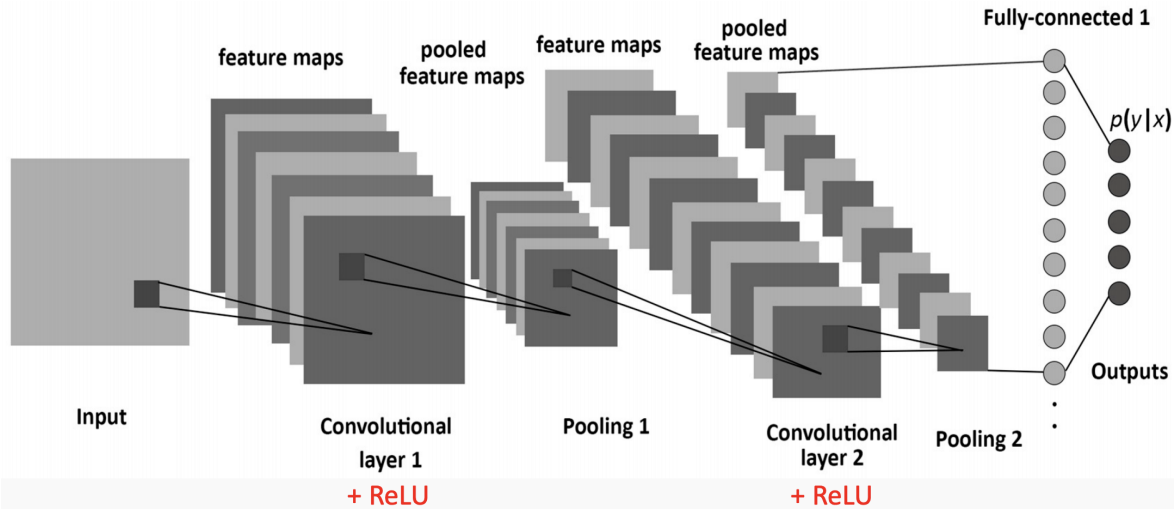


Figure 4.4: Example of a CNN with two convolutional layers, followed by a pooling layer, and a final, fully connected, classification layer, image from (80).

CNNs are a form of deep learning, as they consist of multiple layers. Since every layer performs its own set of calculations, thereby allowing to detect and recognise higher level features, e.g. persons and medical instruments, adding more layers to the network, allows for the network to recognise increasingly difficult features. However, by adding more layers to the network, more training data is needed to train the network correctly, thereby also increasing the needed time and computing power to train the network.

4.4. Video representation using a trained CNN

We need to represent our input mathematically, such that we can train and test our classifier. While doing so, we have to take into account the challenges as described in Chapter 3.3. This representation should be able to generalise over inter-class variability, and distinguish between intra-class variability. Thereby, it should be able to overcome differences in backgrounds, viewpoints, lighting conditions, and the appearance of individuals.

An open-source and trained CNN can be used to extract this valuable information. OpenPose (84) is such a trained CNN, and is available for scientific purposes. OpenPose has been trained on the COCO (85) and MPII (86) datasets and is trained to identify persons in images and videos. Since it is pretrained on a big dataset, it is invariant to background, camera viewpoints, and lighting conditions. After detecting the persons in the videos, it generates the keypoints of these persons. These keypoints are the location of the joints and limbs, e.g. the position of the neck, the hands, and the hip, an example is shown in Figure 4.5.

The coordinates of the skeletons in itself may not contain valuable informtaion, since the exact same movement recorded with the camera results in entirely different coordinate points compared to the recordings of the GoPro, and a shift of the camera may also result in entirely different coordinate points. However, from the coordinates of these keypoints, we can subtract valuable information, such as the distances from one person to the other, which can be used to train and test our classifiers.

The keypoints of the persons in the video represents the skeleton of these persons. In the work of (87), depth sensors are used to create a complex human activity dataset, consisting of two person interactions. Skeleton features are used to calculate the distance between all pairs of joints, and using these as the representation yields the best outcome of their Support Vector Machine (SVM) classifier.

In the work of (88), the authors use a regularized deep Long Short-Term Memory

Figure 4.5: Extracted Keypoints using OpenPose.

(LSTM) network to recognize action from the 3D skeletons of the persons in the videos. They use the co-occurence of joints that can characterize a human action, i.e. the joints that are moved simultaneously while humans perform certain actions. They show that the dropout algorithm ensures effective learning of the LSTM neurons.

In (89), the authors propose an approach using joint angles from three dimensional (3D) skeleton features to recognise human actions with a linear SVM. From the 3D skeleton joints, spatial features and spatio-temporal features are extracted. The algorithm was tested on the MSR-Action 3D Dataset ((90)), the MSR-Hand Gesture Dataset ((91)), and the UCF-Kinect Dataset ((92) and achieved state-of-the-art results).

Skeleton data was used to extract spatial and temporal features in (93). In their work, a Recurrent NN with Long Short-Term Memory was build. The model was trained to recognise which joints were dominant in certain actions, using a temporal attention module.

In (94), the authors extract geometric relations amongst all joints from the 3D skeletons obtained from the dataset. The LSTM achitecture is able to capture the long-term dependencies. From the 3D skeletons, both lines between two joints, as well as planes between three joints were used to extract the geometric relations. Experiments on the SBU-Kinect dataset (87), NTU-RGB+D dataset (95), Berkeley MHAD dataset (96), and the UT-Kinect dataset (97) show that the distance between joints combined

with selected lines achieve the best results amongst all the tested features, resulting in stat-of-the-art results on these datasets.

For our dataset, we use the Euclidean distances between joints, as described in (87; 94), as well as the angle between two joints (94). We describe this more thoroughly in Section 7.2.

5. A method for Human Medical Action Recognition

In Figure 5.1, our approach is shown. The processes are further specified in Table 5.1 and the deliverables of the product delivery diagram (PDD) are specified in Table 5.2.

Firstly, research is conducted with regards to the clinical guidelines of the GPs. In these guidelines, we find the relevant medical actions as can be found at the GP's office. Secondly, a discussion with a medical student takes place, in order to find out in which sequence multiple medical actions can occur in a single session, these actions will be recorded in a session. A session can consist of a single medical action or of a sequence of medical actions, i.e. auscultation of both the heart and the lungs. We record multiple sessions in a row, and edit the videos into singular sessions. Thirdly, we conduct research on how to place the cameras, such that the best viewpoint is created. We use three different (offline) cameras, namely a Panasonic HC-V770, a GoPro Hero 5, and an iPad.

After thoroughly researching the relevant medical actions and recording settings, we start filming. Since we film multiple sessions per recording, we cut them into singular sessions in a later stadium. After editing the sessions, we annotate them, using the tool ELAN (76). By doing so, we annotate different occurences, namely the 'posture of the patient', the 'area of investigation', whether 'the doctor touches the patient' or not, and lastly the 'medical procedure'. The posture of the patient is either laying with the legs flat, laying with the knees bend, or sitting upright. The area of investigation is either chest, upper back, abdomen, or arm. Lastly, the medical procedure is either auscultation of the heart, auscultation of the lungs, auscultation of the abdomen, percussion of the abdomen, palpation of the abdomen, blood pressure measurement, or heart rate measurement. Choices with regards to these medical actions are discussed in Section 6.2.

After annotating the dataset, we can start with the medical action detection part.

By using OpenPose we account for differences in viewpoints, changes in background and lighting conditions and we can extract the body joints of the people in the videos, we discuss this in more detail in Section 7.1. These keypoints can be used to obtain valuable information, e.g. the angle of the body of the patient, as discussed in Section 7.2. With this mathematical representation, we can create feature sets, which will be used to train our classifier, e.g. to distinguish the posture of the patient as either sitting or laying down. Experiments and results are discussed in Chapter 8.

After dividing the dataset into a training, validation, and test set, we can train the classifier with the training set. Then, we can use the validation set, to tune the parameters of the classifiers. After finding the optimal parameter settings, we can test the performance of the algorithm with the test set. It is important that none of the videos in the test set have been used in the training and validation set, as this would influence the accuracy of the algorithm.

For classification, we train and test three different classifiers, namely a k-nn, a DT, and a RF classifier. We discuss this more elaborately in Section 4.1

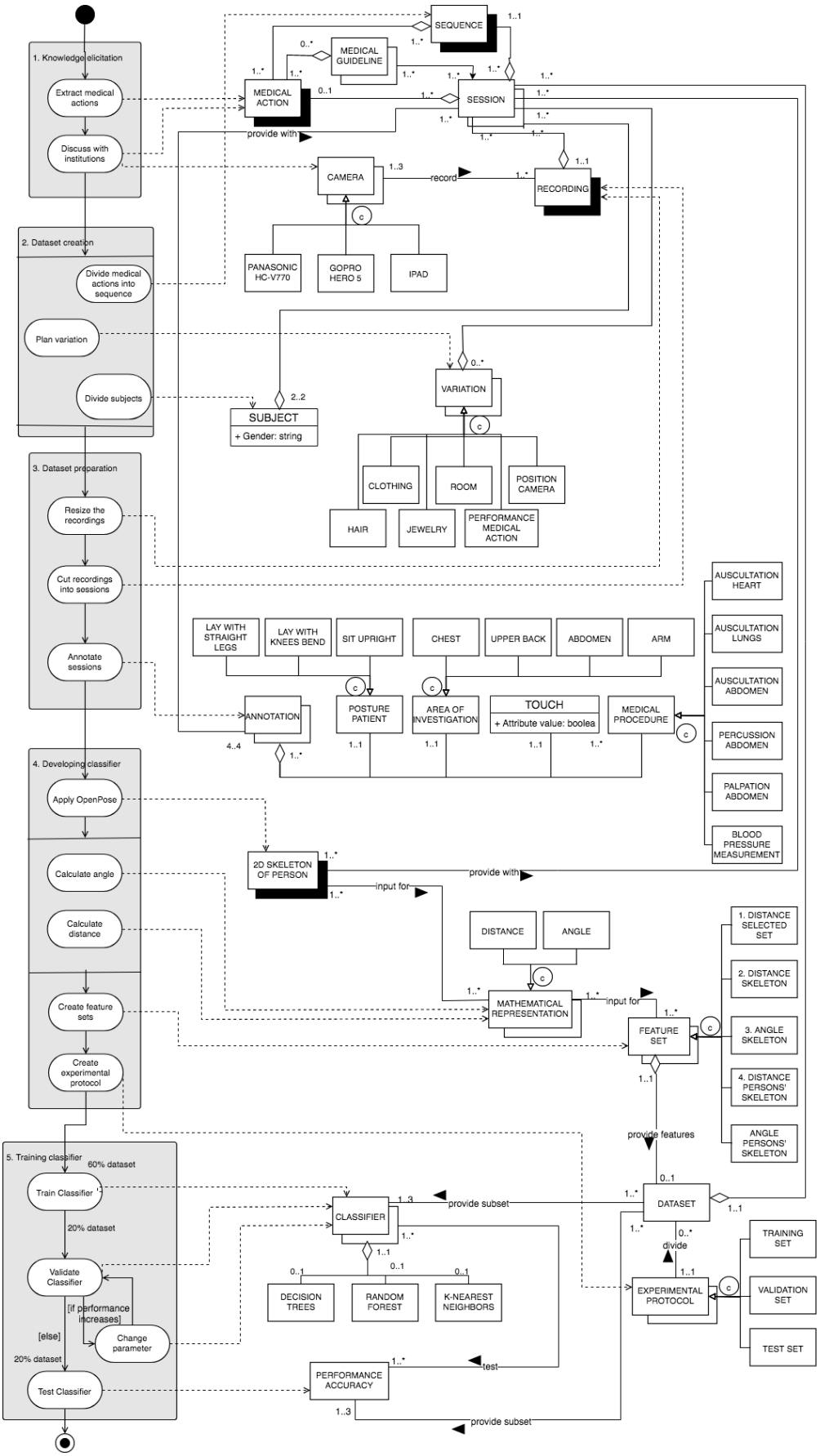


Figure 5.1: PDD that depicts the method used to create the dataset and to train and test the classifier.

Process	**Definition**
Extract medical actions	From the MEDICAL GUIDELINES[8] we extract the relevant MEDICAL ACTIONS that can be represented in our DATASET.
Discuss with institutions	The relevant MEDICAL ACTIONS are discussed with a senior medical student in order to find the SEQUENCES in which multiple MEDICAL ACTIONS can occur. At the Kinder Kennis Centrum (Child's Knowledge Centre), we obtained more information using multiple CAMERAS simultaneously.
Divide medical actions into sequences	The MEDICAL ACTIONS are divided into SEQUENCES, such that a SEQUENCE consists of one or more MEDICAL ACTIONS that are performed during a SESSION.
Plan variation	We plan for variation in our DATASET, to create more intra-class variation. Explained in VARIATION.
Divide subjects	We divided the SUBJECTS that act in our recordings, with regards to MALE and FEMALE.
Resize the recordings	We resized the RECORDINGS such that they all have the same resolution and frame rate
Cut recordings into sessions	After filming multiple SESSIONS in one RECORDING, we have to cut these RECORDINGS into separate SESSIONS, such that every SESSION consists of one medical consultation between a GP and a patient.
Annotate sessions	We make ANNOTATIONS of the SESSIONS, such that all the frames in the videos have a label, as described in ANNOTATION.
Apply OpenPose	The SESSIONS are fed into the open source CNN OpenPose. This results in a 2D SKELETON OF THE PERSON, i.e. the position of the joints of the persons in the videos.
Calculate distance	The distance between two joints from the 2D SKELETON are calculated and saved as a MATHEMATICAL REPRESENTATION
Calculate angle	The angle between two joints from the 2D SKELETON are calculated and saved as a MATHEMATICAL REPRESENTATION
Create feature sets	The MATHEMATICAL REPRESENTATION is stored in FEATURE SETS
Create experimental protocol	The DATASET is divided in TRAINING, VALIDATING, and TESTING sets, as represented in our EXPERIMENTAL PROTOCOL, more details in Section 6.9.
Train Classifier	According to our EXPERIMENTAL PROTOCOL, we use our TRAINING SET to train our CLASSIFIER. We train our CLASSIFIER, using a FEATURE SET, and the ANNOTATION of the corresponding SESSIONS.
Validate Classifier	After training the CLASSIFIER, we validate it using the VALIDATING SET, as found in the EXPERIMENTAL PROTOCOL. From the VALIDATING SET, we use the FEATURE SET and the ANNOTATION of the corresponding SESSIONS.
Change parameter	If the performance increases for certain parameter settings, we update the parameters of the CLASSIFIER accordingly.
Test Classifier	The optimal tuning as found in the validation phase is used to test the CLASSIFIER on the TESTING SET of the DATASET, as found in the EXPERIMENTAL PROTOCOL. Note, that the test set has not been previously used in the training or validation step.

Table 5.1: Definitions of the processes.

MEDICAL ACTION	These are the relevant MEDICAL ACTIONS that can be found in our DATASET. This is further specified in Section 6.3.
MEDICAL GUIDELINE	From the MEDICAL GUIDELINES [9] , we can extract the relevant MEDICAL ACTIONS.
SEQUENCE	A SEQUENCE is a combination of one or more MEDICAL ACTIONS. This is further specified in Section 6.4
SESSION	A SESSION is one medical consultation between a GP and a patient.
CAMERA	The CAMERAS are used to record the sessions. The camera is either a PANASONIC HC-V770, a GOPRO HERO 5, or an IPAD.
RECORDING	A RECORDING contains multiple SESSIONS as recorded with one of the CAMERAS.
SUBJECT	The SUBJECTS are the actors in the videos. A SUBJECT is either a MALE or a FEMALE person.
VARIATION	The VARIATION in our dataset is created by using multiple CLOTHING and HAIR styles and the use of JEWELRY per SUBJECT. VARIATION is also created by changing ROOMs for the SESSIONS, by the POSITIONING of the CAMERAS, and by the PERFORMANCE of the MEDICAL ACTIONS.
ANNOTATION	An ANNOTATION is a labeling of a SESSION. There are four different annotations, namely the POSTURE OF THE PATIENT, the AREA OF INVESTIGATION, the MEDICAL PROCEDURE, and the DISTANCE TO THE PATIENT. See also Section 6.6
POSTURE PATIENT	The annotated POSTURE OF THE PATIENT can be either LAYING WITH STRAIGHT LEGS, LAYING WITH THE KNEES BEND, or SITTING.
AREA OF INVESTIGATION	The AREA OF INVESTIGATION is the part of the patients' body that the GP examines or where the MEDICAL ACTION is being performed. The area of investigation can be either the CHEST, the UPPER BACK, the ABDOMEN, or the left or right ARM of the patient.
TOUCH	The ANNOTATION 'TOUCH' shows whether the doctor touches the patient on a specific part of the body or not, with either the hands or a medical instrument,.
MEDICAL PROCEDURE	The MEDICAL PROCEDURE that the GP performs can be either AUSCULTATION OF THE HEART, the LUNGS, or the ABDOMEN, PERCUSSION or PALPATION OF THE ABDOMEN, BLOOD, or PRESSURE MEASUREMENT. The MEDICAL ACTIONS are further specified in Section 6.3.
2D SKELETON PER PERSON	OpenPose (98) generates the 2D SKELETON OF THE PERSONS in the videos. This represents the joints of the body, e.g. wrists, elbows, and shoulders.
MATHEMATICAL REPRESENTATION	The MATHEMATICAL REPRESENTATION is obtained by a calculation with the 2D SKELETON OF THE PERSONS in the video. This can either represent the DISTANCE between two joints, or the ANGLE of these joints. This is further discussed in Section 7.2
FEATURE SET	The FEATURE SET consists of a combination of the MATHEMATICAL REPRESENTATION. It either represents the 1. DISTANCE for a SELECTED SET, 2. DISTANCE of two joints of the SKELETON, 3. ANGLE of two joints of SKELETON, 4. DISTANCE between two PERSONS' SKELETON, or 5. ANGLE between to joints of two PERSONS' SKELETON and is explained in more detail in Section 4.4
DATASET	Our created DATASET consists of all the SESSIONS with their corresponding ANNOTATIONS.
EXPERIMENTAL PROTOCOL	We separate our DATASET according to our EXPERIMENTAL PROTOCOL in a TRAINING, VALIDATING, and TESTING set.
CLASSIFIER	A CLASSIFIER is able to learn to predict a class, given some datapoints. The classifiers that we use are a DECISION TREE, a RANDOM FOREST and a K-NEAREST NEIGHBOR CLASSIFIER. These are specified in Section 4.1
DECISION TREE	Explained in Section 4.1
RANDOM FOREST	Explained in Section 4.1
K-NEAREST NEIGHBOR	Explained in Section 4.1
PERFORMANCE ACCURACY	After training and validating the CLASSIFIERS, it is tested on the TEST SET. During the test phase, the CLASSIFIER predicts the classes of the input data. These predictions are then compared to the ANNOTATIONS, that is considered the ground truth.

Table 5.2: Definitions of the deliverables.

6. Collecting the dataset

We aim to collect and annotate a dataset of one-on-one medical encounters, which will be used as a training and testing set for our action recognition pipeline. We start by conducting research at Nivel and NHG. Nivel is the Dutch Institute for research in healthcare (99). The institute researches the effectiveness and quality of healthcare in the Netherlands and the relationships between healthcare providers, consumers, and insurers, and the government. NHG is the 'Nederlands Huisartsen Genootschap' (Dutch General Practitioners Society)[10] . Their aim is to facilitate scientifically justified professional practices by the GPs of the Netherlands. They have clinical guidelines for various diseases and physical symptoms. The guidelines provide a basis for identification of the relevant, and most occurring, medical actions to represent in V2R, as characterized in Section 6.2. Furthermore, Section 6.3 addresses the consultation of a senior medical student in the final phase of her Master education on the medical actions, to assure a representative collection. From here on we will refer to her as the medical student. After recording the relevant medical actions, they are be processed and annotated. Finally, we have an annotated dataset that consists of one-on-one medical consultations between GPs and their patients.

6.1. Gaining insights at NIVEL

To fully understand the type of actions that are performed in healthcare, we examine the dataset of NIVEL (99). NIVEL has agreed upon a cooperation with C2R and therefore, under certain (privacy) conditions, their dataset is available for research in the scope of C2R. NIVEL has recorded data from healthcare providers in the Netherlands, which can be used to gain a better insight in what kind of medical care and treatments are provided. Examining their dataset increases our insights in the difference of treatments, e.g. how they can be recognized, with regards to movements and medical instruments that are used.

[10]https://www.nhg.org/nhg-standaarden

The dataset consists of clinical encounters between GPs and their patients. After examing the dataset, we found a lot of occlusion in the videos and a lot of medical actions are performed outside the view of the camera. Therefore, this dataset is not well suited for using it to train and test the action recognition model. Moreover, the dataset is only available at NIVEL, under privacy constrictions, and no data is allowed to leave the premises, making it less flexible to use for our work.

Even though the dataset of NIVEL is not usable for training and testing, we gained insights in the different movements of the treatments. One treatment may consist of different medical actions and the movements of the medical actions may vary between health practitioners.

6.2. Selecting medical actions from the medical guidelines of NHG

The Dutch GPs have clinical guidelines available online, for ninety diseases[11] . The structure of the guidelines can be found in Figure 6.1. In the clinical guidelines, we can find the diagnostic guidelines and the guidelines policy. The diagnostic guidelines are specified in anamnesis, physical examination, additional research, and evaluation. In the physical examination, information can be found on the medical actions that GPs are supposed to perform, e.g. blood pressure measurement, auscultation of heart and lungs, etc.

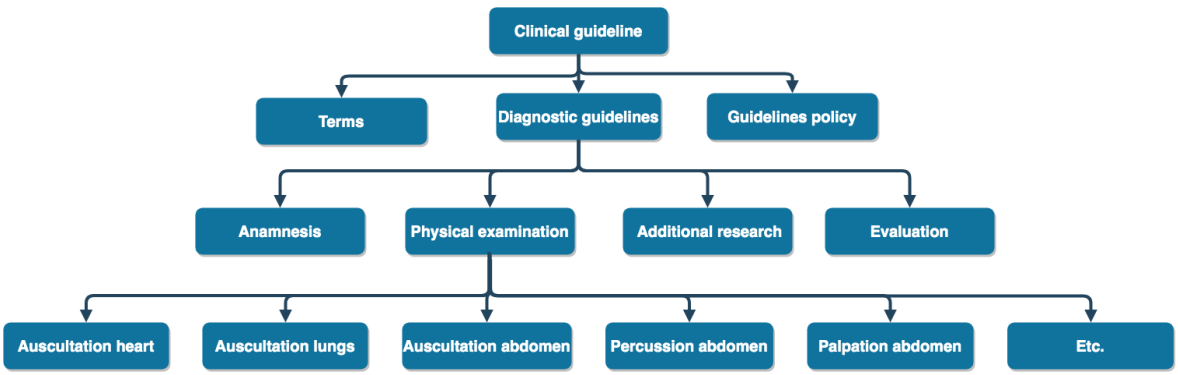


Figure 6.1: Structure of clinical guidelines at NHG.

We have examined the medical actions that occur in the clinical guidelines. While

[11]https://www.nhg.org/nhg-standaarden

looking into these actions, we have to take into account that some of the actions performed by the GPs are quite intimate and private. For privacy reasons, these actions will most likely not be filmed during actual consultations. Therefore, we have eliminated the examinations and treatments that we consider too private to film, e.g. examinations in the genital area. Furthermore, some of the actions consist of inspection (with the eyes) by the GP. This medical observation is not an active movement by the doctor. For this stage of the C2R program, these actions are too advanced to identify. Therefore, we leave these actions out of the relevant medical actions as well.

After manually eliminating the guidelines that contain only medical actions that are either too intimate to record or are not 'active' medical treatments, we end up with sixty one guidelines that contain medical actions that we can record. In Table 6.1, we listed the medical actions, and their percentage of occurence in the guidelines, that can be used to record in V2R. The percentage of occurrences in these guidelines have been manually obtained. Measuring the blood pressure occurs most frequently, namely in 44% of the 61 guidelines. Palpation of the abdomen is the runner up, with 28%, followed by heart rate measurement in 25% of the 61 guidelines. Note that some of the guidelines consist of multiple medical actions, e.g. the clinical guideline for 'acute coronary syndrome' consists of heart rate measurement, blood pressure measurement, auscultation of both the heart and the lungs, and palpation of the chest. In Section 6.4, we go into more detail of what medical actions may follow each other in a sequence during a consultation.

For accurate collection of the dataset, medical instruments that GPs normally use are indispensable. For this purpose, C2R has a MySignals kit, (29), at her disposal. The MySignals kit consists of Bluetooth instruments, of which the output can be stored on the computer. To use these instruments, they have to be installed and run on a special program. Unfortunately, the MySignals kit that is at our disposal, does not contain all the instruments that are needed for the medical actions as listed in Table 6.1. For instance, a screening audiometer (for checking the hearing) is not available,

Medical action	**Occurrence**
Bloodpressure measurement	44%
Palpation abdomen	28%
Heartrate measurement	25%
Percussion abdomen	21%
Body mass index	20%
Temperature measurement	18%
Auscultation lungs	16%
Auscultation heart	15%
Testing eyes	15%
Palpation other areas	13%
Testing ears	11%
ECG	7%
Auscultation abdomen	3%

Table 6.1: Percentage of actions in the 61 guidelines after eliminating guidelines that do not contain usable medical actions.

nor is an electrical cardiogram monitor [12] ('ECG') or a thermometer[13] . Therefore, physical examinations and actions that involve these instruments are not taken into account for the collection of V2R. At the time of writing this thesis, not all of the instruments have been installed yet. However, we decided to use the blood pressure monitor. Therefore, in order to still use it for the creation of V2R, when we use this instrument, we put the cuff around the arm of the patient, wait for about 40-60 seconds, since this is the time it normally takes to automatically pump and release the air in the cuff, and then remove it, without it actually being turned on. This approach is considered to be representative for real life usage.

For the creation of V2R, we have asked the medical student for assistance. She agreed to help and to contribute to the videos by acting both as a GP and as a patient. By asking her to act as a GP, we gain more insights in the movements of a real GP.

[12] for assessment of electric cardial acitivity

[13] GPs use a special ear thermometer: https://www.vipermedical.nl/38276_genius-3-tympanic-thermometer?gclid=EAIaIQobChMI1eGXsqX-5gIVFYjVCh3HZAPXEAQYASABEgJj8_D_BwE

With the stethoscope we can perform auscultations of the heart, the lungs, and the abdomen and with the blood pressure monitor we can measure the blood pressure. Furthermore, we can perform percussion and palpation of the abdomen, since we do not need instruments for those medical actions.

After a first trial shoot, we eliminated 'heart rate measurement' as a separate action, because the action is quite subtle and often occluded from sight, since the GP takes the pulse at the wrist of the patient or during heart auscultation. Therefore, either the GP or the body of the patient may block the view. Whatsmore, heart rate measurement occurs in 15 of the 61 guidelines. Out of these 15 occurrences, it is combined with blood pressure measurement 12 times. In most blood pressure monitors, the heart rate is also measured (29). Therefore, by measuring blood pressure, we also capture the heart rate. Lastly, since 'palpation other areas' is too broad, e.g. palpation of the chest, the arm, or the legs are a few examples, we also eliminated that from possible actions in V2R. Therefore, our dataset consists of the medical actions that are listed in Table 6.2. These medical actions occur in 42 of the medical guidelines, accounting for a total 46% of all the medical guidelines. Note that in some of the guidelines, other medical actions may also occur, which we have excluded from our dataset. Therefore, a recorded session in V2R may not contain all medical actions that a GP would perform during a consultation. In the Appendix 10, Chapter C, we find an overview of the sequences of the medical actions from Table 6.1.

Medical action
1. Blood pressure measurement
2. Palpation abdomen
3. Percussion abdomen
4. Auscultation lungs
5. Auscultation heart
6. Auscultation abdomen

Table 6.2: Medical actions that are represented in V2R.

6.3. Overview of selected medical actions for V2R

After careful selection, a list of appropriate medical actions emerged, as displayed in Table 6.2. The typical duration of the movements, the order in which the medical actions are performed, and specific gestures and postures of the doctor and patient are not specified in the medical guidelines. Therefore, we consulted the medical student.

In the Sections below, we specify which medical instruments are used for the medical actions and the gestures by both the doctor and the patient. Note that each action can be performed in several ways, e.g. auscultation of the lungs can start on either the left or the right side of the body.

Blood pressure measurement. During blood pressure measurement, the blood pressure is measured using a monitor. The cuff is put around either the left or the right arm of the patient. After turning it on, the cuff is filled with air. Then, after about 30-40 seconds the air is automatically released, and the blood pressure is shown on a screen. The patient is preferably sitting upright during this medical action, or is lying down, and may not talk or move during the measurement.

Palpation of the abdomen. Palpation of the abdomen can be done in two steps. The first is the superficial palpation. The GP uses one hand to touch and press the abdomen slightly on all six regions of the abdomen. The second part is deep palpation, in which the GP places his/her dominant hand on the abdomen, and the other hand on top of the first hand. He/She then exerts more pressure on the abdomen so that he/she can feel internal structures. The abdomen is roughly divided into six regions, as shown in Figure 6.2. All six regions of the abdomen are addressed, however the order in which they are addressed depends on the complaints of the patient. The part where the pain is experienced, is usually addressed lastly. The patient lays down on their back during the palpation. If the GP can not palpate the abdomen due to active or passive increased tension of the abdominal muscles, the patient is asked to bend the

knees, since this helps to relax the abdominal muscles of the patient.

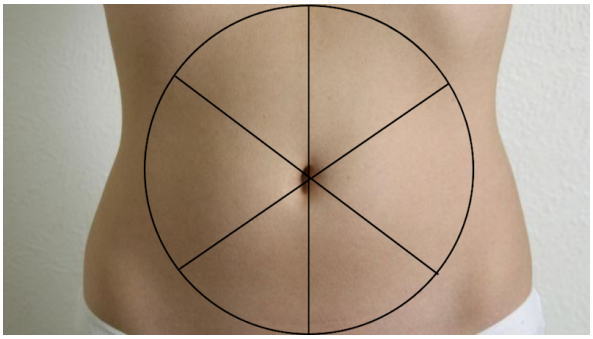


Figure 6.2: Division of the abdomen in six parts. Modified from image (100).

Percussion of the abdomen. Percussion of the abdomen is done using both hands. One hand of the GP is placed on the abdomen, with the fingers spread out and extended, applying slight tension on the skin. With the index or middle finger of the other hand, the GP ticks on the middle finger of the first hand. This way, he/she can hear a sound. The GP moves his/her hands around the abdomen and ticks in every region of the abdomen. Again, all six regions of the abdomen are addressed, as shown in Figure 6.2, and the order in which all six parts are touched differs per session, since it depends on where the pain is experienced in the abdomen. The patients lay down on their back during the percussion.

Auscultation of the lungs. Auscultation [14] of the lungs happens at approximately eight different positions and can be performed on the front and the back of the patient. These eight points are divided over the left and right side of the body. In other words, on both the left side and the right side, there are at least four listening points. Starting at the top of the lungs, the listening points are the same on both sides of the body, and are shown in Figure 6.3. So point 1a and b, followed by 2a and b, then 3a,b and 4a,b. The GP may start at point 1 on the left (a) or the right (b) side of the body. So the sequence may be either 1a-1b-2a-2b-3a-3b-4a-4b or 1b-1a-2b-2a-3b-3a-4b-4a. These listening points and sequences are the same for the back of the patient. The patient

[14] 'the process of listening to someone's breathing using a stethoscope', https://www.oxfordlearnersdictionaries.com/definition/american_english/auscultation

is laying down during auscultation on the front of the body, and sitting upright when the GP listens to the back of the lungs.

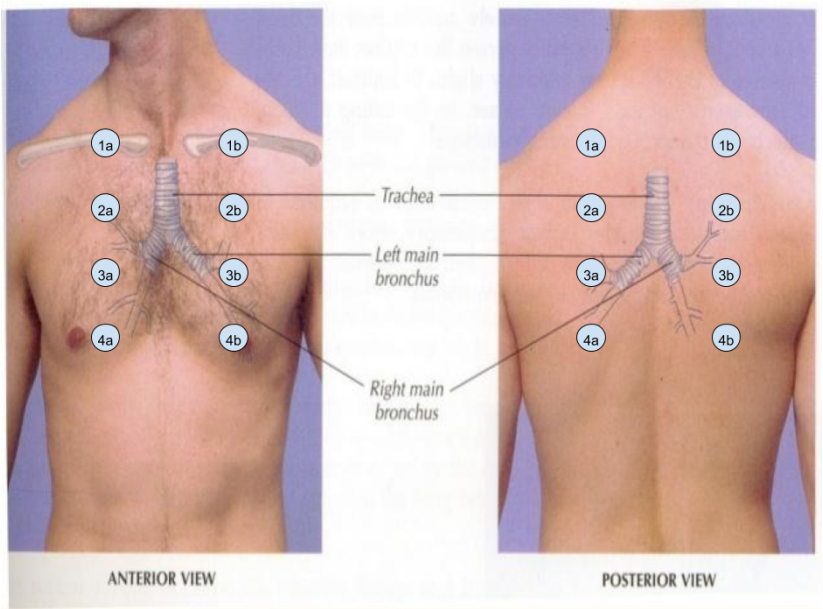


Figure 6.3: Auscultation of the lungs. Modified from image (101).

Auscultation of the heart. Auscultation of the heart is mostly similar in every session. An overview of the listening points is shown in Figure 6.4. First the doctor listens to point 1, then 2 followed by 3, 4, and 5, as shown in Figure 6.4. Depending on the clothing of the person, point 4 and 5 may vary a bit in position. The patient lays down during this medical action.

Auscultation of the abdomen. Auscultation of the abdomen consists of listening for 30 seconds directly next to the belly button. Variation lies on the exact listening spot, which can be both on the left or the right of the belly button. The patient lays down during the auscultation, either with the knees bend or with straight legs.

6.4. Sessions in our recording

One session represents one medical consultation and this may contain multiple medical actions in a sequence. Our dataset covers medical actions that occur in forty

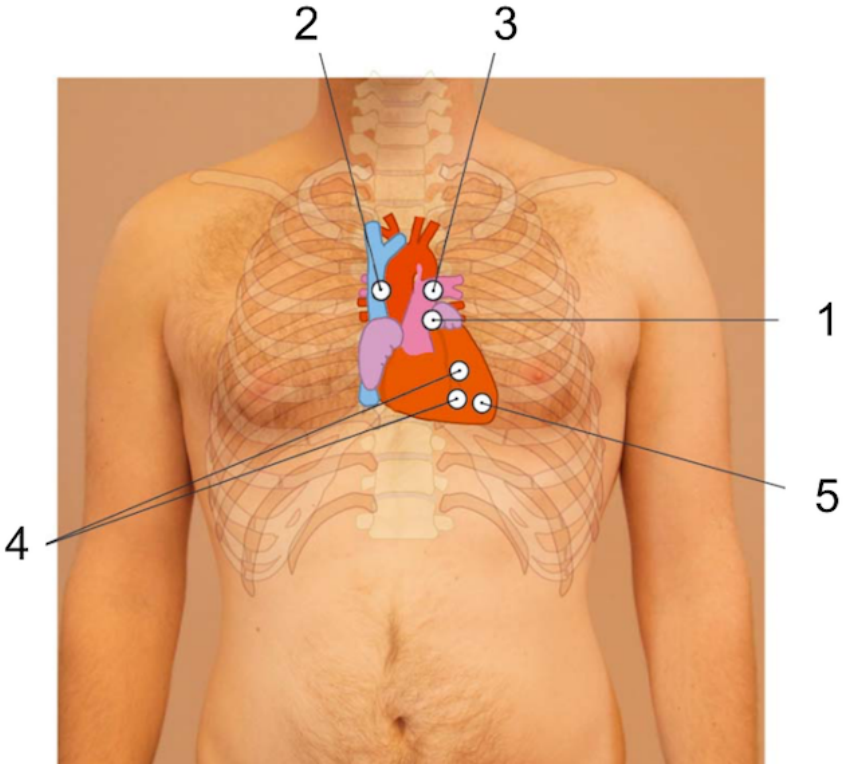


Figure 6.4: Auscultation of the heart. Modified from image (102).

two of the medical guidelines. In Table 6.3, we listed the combination of medical actions as they occur in these guidelines. Since we want the dataset to represent the normal doctor-patient interactions as accurately as possible, we included these sequences in V2R. Medical actions that are combined most often during a consultation, are auscultation of the heart and lungs, as well as auscultation, percussion, and palpation of the abdomen. In an orienting physical examination, all these five medical actions are combined in a single consultation. Moreover, measuring blood pressure is often performed together with either auscultation of the heart or the lungs. Therefore, the following sequences can be found in V2R:

(i) Blood pressure measurement

(ii) Palpation abdomen

(iii) Percussion abdomen

(iv) Auscultation lungs

- Front
- Back

# Occurrences \medical action	BPM	PaA	HRM	PeA	AL	AH	AA
11 times	X						
6 times		X		X			
5 times	X		X		X	X	
4 times					X		
4 times	X	X	X	X			
2 times	X					X	
2 times	X	X	X				
2 times	X				X	X	
1 time		X					
1 time			X				
1 time		X					X
1 time	X	X		X			
1 time		X		X			X
1 time	X	X	X	X	X	X	

Table 6.3: Total amount of occurrences per sequence in the ninety one medical guidelines, e.g. the combination Palpation Abdomen with Percussion Abdomen occurs six times in the medical guidelines. *BPM = Blood Pressure Measurement, PaA = Palpation Abdomen, HRM = Heart Rate Measurement, PeA = Percussion Abdomen, AL = Auscultation Lungs, AH = Auscultation Heart, AA = Auscultation Abdomen.*

(v) Auscultation heart

(vi) Auscultation abdomen

(vii) Blood pressure measurement and auscultation of the heart

(viii) Blood pressure measurement and auscultation of the lungs

(ix) Auscultation of heart and lungs

- Lungs front
- Lungs front and back

(x) Auscultation, percussion and palpation of abdomen

(xi) Auscultation of heart, lungs, and abdomen, percussion and palpation of abdomen

When a sequence of medical actions occurs in a session, the order in which they

are performed are always the same per session. If the entire body is inspected in an orienting physical examination, the order in which this happens is as follows: auscultation heart - auscultation lungs (if the lungs are auscultated at both the front and the back, then the front is auscultated first) - auscultation abdomen - percussion abdomen - palpation abdomen. For the sequences that consists of a subset of these five actions, this is the order in which they occur. So, during auscultation of the heart and lungs, the GP first listens to the heart and then to the lungs. Lastly, an orienting examination of the abdomen is done in the order auscultation, percussion, and then palpation of the abdomen. When the blood pressure is measured in combination with auscultation, then the measurement is performed first, followed by the auscultation of the specific region.

6.5. Recordings of the one-on-one consultations

To easily increase the amount of videos, while keeping variety, we use multiple cameras while filming. This way, we can also investigate what position of the camera is most convenient to use for C2R. At the Kinderkenniscentrum[15] , we have conducted research about the use of multiple cameras. They told us that certain software could synchronise online cameras that are linked to the computer. Unfortunately we do not have such cameras at our disposal. Instead we used three different offline cameras. These are a Panasonic HC-V770 (which we will refer to as the 'camera' from now on), a GoPro Hero 5, and an iPad. During recording, the cameras ran for multiple sessions in a row, creating recordings with multiple consultation sessions. With Adobe Pro, the recordings were cut into separate sessions.

We decided to position the cameras at different heights and in different locations. They remained at the same position for the entire shooting per day. We had a total of five recording days, and since we could not leave the cameras in the room, the height and positioning of the cameras differs per shooting day.

[15]https://www.uu.nl/organisatie/faculteit-sociale-wetenschappen/contact/ kinderkenniscentrum-utrecht

For filming, we reconstructed an approximate representation of a GPs' room, by using one of the classrooms of the university. To create an examination bank, we placed two Tables next to each other and the walls of the room were white. Note that these classrooms differ per day of shooting, but that the setup of the rooms are similar. In Figure 6.5, the field of overview of the setup is shown. Differences between our setup and a GPs' office should not matter, since we focus on the medical actions that are performed, rather than the room it is performed in.

In order to create the maximum overview with the least amount of occlusion, the camera is positioned slightly higher, such that a bird's eye view is created. The GoPro has an 170 degree angle and is positioned at eye height. The iPad is positioned in the corner at eye height.

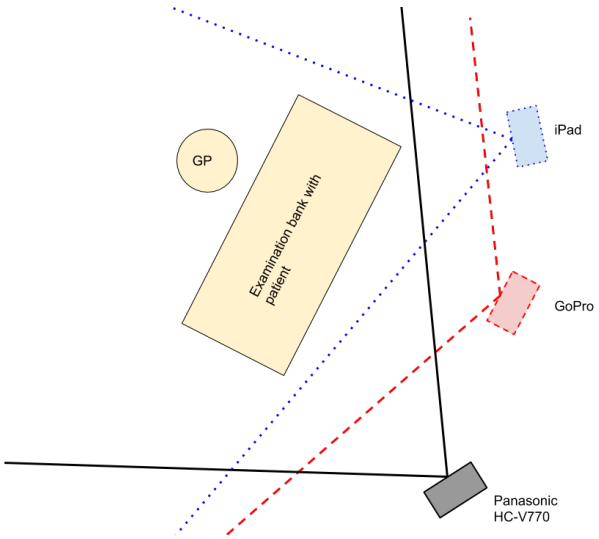


Figure 6.5: Field of View for the setup of the recording sites.

Since we made use of different cameras, the settings per camera were different. In order to get the same resolution, we edited the recordings with the Adobe Premiere Pro. We set the resolution of the recordings at 1920 x 1080 pixels, with a frame rate of 30 fps, as shown in Figure A.1 in Section 10. Furthermore, Adobe Pro has the ability to synchronise videos and then cut them into separate videos. We used this to cut the recordings into separate sessions. In the end, we shot 192 unique sessions and a total of 450 videos.

6.6. Annotation

After cutting all the recordings into separate sessions, we annotated the videos using ELAN (76), an online annotation tool. The synchronisation mode, as shown in Figure B.1, can be used to synchronise multiple videos at once, such that they can be annotated in a single file. This way, we can synchronise the sessions that were recorded simultaneously. Since some consultations were recorded with a single camera and others with two or three cameras, some of the sessions are annotated individually while others are annotated in pairs or in triples. In Table 6.4 we can find the amount of sessions and videos.

We annotated four different occurrences, as shown in Figure 3.2. Since the goal is to recognise medical actions, we annotated the medical actions as consisting of one of the six medical actions, as discussed in Section 6.3.

However, recognising medical actions can be subdivided into recognising the posture of the patient and the area of investigation. Therefore, we annotated the posture of the patient, which could be either laying down with flat legs or with the knees bent, or sitting upright. Thirdly, the area of investigation is annotated. The area of investigation can be either chest, upper back, abdomen, or arm. Lastly, we annotated the distance of the GP to the patient, which could be either touching or not touching. All these annotations are saved in a .CSV file.

The agreements on annotating the files were as follows:

(i) The medical action starts from the moment the GP touches the patient, either with her hands or with a medical instrument. It lasts until the GP no longer touches the patient with either the hands or the medical instrument.

(ii) The GP is considered to touch the patient either when the hands or a medical instrument touches the patient at the part of the body where examination takes place. For all medical actions, except blood pressure measurement, the GP is considered to touch the patient from the start until the end of the medical action,

since moving the hands or medical instruments takes only a few seconds, and the hands or instruments do not deviate from the patients' body much. While measuring the blood pressure, the GP touches the patient only while putting the cuff around the arm, or while removing it. For the duration of this measurement, the GP could be anywhere in the room, therefore, the distinction between touching or not touching during the medical action is clearer, and we consider the GP to touch the patient only when putting on the cuff and thus touching the patient with the hands and the instrument, or when removing it.

(iii) The area of investigation is the part of the patients' body where the medical action takes place, and is annotated for the entire duration of the medical action. An exception is made for blood pressure measurement, for which we annotate the area of investigation for the duration of the medical action, as well as for only when the GP is considered to touch the arm. We have annotated this as 'Arm' and 'ArmTouch' respectively.

(iv) The posture of the patient is only defined at the static moments, and not in the transition phase. Sitting upright is annotated when the patient body is vertical, while lying down is annotated when the patients' body is horizontal.

Note that activities can be static (e.g. sitting and lying), dynamic (e.g. walking), or transitional (e.g. lying to sitting) (6). We decided to annotate only the static postures of the patient. The annotations of the medical action consist of static, dynamic and transitional movements, since we annotated these actions from the moment the GP start touching the patient, until he/she is done with the medical action.

6.7. Analysis

In Table 6.4, the statistics of the sessions are listed. In total, we shot 192 unique sessions, which were recorded with either one, two, or three cameras. In total, we recorded 451 videos. Moreover, in Figure 6.6, the distribution of the sessions is shown. On the left, the distribution of medical actions is shown. Auscultation of the heart and lungs occur most often, while the distribution amongst the other four medical actions are evenly divided. In the middle, the distribution of the posture of the patient is

Action	**Unique sessions**	**Total videos**	**Average**	**Shortest**	**Longest**
Blood pressure measurement	23	69	01:37	00:43	02:40
Palpation of abdomen	18	42	00:34	00:18	01:01
Percussion abdomen	14	32	00:23	00:18	00:28
Auscultation lungs	19	46	00:50	00:28	01:23
Auscultation lungs back	22	49	00:42	00:25	01:15
Auscultation heart	21	50	00:32	00:25	00:45
Auscultation of abdomen	18	44	00:23	00:14	00:35
Blood pressure measurement and auscultation of heart	19	43	02:02	01:42	02:30
Blood pressure measurement and auscultation of lungs	6	12	02:23	02:11	02:47
Auscultation, percussion, and palpation of abdomen	7	14	01:25	01:04	01:58
Auscultation of heart and lungs (front and back)	6	12	01:50	01:27	02:30
Entire body (except blood pressure)	19	38	02:50	01:53	03:50
Subtotal	192	451			

Table 6.4: List of the created videos, including their average time, and the shortest and longest video of each session.

depicted. We see that the patient appears to be laying down most often. Lastly, on the right, the distribution of area of investigation is depicted. Most medical action appear to happen on the chest, while the abdomen is investigated least often.

6.8. Variations in the sessions

We aim to collect a dataset that best represents the natural way consultations are held. While doing so, we tried to get as much variety as possible and also make the setting as reflective of the GPs' office as possible. Since we did not have an ex-

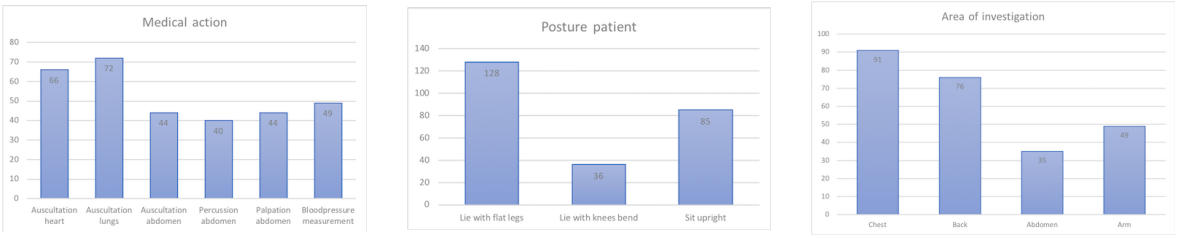


Figure 6.6: Statistics in sessions. On the left, the distribution of medical actions is shown, in the middle the posture of the patient, and on the right the area of investigation. Note that in one session, multiple medical actions, postures, and areas of investigation can occur.

amination bank at our disposal, two tables were put together. This served as a good representation of an examination bank. Furthermore, we made use of an official (red) medical stethoscope from the medical student. Other common colours of stethoscopes are blue and black. The official (red) stethoscope is used in the videos that do not contain blood pressure measuring, while the sessions with a combination of blood pressure measurements and auscultation are recorded using a black (toy) stethoscope.

In Figure 6.7, the three different views that were created for one of the sessions, at the same time frame, is shown. As can be seen from the images, the lighting conditions are different for all three different cameras. Also the angle in which the cameras was recorded is different.

Figure 6.7: Images that were captured by the three different cameras. Left: camera, middle: GoPro, right: iPad

A total of four subjects occur in the videos, and they switch positions and clothes, to account for variety in the videos. The clothes were changed at least five times per subject per day throughout the recording days, but remain the same for the duration

of a session, and we also changed our hairdo accordingly. The GP needs to wear the hair in a bun, but the patient is allowed to have the hair any way he/she likes. Also, the patient is wearing a watch in some of the videos, whereas, the doctor is not allowed to wear jewelry on his/her hands, for safety and hygienic reasons. We did make use of glasses and switched those for both the GP and the patient. Moreover, variance also occurs in the position of the patient. This can be either laying flat or with the knees bend, or sitting upright. Figure 6.8 shows variances that occur within the sessions.

During the different actions, we changed the positions of both the GP and the patient to account for variation. For instance, while listening to the lungs, the doctor would listen to the patient from different positions alongside the patients' body. Also, the patient might have her legs bent or stretched out while the GP is performing the examination. Moreover, even though the sequences are the same for multiple actions in one session, there is room for some variation per medical action. For instance, during auscultation of the lungs, the GP can start by either listening to the right or to the left side of the patients' body. This is represented in the videos.

Our dataset was recorded using multiple actors. The majority of the videos were recorded with both a female GP and doctor (131 out of 192 sessions, 68,2%), while 15,6% of the videos had a female-GP/male-patient distribution (30 out of 192 session), and 16,7% of the sessions had a male-GP/female-patient (32 out of 192).

6.9. Experimental protocol

For training and testing the classifiers, it is important to use separate subsets of the dataset. Therefore we split the dataset into 60% training, 20% validation, and 20% test set. In Table 6.5, this division is shown. The sessions that were recorded simultaneously, with different cameras, are put in the same subset. So for example, the training set for blood pressure measurement consists of videos 1-14 ('a', 'b', and 'c'), the validation set consists of videos 15-18 ('a', 'b', and 'c'), and the test set consists of videos 19-23 ('a', 'b', and 'c').

Figure 6.8: Images that were captured by the three different cameras; left: camera, middle: GoPro, right: iPad. Variance can be found in position, clothing, and hairdo of GP and patient. Moreover, the rooms in which the videos were recorded also provide variation.

Medical action	Training # sessions	Training # videos	Validation # sessions	Validation # videos	Testing # sessions	Testing # videos
Blood pressure measurement	14	42	3	12	4	15
Palpation of abdomen	10	30	3	7	5	5
Percussion abdomen	8	14	3	9	3	9
Auscultation lungs	11	28	4	12	4	6
Auscultation lungs back	13	31	4	8	5	10
Auscultation heart	13	29	4	11	4	10
Auscultation of abdomen	11	25	3	9	4	10
Palpation of abdomen	10	30	3	7	5	5
Blood pressure measurement and auscultation of heart	12	29	4	8	3	6
Blood pressure measurement and auscultation of lungs	4	8	1	2	1	2
Auscultation, percussion, and palpation of abdomen	4	8	2	4	1	2
Auscultation of heart and lungs (front and back)	4	8	1	2	1	2
Entire body (except blood pressure measurement)	11	22	4	8	4	8
Total	115	274	37	92	40	85

Table 6.5: Division of the dataset into 60% training set, 20% validation set, and 20% test set.

7. Action detection and recognition

We divided V2R in a training, a validation, and a test set. The training set is used to train the classifier which features are most valuable. In the training step, we apply supervised learning, in which we provide the algorithm with both the representation of the data and the correct classes, in the form of the annotation. While learning, the algorithm receives feedback on its performance by means of the annotation, and thereby it can learn which variables contain the most relevant information and it can give more weight to that input.

Each classifier comes with certain parameters that can be adjusted accordingly to the user's preference. After training, we use the validation set to optimize the parameters of the classifier. When we find the optimal tuning for (a subset of) the parameters, we can test the performance on the test set. It is important that none of the input data in the test set has been used in either the training or validation step, for this may influence the accuracy, such that the accuracy in the test step seems higher than it actually is.

While training the parameters, we have to find the right balance between overfitting and underfitting. Overfitting means that the classifier is too much tuned in on the noise of a certain dataset, thereby, it loses the ability for generalization. On the other hand, underfitting means that the classifier is not able to distinguish between certain movements and features, and thereby it is not able to detect patterns in the data. The validation step is used to resolve the problem of over- and underfitting.

Cross validation can also be used to reduce this problem. In k-fold cross-validation, the dataset is split into a training and a testing set. The training set is further divided into a training and validation set and this split is rotated for k times. Then the average accuracy is taken for the parameters, and with these settings, the algorithm is tested on the test set, which it has not been trained on.

7.1. Extracting keypoints

We have used OpenPose (84) to extract the keypoints from the persons in the video. OpenPose can accurately detect humans in videos. By extracting the keypoints from the videos, we have a representation that is independent of the gender or ancestry of the persons that occur in the video. Therefore, we account for the lack of variation in actors with this representation.

We store the keypoints of the persons in the video, as shown in Figures 7.1 and 7.2. While doing so, we assume that there are at most 2 persons in the video. If there is only one or no persons in the video, then we store the (X,Y)-coordinate keypoints for the second person as being '(-1000, -1000)'. Thereby, we ensure that we can store the same amount of features for all frames, which is necessary for classification, while also keeping track of the video-frames that do not contain two persons.

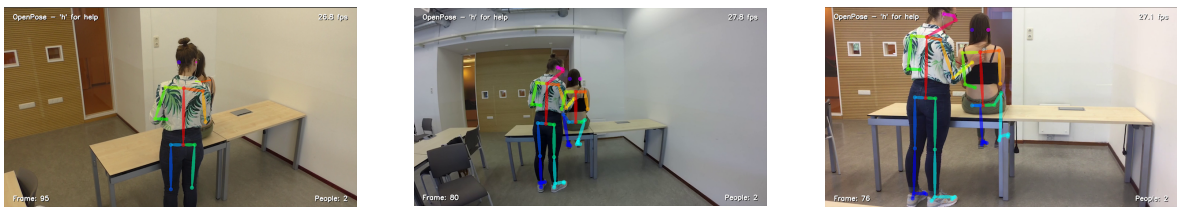


Figure 7.1: Extracted 2D skeleton joints for auscultation of the lungs on the back of the patient. Left: camera, middle: GoPro, right: iPad

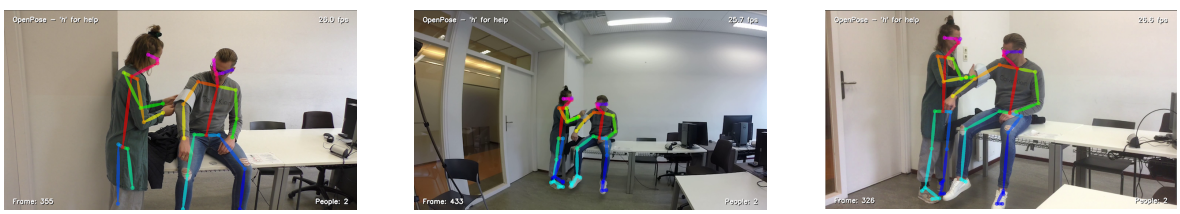


Figure 7.2: Extracted 2D skeleton joints for measuring the blood pressure. Left: camera, middle: GoPro, right: iPad.

However, Figure 7.3 shows two images of similar frames, filmed with the camera and the GoPro, in which OpenPose did not correctly recognise the legs of the patient.

We noticed that OpenPose sometimes has difficulty with correctly identifying the lower body parts, e.g. the knees and feet of the persons. However, the upper body parts are mostly correct. Since the medical actions in V2R all occur at the upper body of the patient, we expect the effect of this error may remain minimal.

Figure 7.3: Wrongly obtained 2D skeleton for two camera positions. Especially the legs of the patient are difficult to identify. Left: camera, right: GoPro

Since we recorded the sessions with three different camera's, all from different angles, and moved the cameras during the different days of shooting, the representation of the keypoints do not contain valuable information. Even though the actions from the same class are semantically similar, they are not necessarily numerically similar (103). Therefore, we have to do some mathematical calculations, e.g. calculate the distance or the angle between the keypoints (104).

7.2. Mathematical manipulation

The keypoints, as extracted using OpenPose, need mathematical manipulation in order to be numerically meaningful. (94) suggests to use multiple geometric features from skeleton based representations. One can use the relations between a selection of joints, or between all joints. A relation of the joints can be the distance between the joints of a person or the angle of the joints. We can calculate the Euclidean distance between two (or more) joints for a person (94) and for the distances between the joints of two persons (87).

Therefore, we store the 2D joint coordinates (J_c, Eq. 7.1) and calculate the

Euclidean distances (JJ_d, Eq. 7.2) between two joints (J_1 and J_2) as follows:

$$J_c = J_c(J) = (J_x, J_y) \tag{7.1}$$

$$JJ_d = J_1J_{2d}(X,Y) = \overrightarrow{||J_1J_2||} = \sqrt{(J_1(X) - J_2(X))^2 + (J_1(Y) - J_2(Y)^2} \tag{7.2}$$

The angle JJ_a of the body joints (J_1 and J_2) can be calculated as follows (105).

$$JJ_a = J_1J_{2a}(X,Y) = tan^{-1}(\frac{J_1(X) - J_2(X)}{J_1(Y) - J_2(Y)}) \tag{7.3}$$

We can calculate the distances between body joints and their angles per person, and we can store this between the body joints of two different persons as well. Both of these set contain valuable information.

Adding temporal information may help increase the accuracy of the classifier. We do so by segmenting our dataset, thereby we take multiple frames at once, and calculate the average and variance of the feature set, as well as the minimum and maximum value, e.g. the average angle of the upper body of a person and its variance, minimum and maximum value. Thereby, we are reducing the dimensionality of V2R, thus decreasing the training and testing time significantly.

7.2.1. Sets of features

We choose to run experiments with the following sets of features, to experimentally find which set contains the most valuable information.

Set 1. Pre-selected group of features, namely:

(i) Angle between the neck-midhip for both persons.
(ii) Distances between both hands for PersonA to a specific body part of PersonB, and both hands of PersonB to a specific body part of PersonA. These specific body parts are:
  (a) Chest
  (b) Abdomen
  (c) Arm
  (d) Left hand
  (e) Right hand

Feature set 2. Distances of all keypoints within PersonA and within PersonB, as shown in Figure 7.4.

Feature set 3. Angle between all keypoints within a person of both persons, as shown in Figure 7.5.

Feature set 4. Distances between both hands of PersonA to all upper body part of PersonB, and both hands of PersonB to all upper body part of PersonA, as shown in Figure 7.6.

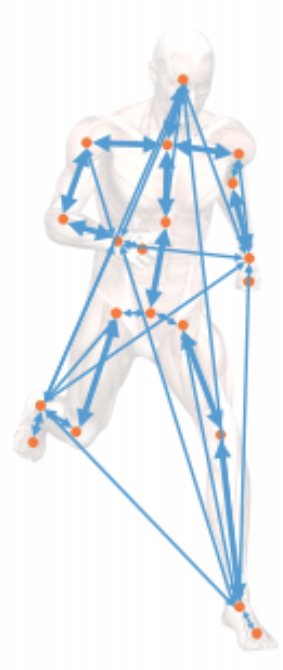


Figure 7.4: Feature set 2: Distances between all keypoints for a person, image from (94).

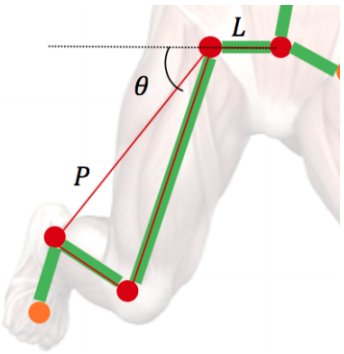


Figure 7.5: Feature set 3: The angle θ between two joints for a person, image from (94).

Feature set 5. Angle between the hands of PersonA to upper body part of PersonB and vice versa, as shown in Figure 7.7.

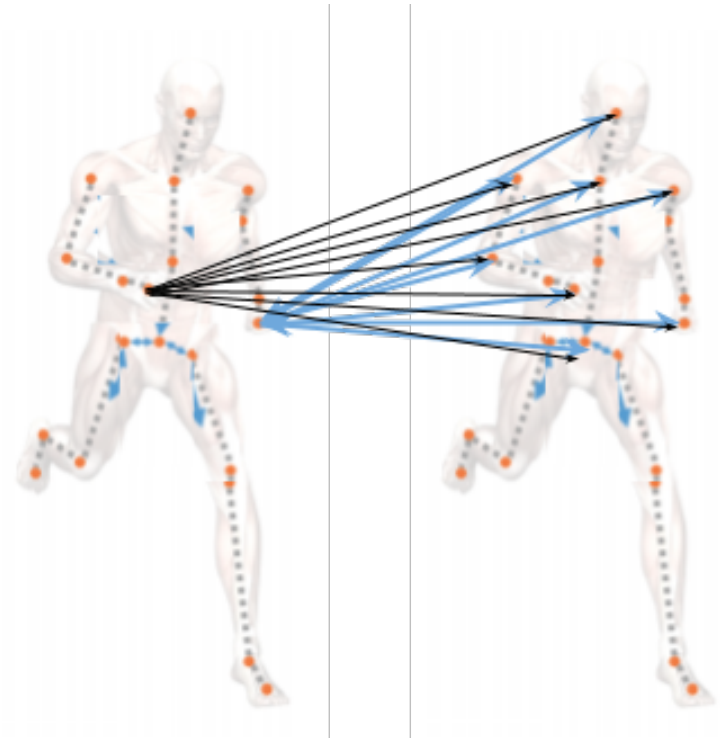


Figure 7.6: Feature set 4: The distance between both hands of one person to the upper body part of the other person. Image modified from (94).

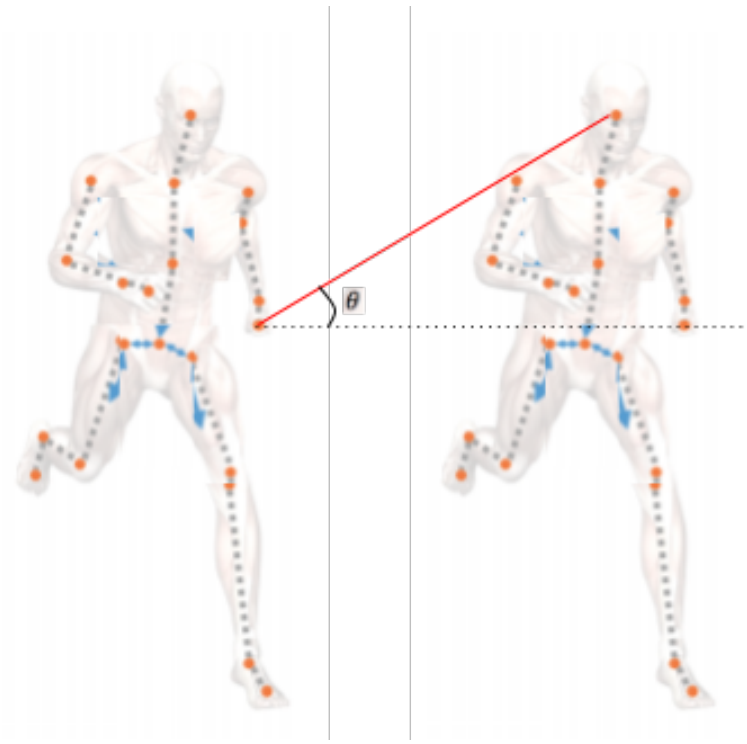


Figure 7.7: Feature set 5: The angle θ between the hands of one person to an upper body part of the other person. Image modified from (94).

8. Experiments and results

To train and test our classifiers, we ran multiple experiments. Since we are most interested in recognising the medical actions, rather than the posture of the patient, distance to the patient or the area of interest, we describe the best performing classifiers, i.e. the RF classifier, on these experiments and results first. The results that we have obtained, and present in this Section, can be used as a baseline for future research. In Section 8.2 we describe the experiments and results with all the classifiers.

Our dataset is split into 60% training-set, 20% validation-set, and a 20% test-set, as described in Section 6.9. In order to divide the different medical actions evenly, we first randomly select the medical actions, and divide these into the 60-20-20 sets. All videos occur once and only in one of these training, validation, or test sets.

Normally, cross-validation is used, to ensure that a classifier is not overfit to particular input features. However, since we do not have a a lot of computing power at our disposal, the different medical actions are not evenly divided over the sessions, and have limited time for the thesis, we will not apply a cross-validation. We report on the accuracy on the validation and test set.

We did a grid search with the parameters, and ran experiments with the feature sets individually, as well as combinations of them. In Table 8.1 we present the validation and test accuracies on predicting the medical actions for the DT and RF classifier. The RF achieves the best results, with 69,7% accuracy on the test set, on a combination of feature sets, namely 3,4, and 5 combined.

We wanted to combine feature sets 2, 3, 4, and 5 as well, however, since we did not have the computing power, and combining these 4 sets increases the input features considerably, we were not able to run experiments on this combination of feature sets.

Figure 8.1 shows the corresponding CMs for the best performing classifier, i.e. RF

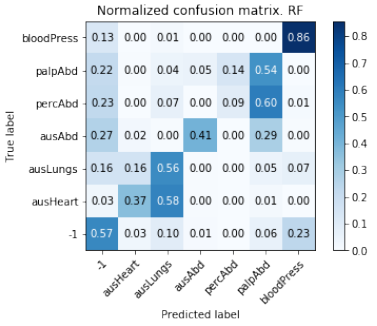
Feature set	Decision Tree			Random Forest		
	Validate	Test	Difference	Validate	Test	Difference
1	0.650	0.578	0.072	0.687	0.610	0.077
2	0.634	0.570	0.063	0.695	0.650	0.045
3	0.676	0.566	0.110	0.722	0.673	0.049
4	0.680	0.615	0.066	0.726	0.634	0.092
5	0.667	0.566	0.101	0.729	0.669	0.060
4,5	0.671	**0.643**	0.028	0.751	0.673	0.078
2,4,5	**0.708**	0.605	0.103	0.760	0.686	0.074
3,4,5	0.704	0.624	0.080	**0.774**	**0.697**	0.077

Table 8.1: Validation and test accuracies on the different feature sets for predicting the Medical Actions.

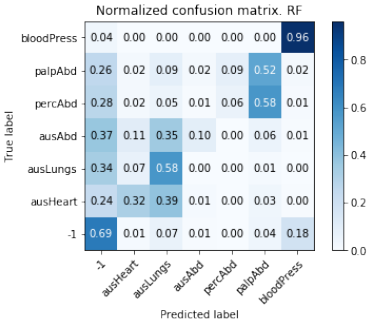
classifier, for the test results. Note that normally, the CMs are rotated 45°. However, for our CMs, the diagonal from the left corner in the bottom, to the top on the right are the true positives for our classes.

The classifier predicts the blood pressure measurement correctly most often, whereas medical action in the abdomen area are more difficult to recognise. Distinguishing palpation from percussion of the abdomen is most often mixed up. These two movements are rather similar if one is to look at the individual frames of the videos, rather than a segment of it. During palpation, both hands are pressed on the abdomen, while for percussion of the abdomen, one hand is not released from the abdomen, while the other is. Therefore, percussion of the abdomen is more easily confused as being palpation of the abdomen, as is also evident from the CMs.

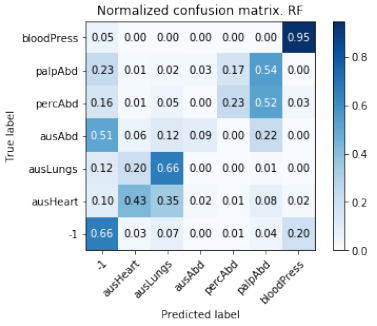
Similarly, auscultation of the heart and the lungs are alike for individual frames as well. Auscultation of the heart covers only the chest around the heart, whereas auscultation of the lungs covers the entire chest. Therefore, auscultation of the heart is more likely to be wrongly recognised as auscultation of the lungs, then vice versa.



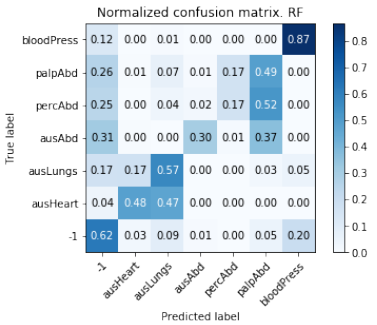
On feature set 1.



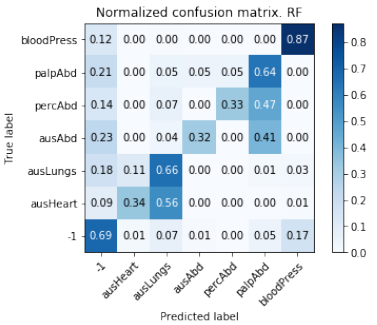
On feature set 2.



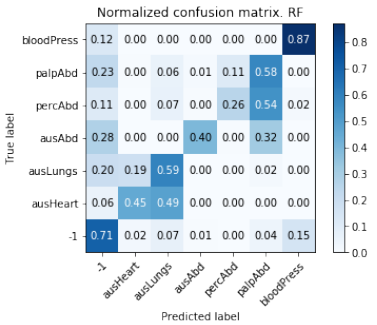
On feature set 3.



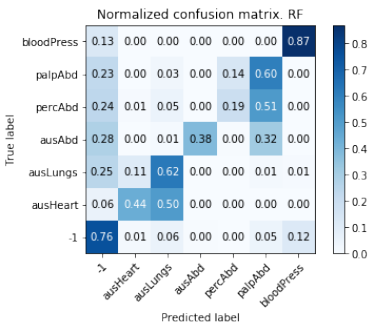
On feature set 4.



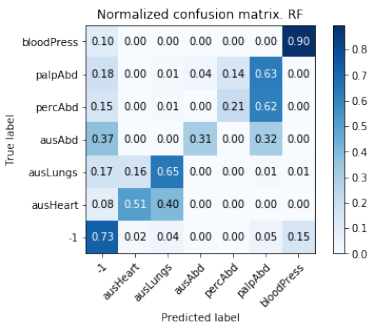
On feature set 5.



On feature set 4 and 5.



On feature set 2, 4, and 5.



On feature set 3, 4, and 5.

Figure 8.1: Confusion Matrices for the RF classifier, on various feature sets.

This is also represented in the CMs, in which the true label auscultation of the heart is wrongly predicted as being auscultation of the lungs, e.g. on feature set 3, 4, and 5 it correctly classifies it 51% of the instances and predicts it as being auscultation of the lungs in 40% of the occurrences.

8.1. Segmentation of the input features

V2R consists of videos rather than images. These have an extra dimension, namely time. Certain medical actions will be easier to detect when we take in the temporal aspect, e.g. auscultation of the heart compared to the lungs. To do so, we can feed the classifier with multiple frames at once, say x frames. Then, we assign the most occurring action among these x frames to all of them, e.g. $x = 30$ and 20 frames have been detected as auscultation of the heart, then all of these 30 frames will be assigned the label auscultation of the heart. By deciding on the segmenting length, we have to take the frame rate of the videos into account. Since we set all our recorded to have the same frame rate, we can use the same segments for all videos.

We ran experiments with the best performing combination of feature sets, i.e. feature sets 3, 4, and 5 combined, and added the temporal aspect in our experiments. Table 8.2 shows the training results, both for individual frames, as well as various segmentation lengths. Table 8.3 lists the highest accuracies during the training phase and the test results with these parameter settings. The test results increases with roughly 6% when we use segments of 120 frames, i.e. 2 seconds, and a sliding window of 20 frames.

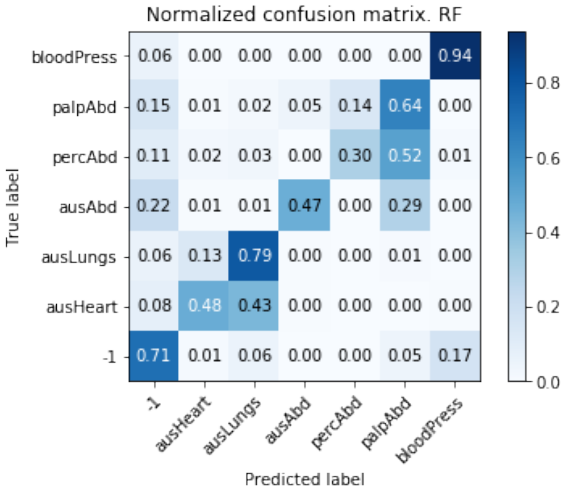
The corresponding CMs for these results are shown in Figure 8.2. We see an increase in accuracy on all medical actions. For comparison reasons, Figure 8.3 depicts the CMs for feature sets 3, 4, and 5 combined for the individual frames, and for the best performing segmentation window, i.e. 120 frames.

Classifier	Smoothing, N =	Nodes	Depth	Accuracy						
				No segment	Skip: 15, segment: 30	Skip: 20, segment: 30	Skip: 20, segment: 60	Skip: 20, segment: 90	Skip: 20, segment: 120	Skip: 20, segment: 150
Decision Tree	5	200	15	**0,704**	0,702	**0,691**	0,683	0,700	0,707	0,711
			20	0,704	**0,703**	0,690	0,684	**0,703**	**0,713**	**0,712**
			25	0,704	0,702	0,690	**0,685**	0,695	0,711	0,711
		400	15	0,695	0,693	0,689	0,677	0,691	0,694	0,704
			20	0,695	0,694	0,690	0,673	0,695	0,692	0,705
			25	0,695	0,697	0,684	0,671	0,694	0,696	0,702
		600	15	0,693	0,693	0,690	0,660	0,678	0,685	0,689
			20	0,692	0,687	0,690	0,663	0,680	0,690	0,700
			25	0,692	0,690	0,679	0,667	0,685	0,684	0,689
		800	15	0,683	0,685	0,678	0,662	0,681	0,698	0,696
			20	0,682	0,686	0,677	0,655	0,672	0,679	0,685
			25	0,682	0,687	0,679	0,647	0,671	0,686	0,690
Random Forest	5	200	15	0,750	0,756	0,767	0,760	0,770	0,773	0,768
			20	0,750	0,758	0,767	0,759	0,774	0,771	0,770
			25	0,750	0,758	0,767	0,759	0,774	0,771	0,770
		400	15	0,764	0,763	0,768	0,770	0,775	0,783	**0,777**
			20	0,769	0,765	0,770	0,769	0,775	0,787	0,769
			25	0,769	0,764	0,770	0,770	0,776	**0,787**	0,768
		600	15	0,768	0,762	0,768	**0,775**	0,776	0,781	0,774
			20	0,770	**0,769**	0,775	0,765	**0,780**	0,785	0,768
			25	0,770	0,765	0,777	0,767	0,780	0,783	0,768
		800	15	0,771	0,763	0,769	0,773	0,774	0,777	0,773
			20	0,773	0,766	0,776	0,775	0,779	0,782	0,773
			25	**0,774**	0,761	**0,777**	0,769	0,779	0,781	0,775

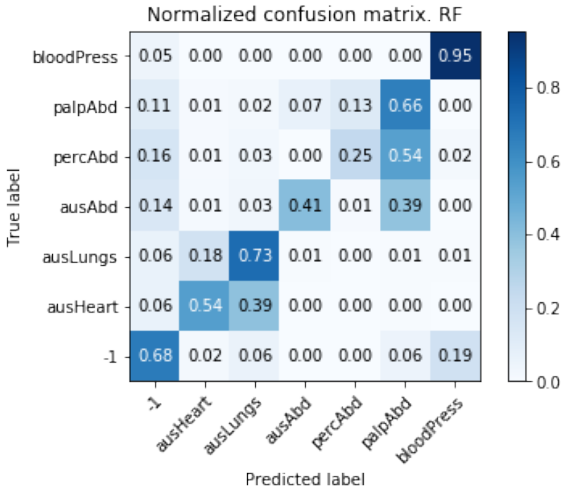
Table 8.2: Training accuracies for the DT and RF classifier, on feature sets 3, 4, and 5 combined. The classifier achieves highest training results for a segmentation length of 120 frames, with a skip length of 20 frames.

		Decision Tree			Random Forest		
Skip	Segments	Validate	Test	Difference	Validate	Test	Difference
15	30 frames	0.703	0.644	0.059	0.769	0.644	0.125
20	30 frames	0.691	0.614	0.077	0.777	0.740	0.037
20	60 frames	0.685	0.628	0.057	0.775	0.731	0.043
20	90 frames	0.703	**0.670**	0.033	0.780	0.749	0.031
20	120 frames	**0.713**	0.643	0.070	**0.787**	**0.756**	**0.031**
20	150 frames	0.712	0.684	**0.028**	0.777	0.727	0.050

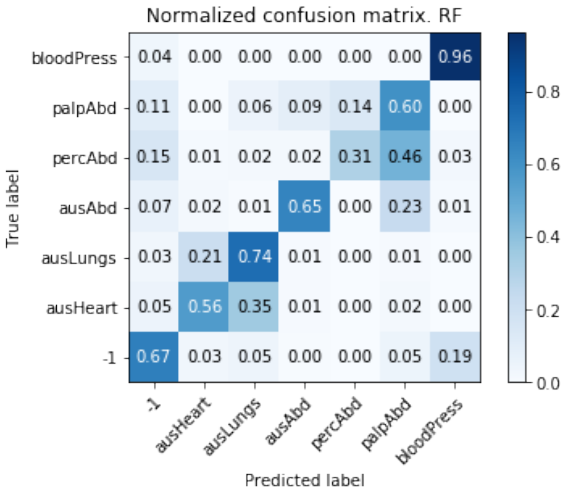
Table 8.3: Validation and test accuracies for the RF classifier on the medical actions, while taking into account the temporal aspect of the videos, by using segments of the videos as input, rather than individual frames.



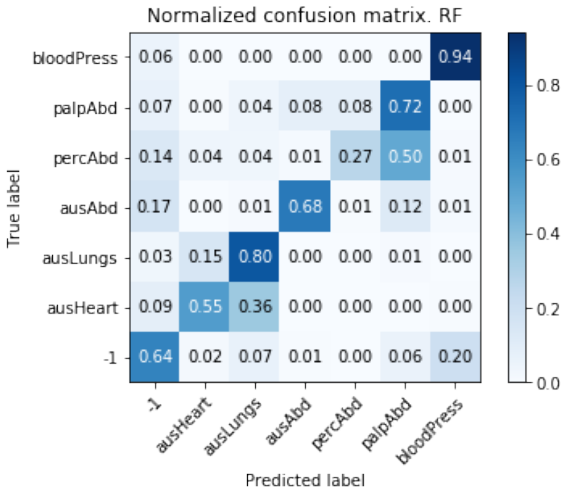
With segments of 30 frames.



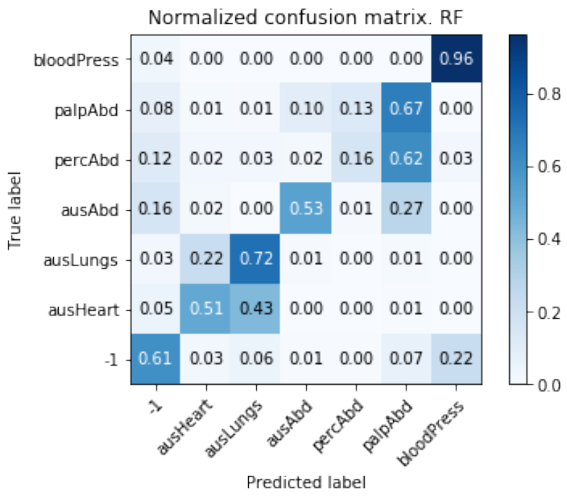
With segments of 60 frames.



With segments of 90 frames.

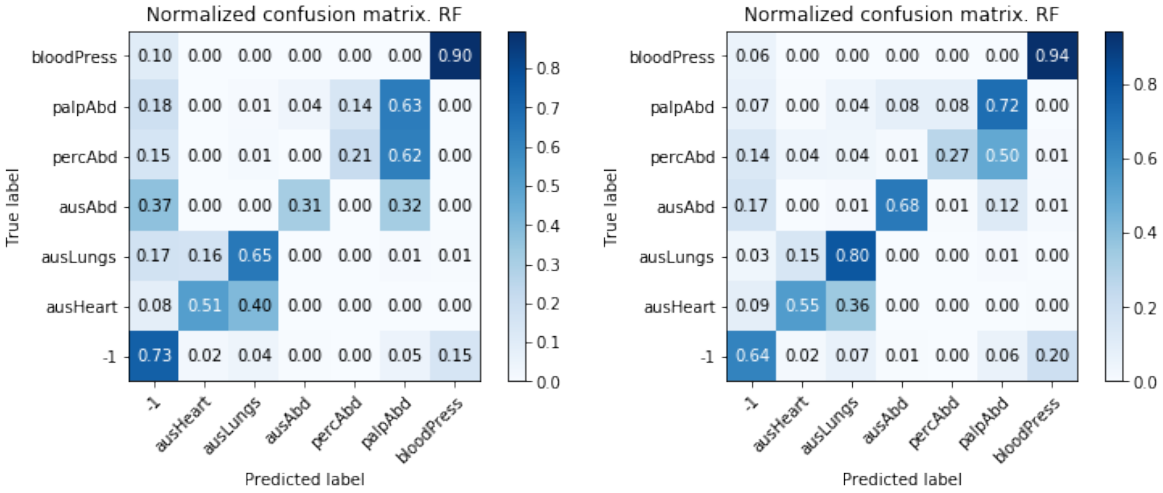


With segments of 120 frames.



With segments of 150 frames.

Figure 8.2: Confusion Matrices for the RF classifier, on feature sets 3, 4, and 5 combined, while taking into account the temporal aspects of the videos. The sliding window is 20 frames.



No segmentation With segments of 120 frames.

Figure 8.3: CMs for the RF classifier, on feature sets 3, 4, and 5 combined.

We see an increase of nearly 10% on palpation of the abdomen, and a 6% increase on percussion of the abdomen. Moreover, auscultation of the lungs is predicted correctly in 80% rather than 65% when we segment the input, and the classifier performs better on auscultation of the heart with an increase of 4%. Furthermore, a remarkable increase of $\leq$ 35% on auscultation of the abdomen for segmenting the video. This indicates that the temporal aspect of the videos captures valuable information.

8.2. Other results

We ran multiple other experiments, with varying parameters. We ran experiments to predict the posture of the patient, the distance from the GP to the patient, the area of investigation, and the medical action. Since calculation of the k-nn was costly, and did not result in better accuracies on the instances for the individual feature sets, we decided further experiment only with the DT and RF classifier.

In Chapters **??** and D, we displayed the training parameters for the classifiers with their results, and CMs on the testing results respectively, on the medical actions. In Chapters **??** and E, we displayed this for the area of investigation as well. Table 8.4

shows an overview of these acquired results, combined with the results on posture of the patient, and distance to the patient.

In Table 8.4, we listed all the acquired best results on the training and testing sets. From these results, we see that combining feature sets results in higher accuracies on both the training and test sets. The RF classifier scores highest on all predictions compared to the other classifiers. Combining feature sets 3, 4, and 5 achieves the best results, both with and without the temporal aspect taken into account. The posture of the patient, distance to the patient, and area of investigation is best predicted by taking into account the temporal aspect, while for medical action, the RF classifier achieves best results, by not taking into account the temporal aspect.

Action	Feature set	Decision Tree		Random Forest		k-nn	
		Validate	Test	Validate	Test	Validate	Test
Medical Action	1	0.650	0.578	0.687	0.610	0.688	0.596
	2	0.634	0.570	0.695	0.650	0.612	0.570
	3	0.676	0.566	0.722	0.673	NA	NA
	4	0.680	0.615	0.726	0.634	0.690	0.614
	5	0.667	0.566	0.729	0.669	0.680	0.648
	4,5	0.671	0.643	0.751	0.673	NA	NA
	2,4,5	0.708	0.605	0.760	0.686	NA	NA
	3,4,5	0.704	0.624	**0.774**	**0.697**	NA	NA
	Temp 4,5	0.695	0.654	0.748	0.684	NA	NA
	Temp 2,4,5	0.696	0.648	0.687	0.687	NA	NA
	Temp 3,4,5	0.703	0.644	0.769	0.644	NA	NA
Posture Patient	1	0.986	0.980	0.988	0.990	0.986	0.982
	2	0.968	0.970	0.994	0.992	0.982	0.980
	3	0.992	0.984	0.997	0.994	NA	NA
	4	0.981	0.972	0.989	0.988	0.990	0.988
	5	0.983	0.974	0.991	0.988	0.984	0.983
	4,5	0.983	0.973	0.992	0.993	NA	NA
	2,4,5	0.988	0.975	0.994	0.993	NA	NA
	3,4,5	0.992	0.988	0.997	0.995	NA	NA
	Temp 4,5	0.979	0.976	0.991	0.989	NA	NA
	Temp 2,4,5	0.975	0.973	0.994	0.992	NA	NA
	Temp 3,4,5	0.988	0.980	**0.998**	**0.996**	NA	NA
Distance to Patient	1	0.901	0.862	0.915	0.889	0.901	0.864
	2	0.877	0.828	0.908	0.878	0.835	0.822
	3	0.898	0.853	0.924	0.904	NA	NA
	4	0.899	0.880	0.919	0.887	0.902	0.871
	5	0.888	0.880	0.916	0.904	0.891	0.865
	4,5	0.907	0.888	0.927	0.901	NA	NA
	2,4,5	0.900	0.872	0.924	0.894	NA	NA
	3,4,5	0.910	0.876	0.933	0.906	NA	NA
	Temp 4,5	0.903	0.894	0.922	0.905	NA	NA
	Temp 2,4,5	0.905	0.857	0.930	0.913	NA	NA
	Temp 3,4,5	0.910	0.887	**0.935**	**0.910**	NA	NA
Area of Investigation	1	0.859	0.832	0.876	0.846	0.868	0.833
	2	0.804	0.760	0.847	0.807	0.783	0.762
	3	0.860	0.781	0.882	0.863	NA	NA
	4	0.860	0.843	0.891	0.864	0.882	0.852
	5	0.848	0.834	0.900	0.869	0.868	0.825
	4,5	0.884	0.850	0.906	0.881	NA	NA
	2,4,5	0.891	0.856	0.910	0.883	NA	NA
	3,4,5	0.895	0.871	0.917	0.890	NA	NA
	Temp 4,5	0.866	0.838	0.904	0.885	NA	NA
	Temp 2,4,5	0.883	0.839	0.915	0.885	NA	NA
	Temp 3,4,5	0.892	0.847	**0.926**	**0.908**	NA	NA

Table 8.4: Accuracy on the training and testing sets for the DT, RF, and k-nn classifiers. The RF classifier achieves the highest accuracies on all feature sets, and on all action classes. For the temporal aspect, we chose a sliding window of 15 frames, and segmentation length of 30 frames

9. Discussion

V2R consists of sequences of medical actions. The order of these sequences are the same in all our sessions, however in real life the order might change a bit, with an exception to the examination of the abdomen and the entire inspection of the body. The order as found in V2R is also the order in real life for these sequences.

What is more, we do not track or identify the GP and patient in the recordings. If we were to do so, this might have increased the accuracy of the classifier.

9.1. Limitations on our dataset

We carefully selected our medical actions, based on privacy issues, relevancy, occurrence in medical guidelines, and available medical instruments. Since blood pressure measurement was a frequently occurring medical action, and we did have a blood pressure monitor at our disposal, we decided to use this as well. However, the blood pressure monitor did not work. To use it, we put the cuff around the arm and waited roughly 30 seconds, as if it were working. However, guessing the time while acting was not as reliable as previously assumed. On average, measuring blood pressure took 1:37 min, with the shortest measurement being 0:43 min, while the longest took 2:40 min. This is not representative for the real measurement, in which the measurement takes around 40 seconds. Even though the measurement was not representative, measuring blood pressure was correctly recognised 94% of its occurrences. Most likely, this is the case since measuring blood pressure was the only medical action in V2R that was performed on the arm of the patient.

Even though we tried to account for diversity in V2R, there were some limitations while recording. For instance, the number of subject is limited and therefore, the diversity in actors is also very limited. Most of the videos are recorded with two female actors. So, there is limitation in the gender, as well as the age, body type, and origin of the actors. Moreover, we have small variance in the set of movements, as one GP

may move his/her body differently than other GPs. Moreover, we have a very limited set of instruments that we use in V2R. This limited the amount of medical actions.

Furthermore, we were unable to schedule a proper facility for this study, therefore, we had to create one ourselves, using classrooms of Utrecht University. Since we use OpenPose to extract the movements of the persons in the video, rather than extract features of the room, this will not interfere with our results.

9.2. Limitations on our Machine Learning approach

We decided to train and test three commonly used basic classifiers. Nowadays, more complex classifiers are available, which could be able to pick up on more complex and dynamic input.

We implemented segmentation of the frames, with a sliding window, to capture the temporal information from the video. However, when training and validating our classifiers, we used a randomly preselected training and validation set, rather than applying cross-validation. This limits the generalisability of our approach, and decreases the certainty that our classifier is overfitted. However, when we apply segmentation for our training and testing steps, the difference in prediction between training and testing accuracy decreases, indicating that utilizing the temporal aspect, causes our model to generalise better.

We utilize OpenPose (84) to extract the 2D skeletons on the persons in the videos. While OpenPose is a pretrained CNN, with accurate results, it is not always able to correctly produce all keypoints accurately. Figure 9.1 shows a sequence of the same session, filmed by the camera, in which OpenPose has difficulty recognising the 2D skeletons of the persons. Especially the legs are recognised wrongly, whereas the upper body is recognised correctly in the majority of the frames.

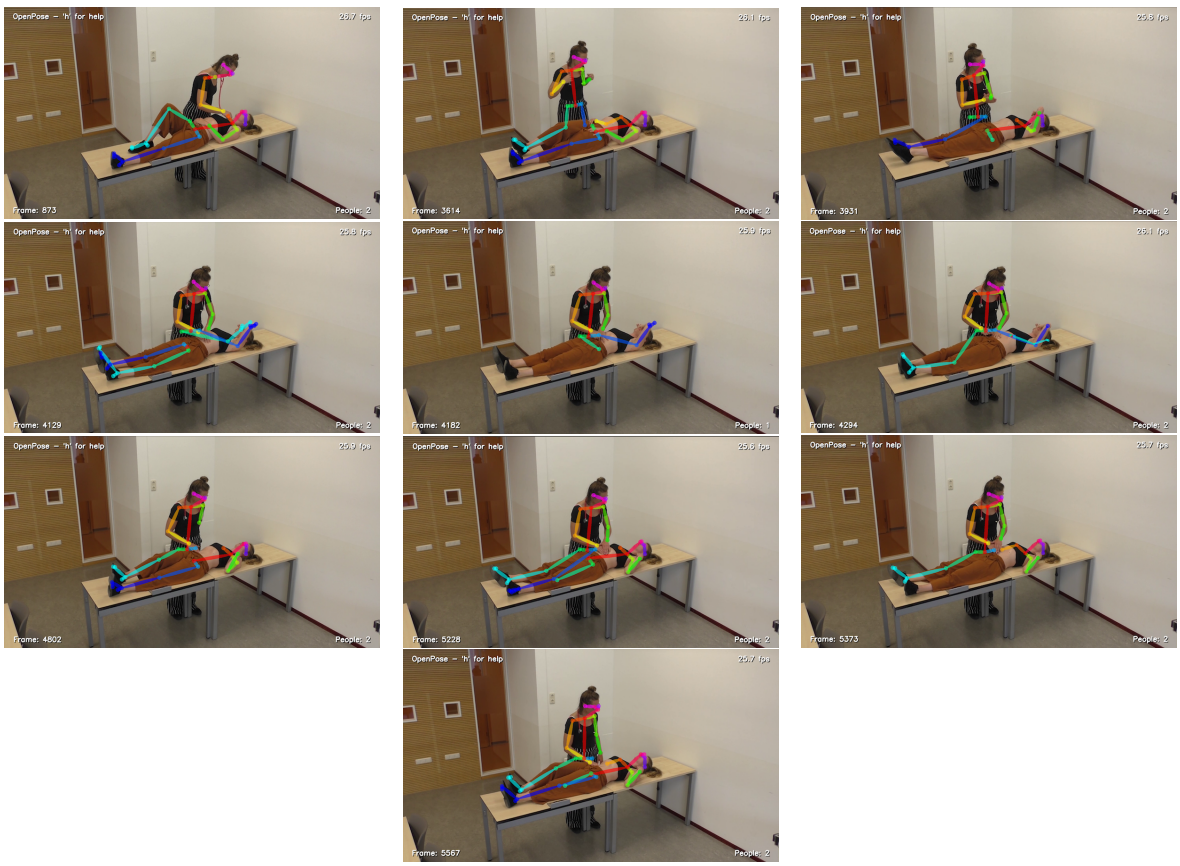


Figure 9.1: A sequence of a session with the acquired 2D skeletons in the video. Especially the legs are difficult to interpret for OpenPose. However, the upperbody remains correctly recognised in most frames.

10. Conclusion

While previous research in healthcare focused on providing aid for elderly care, or detecting fine-grained movements during surgery, our research focuses on recognising medical actions, to support automatically storing patients' files to the EMR.

In order to do so, we focused on medical actions as found in the GPs office, during one-on-one consultations between a GP and a patient. Since there was no available dataset, we had to collect one ourselves. Therefore, we carefully selected the most occurring and relevant medical actions. We recorded 192 individual sessions, and since they were shot with multiple cameras simultaneously, we created 451 videos on these sessions.

With the use of OpenPose, we were able to extract the 2D skeletons from the persons. After mathematical manipulation, we obtained the distances and angles of the skeleton joints. With these features, we trained three commonly used basic classifiers, of which the RF classifier unanimously scored best on all feature sets.

We were able to correctly predict the posture of the patient with more than 99,5% from stand-alone frames. Moreover, we were also able to predict whether the GP touches the patient or not, with a certainty of 90,6%. The RF classifier was able to correctly recognise the area of interest with a certainty of over 89%, based on the individual frames. When taking into account the temporal aspect of the videos, this increased to 91,8%. Lastly, the RF classifier is trained to correctly predict 69,7% of the medical actions, based on individual frames. If we take segments rather than individual frames as the input, this increases to 75,6%.

Future research may increase these scores by identifying and tracking the GP and the patient may increase the prediction scores, since medical actions might be easier to detect. Moreover, ensuring generalizability of the model would achieved by applying cross-validation. Furthermore, expanding the dataset with more sessions,

using varying subjects, medical actions, and medical instruments will further increase the generalizability of the results.

REFERENCES

1. Golob Jr, J. F., J. J. Como and J. A. Claridge, "The painful truth: The documentation burden of a trauma surgeon", *Journal of Trauma and Acute Care Surgery*, Vol. 80, No. 5, pp. 742–747, 2016.

2. Sheppard, J. E., L. C. Weidner, S. Zakai, S. Fountain-Polley and J. Williams, "Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping", *Archives of disease in childhood*, Vol. 93, No. 3, pp. 204–206, 2008.

3. Campanella, P., E. Lovato, C. Marone, L. Fallacara, A. Mancuso, W. Ricciardi and M. L. Specchia, "The impact of electronic health records on healthcare quality: a systematic review and meta-analysis", *The European Journal of Public Health*, Vol. 26, No. 1, pp. 60–64, 2015.

4. Poppe, R., "A survey on vision-based human action recognition", *Image and vision computing*, Vol. 28, No. 6, pp. 976–990, 2010.

5. Shah, M. and R. Jain, *Motion-based recognition*, Vol. 9, Springer Science & Business Media, 2013.

6. Prati, A., C. Shan and K. I.-K. Wang, "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring", *Journal of Ambient Intelligence and Smart Environments*, Vol. 11, No. 1, pp. 5–22, 2019.

7. Moeslund, T. B., A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis", *Computer vision and image understanding*, Vol. 104, No. 2-3, pp. 90–126, 2006.

8. Ajami, S., "Use of speech-to-text technology for documentation by healthcare providers", *The National medical journal of India*, Vol. 29, No. 3, p. 148, 2016.

9. Klann, J. G. and P. Szolovits, "An intelligent listening framework for capturing encounter notes from a doctor-patient dialog", *BMC medical informatics and decision making*, Vol. 9, No. 1, p. S3, 2009.

10. Quellec, G., K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener and G. Cazuguel, "Real-time recognition of surgical tasks in eye surgery videos", *Medical image analysis*, Vol. 18, No. 3, pp. 579–590, 2014.

11. Schuldt, C., I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3, pp. 32–36, IEEE, 2004.

12. Weinland, D., R. Ronfard and E. Boyer, "Free viewpoint action recognition using motion history volumes", *Computer vision and image understanding*, Vol. 104, No. 2-3, pp. 249–257, 2006.

13. Choi, W., K. Shahid and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people", *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1282–1289, IEEE, 2009.

14. Singh, S., S. A. Velastin and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods", *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 48–55, IEEE, 2010.

15. Rodriguez, M. D., J. Ahmed and M. Shah, "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition.", *CVPR*, Vol. 1, p. 6, 2008.

16. Laptev, I., M. Marszałek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", *CVPR 2008-IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

17. Marszałek, M., I. Laptev and C. Schmid, "Actions in context", *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*, pp. 2929–2936, IEEE Computer Society, 2009.

18. Akl, A. and S. Valaee, "Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing", *2010 IEEE International*

Conference on Acoustics, Speech and Signal Processing, pp. 2270–2273, IEEE, 2010.

19. Al-Shabi, M., W. P. Cheah and T. Connie, "Facial expression recognition using a hybrid CNN-SIFT aggregator", *arXiv preprint arXiv:1608.02833*, 2016.

20. Gorelick, L., M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes", *IEEE Transactions On Pattern Analysis and Machine Learning (ICCV'07) Volume 29*, Vol. 29(12), pp. 2247–2253, IEEE, 2007.

21. Wang, Y., K. Huang and T. Tan, "Human activity recognition based on R transform", Vol. 2, pp. 1–8, CVPR, 2007.

22. Woolhandler, S., T. Campbell and D. U. Himmelstein, "Costs of health care administration in the United States and Canada", *New England Journal of Medicine*, Vol. 349, No. 8, pp. 768–775, 2003.

23. Woolhandler, S. and D. U. Himmelstein, "Administrative work consumes one-sixth of US physicians' working hours and lowers their career satisfaction", *International Journal of Health Services*, Vol. 44, No. 4, pp. 635–642, 2014.

24. De Veer, A., K. d. Groot, M. Brinkman and A. Francke, *Administratieve druk: méér dan kwestie van tijd.*, Tech. rep., NIVEL, 2017.

25. Mishra, P., J. C. Kiang and R. W. Grant, "Association of medical scribes in primary care with physician workflow and patient experience", *JAMA internal medicine*, Vol. 178, No. 11, pp. 1467–1472, 2018.

26. Friedberg, M. W., P. G. Chen, K. R. Van Busum, F. Aunon, C. Pham, J. Caloyeras, S. Mattke, E. Pitchforth, D. D. Quigley, R. H. Brook *et al.*, "Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy", *Rand health quarterly*, Vol. 3, No. 4, 2014.

27. Shastri, S., M. Wasserman and V. Chidambaram, "The Seven Sins of Personal-Data Processing Systems under GDPR", *USENIX HotCloud*, 2019.

28. Maas, L., M. Geurtsen, F. Nouwt, S. Schouten, R. Van De Water, S. Van Dulmen, F. Dalpiaz, K. Van Deemter and S. Brinkkemper, "The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare", Unpublished.

29. Libelium Comunicaciones Distribuidas S.L., Zaragoza, Spain, *MySignals SW eHealth and Medical IoT Development Platform Technical Guide*, 4.6 edn., 5 2019.

30. Laptev, I. and B. Caputo, "Recognition of human actions", `http://www.nada.kth.se/cvap/actions/`, October 2019.

31. Gorelick, L., M. Blank, E. Shechtman, M. Irani and R. Basri, "Weizmann actions as space-time shapes", `http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html`, October 2019.

32. of Central Florida, U., "UCF aerial camera, rooftop camera and ground camera dataset", `http://vision.eecs.ucf.edu/data/UCFARG.html`, October 2019.

33. Yuan, J., Z. Liu and Y. Wu, "Discriminative video pattern search for efficient action detection", `http://users.ece.northwestern.edu/~jyu410/index_files/actiondetection.html`, October 2019.

34. of Surrey, U. and CERTH-ITI, "i3dpost multi-view human action datasets", `http://kahlan.eps.surrey.ac.uk/i3dpost_action/`, October 2019.

35. Chen, C.-C., M. S. Ryoo and J. K. Aggarwal, "UT-Tower Dataset: Aerial View Activity Classification Challenge", `http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html`, 2010.

36. Team, C., "Caviar: context aware vision using image-based active recognition", `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`, October 2019.

37. Fisher, R., "Behave: Computer-assisted prescreening of video streams for unusual activities", `http://homepages.inf.ed.ac.uk/rbf/BEHAVE/`, October 2019.

38. Group, V. G., "Tv human interactions dataset", http://www.robots.ox.ac.uk/ vgg/data/tv_human_interactions/index.html, October 2019.

39. Ryoo, M. S. and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)", http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

40. INRIA, "Etiseo video understanding evaluation", https://www-sop.inria.fr/orion/ETISEO/index.htm, October 2019.

41. INRIA, "Inria xmas motion acquisition sequences (ixmas)", http://4drepository.inrialpes.fr/public/viewgroup/6, October 2019.

42. for Biometrics, C. and S. Research, "Casia action database for recognition", http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp, October 2019.

43. Tran, D., A. Sorokin and D. Forsyth, "Human activity recognition with metric learning", http://vision.cs.uiuc.edu/projects/activity/, October 2019.

44. Laptev, I., M. Marszałek, C. Schmid and B. Rozenfeld, "Learning Human Actions from Movies", https://www.di.ens.fr/ laptev/actions/, October 2019.

45. Laptev, I., "Hollywood2: human actions and scenes dataset", https://www.di.ens.fr/ laptev/actions/hollywood2/, October 2019.

46. of Central Florida, U., "UCF sports action dataset", https://www.crcv.ucf.edu/data/UCF_Sports_Action.php, October 2019.

47. Murtaza, F., M. H. Yousaf and S. A. Velastin, "MuHAVi: Multicamera Human Action Video Data", http://velastin.dynu.com/MuHAVi-MAS/, October 2019.

48. of Central Florida, U., "UCF50 - Action Recognition Data Set", https://www.crcv.ucf.edu/data/UCF50.php, October 2019.

49. of Central Florida, U., "UCF101 - Action Recognition Data Set", https://www.crcv.ucf.edu/data/UCF101.php, October 2019.

50. for Informatics, M. P. I., "MPII Cooking Activities Dataset", https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpii-cooking-activities-dataset/, October 2019.

51. Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks", https://cs.stanford.edu/people/karpathy/deepvideo/, October 2019.

52. Heilbron, F., V. Escorcia, B. Ghanem and J. Niebles, "A Large-Scale Video Benchmark for Human Activity Understanding", http://activity-net.org/index.html, October 2019.

53. Gorban, A., H. Idrees, Y.-G. Jiang, A. R. Zamir, I. Laptev, M. Shah and R. Sukthankar, "Action Recognition in Temporally Untrimmed Videos!", http://www.thumos.info/home.html, October 2019.

54. Sigurdsson, G. A., G. Varol, X. Wang, A. Farhadi, I. Laptev and A. Gupta, "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding", *European Conference on Computer Vision*, pp. 510–526, Springer, 2016.

55. DeepMind, "The Kinetics Human Action Video Dataset", https://deepmind.com/research/publications/kinetics-human-action-video-dataset, October 2019.

56. DeepMind, "A Short Note about Kinetics-600", https://deepmind.com/research/publications/short-note-about-kinetics-600, October 2019.

57. DeepMind, "Kinetics", https://deepmind.com/research/open-source/kinetics, October 2019.

58. lab, S., "Hmdb: a large video database for human motion recognition", http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/, October 2019.

59. Fisher, R. B., "The PETS04 surveillance ground-truth data sets", *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, pp. 1–5, 2004.

60. Blunsden, S. and R. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification", *Annals of the BMVA*, Vol. 4, No. 1-12, p. 4, 2010.

61. Nghiem, A., F. Bremond, M. Thonnat and R. MA, "A new evaluation approach for video processing algorithms", *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*, pp. 15–15, IEEE, 2007.

62. Nghiem, A.-T., F. Bremond, M. Thonnat and V. Valentin, "ETISEO, performance evaluation for video surveillance systems", *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 476–481, IEEE, 2007.

63. Yuan, J., Z. Liu and Y. Wu, "Discriminative subvolume search for efficient action detection", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2442–2449, IEEE, 2009.

64. Gkalelis, N., H. Kim, A. Hilton, N. Nikolaidis and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database", *2009 Conference for Visual Media Production*, pp. 159–168, IEEE, 2009.

65. Kuehne, H., H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition", *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

66. Rohrbach, M., S. Amin, M. Andriluka and B. Schiele, "A database for fine grained activity detection of cooking activities", *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194–1201, IEEE, 2012.

67. Reddy, K. K. and M. Shah, "Recognizing 50 human action categories of web videos", *Machine Vision and Applications*, Vol. 24, No. 5, pp. 971–981, 2013.

68. Soomro, K., A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild", *arXiv preprint arXiv:1212.0402*, 2012.

69. Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks", *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

70. Caba Heilbron, F., V. Escorcia, B. Ghanem and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

71. Gorban, A., H. Idrees, Y.-G. Jiang, A. R. Zamir, I. Laptev, M. Shah and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes", `http://www.thumos.info/`, 2015.

72. Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset", *arXiv preprint arXiv:1705.06950*, 2017.

73. Carreira, J. and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset", *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

74. Carreira, J., E. Noland, A. Banki-Horvath, C. Hillier and A. Zisserman, "A short note about kinetics-600", *arXiv preprint arXiv:1808.01340*, 2018.

75. Carreira, J., E. Noland, C. Hillier and A. Zisserman, "A Short Note on the Kinetics-700 Human Action Dataset", *arXiv preprint arXiv:1907.06987*, 2019.

76. Hellwig, B., *ELAN - Linguistic Annotator*, The Language Archive, MPI for Psycholinguistics, Nijmegen, The Netherlands, version 5.4 edn., 12 2018.

77. Salah, A. A., R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten and E. Pauwels, "Multimodal identification and localization of users in a smart environment", *Journal on Multimodal User Interfaces*, Vol. 2, No. 2, pp. 75–91, 2008.

78. Alpaydin, E., *Introduction to Machine Learning, third edition*, Masschusetts Institute of Technology, 2014.

79. Quinlan, J. R., "Induction of decision trees", *Machine learning*, Vol. 1, No. 1, pp. 81–106, 1986.

80. Stuwart, M., "Simple Introduction to Convolutional Neural Networks", `https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac`, December 2019.

81. Han, J. and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning", *International Workshop on Artificial Neural Networks*, pp. 195–201, Springer, 1995.

82. Nair, V. and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

83. Scherer, D., A. Müller and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition", *International conference on artificial neural networks*, pp. 92–101, Springer, 2010.

84. Cao, Z., G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields", *arXiv preprint arXiv:1812.08008*, 2018.

85. Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft coco: Common objects in context", *European conference on computer vision*, pp. 740–755, Springer, 2014.

86. Andriluka, M., L. Pishchulin, P. Gehler and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis", *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

87. Yun, K., J. Honorio, D. Chattopadhyay, T. L. Berg and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning",

2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 28–35, IEEE, 2012.

88. Zhu, W., C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks", *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

89. Ohn-Bar, E. and M. Trivedi, "Joint angles similarities and HOG2 for action recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 465–470, 2013.

90. Li, W., Z. Zhang and Z. Liu, "Action recognition based on a bag of 3d points", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 9–14, IEEE, 2010.

91. Wang, J., Z. Liu, J. Chorowski, Z. Chen and Y. Wu, "Robust 3d action recognition with random occupancy patterns", *European Conference on Computer Vision*, pp. 872–885, Springer, 2012.

92. Ellis, C., S. Z. Masood, M. F. Tappen, J. J. LaViola and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition", *International Journal of Computer Vision*, Vol. 101, No. 3, pp. 420–436, 2013.

93. Song, S., C. Lan, J. Xing, W. Zeng and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data", *Thirty-first AAAI conference on artificial intelligence*, 2017.

94. Zhang, S., X. Liu and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks", *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 148–157, IEEE, 2017.

95. Shahroudy, A., J. Liu, T.-T. Ng and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

96. Ofli, F., R. Chaudhry, G. Kurillo, R. Vidal and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database", *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, IEEE, 2013.

97. Xia, L., C.-C. Chen and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints", *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, IEEE, 2012.

98. Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

99. Nivel, "Nivel. Kennis voor betere zorg", `https://nivel.nl/nl`, October 2019.

100. Unknown, "Zoektocht in Chinese geneeskunde", `https://chinesegeneeskunde.blogspot.com/2015/`, October 2019, image.

101. Aziz, A. M., "thoracic lung assessment", `https://www.slideshare.net/AliMohamedAziz/thoracic-lung-assessment`, October 2019, image.

102. Jongh de, T. O. and B. T. J. Meursing, "Onderzoek van het hart", `https://www.ntvg.nl/artikelen/onderzoek-van-het-hart/volledig`, October 2019, image.

103. Müller, M., T. Röder and M. Clausen, "Efficient content-based retrieval of motion capture data", *ACM Transactions on Graphics (ToG)*, Vol. 24, pp. 677–685, ACM, 2005.

104. Yao, A., J. Gall, G. Fanelli and L. Van Gool, "Does human action recognition benefit from pose estimation?", *BMVC 2011-Proceedings of the British Machine Vision Conference 2011*, 2011.

105. Noori, F. M., B. Wallace, M. Z. Uddin and J. Torresen, "A Robust Human Activity Recognition Approach Using OpenPose, Motion Features, and Deep Recurrent Neural Network", *Scandinavian Conference on Image Analysis*, pp. 299–310, Springer, 2019.

APPENDIX A: Adobe Pro export settings

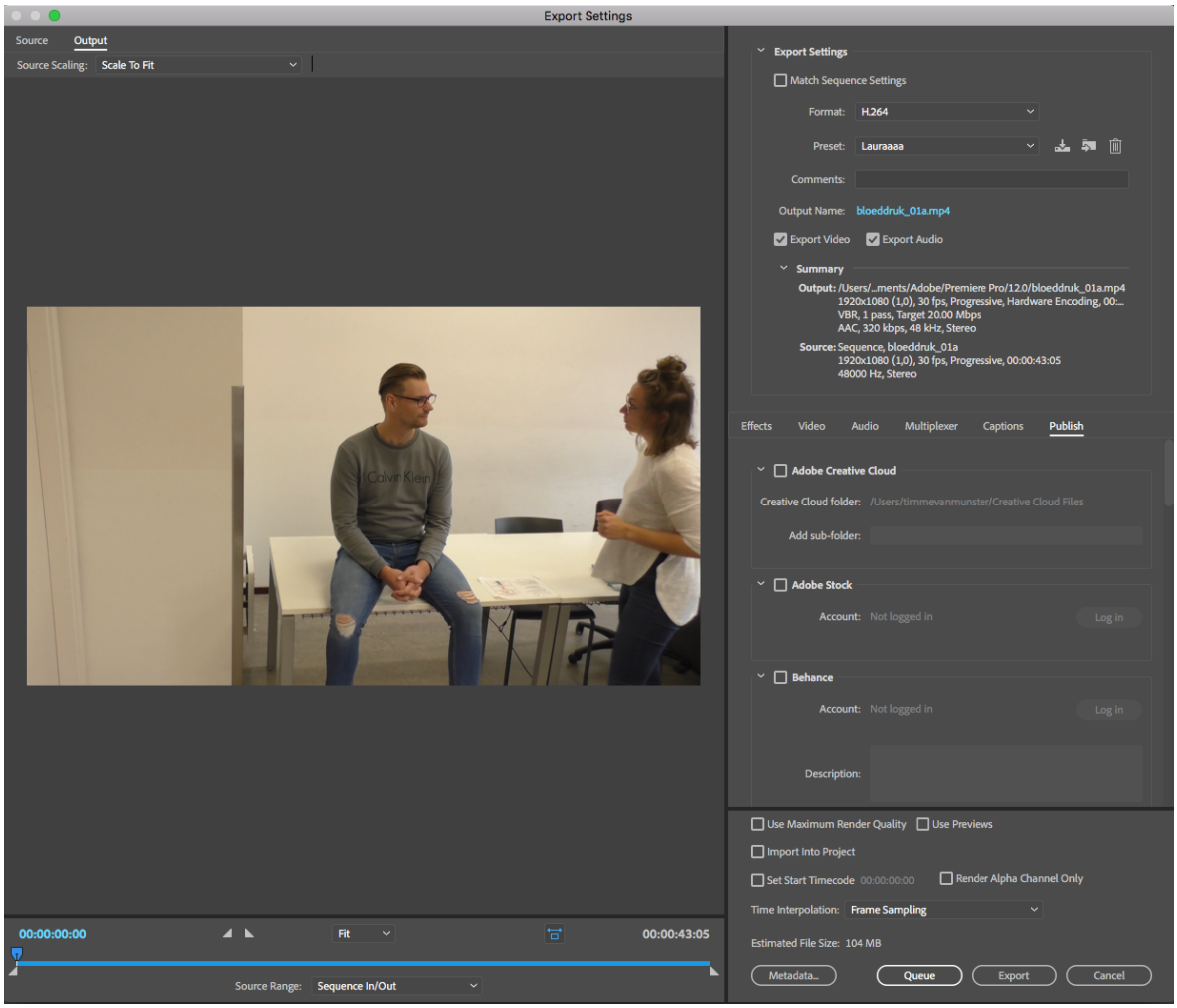


Figure A.1: Adobe Pro export settings

APPENDIX B: Synchronisation mode of ELAN

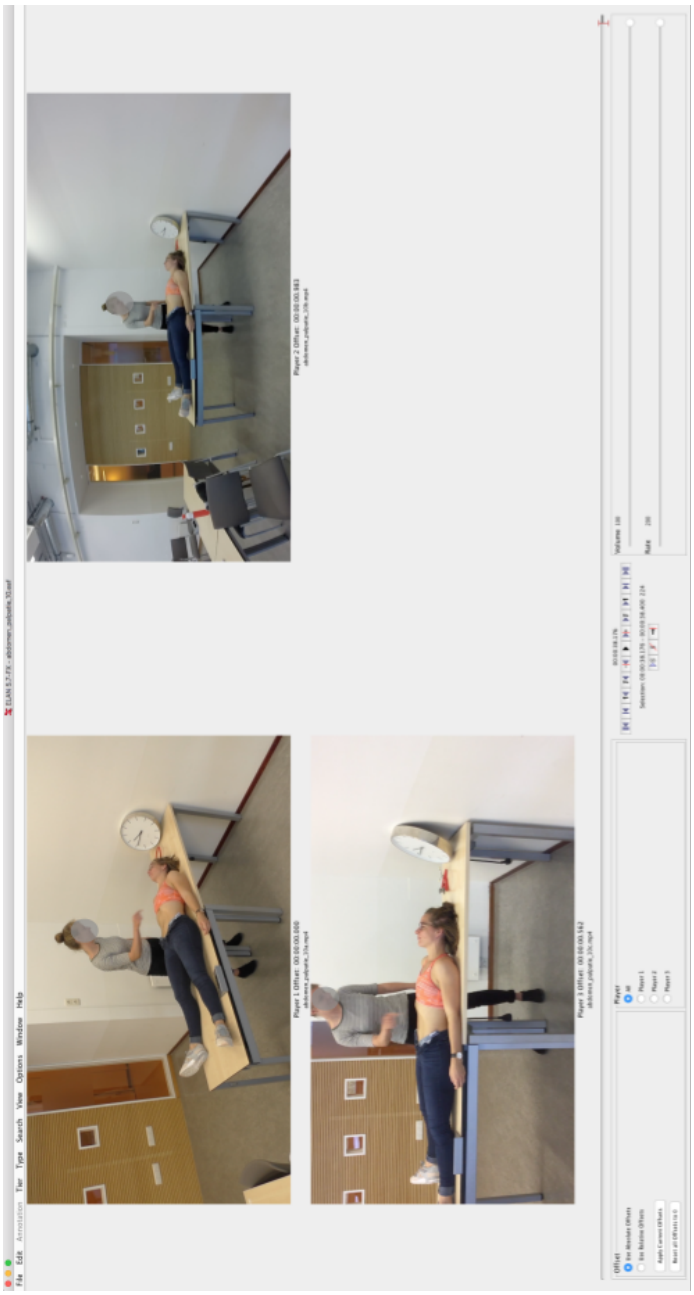


Figure B.1: Synchronisation mode of the ELAN tool

APPENDIX C: Overview of the extracted and chosen medical actions

BPM	Blood Pressure Measurement
PaA	Palpation Abdomen
HF	Heart Frequency
PeA	Percussion Abdomen
BMI	Measuring Body Mass Index
MT	Measuring Temperature
AL	Auscultation Lungs
AH	Auscultation Heart
TEy	Testing the eyes
PaE	Palpation Elsewhere
TEa	Testing the Ears
ECG	ECG
AA	AuscultationAbdomen

Figure C.1: The abbreviations as used in the figures C.2, C.3, and C.4

	BPM	PaA	HF	PeA	BMI	MT	AL	AH	TEy	PaE	TEa	ECG	AA
Occurrence of same medical sequences													
6										PaE			
5		PaA		PeA									
5	BPM				BMI								
4	BPM												
4									Tey				
4											Tea		
3	BPM	PaA	HF	PeA		MT							
3					BMI								
3							AL						
2	BPM	PaA	HF										
2	PBM		HF					AH				ECG	
2									TEy		TEa		
1			HF			MT							
1		PaA		PeA									AA
1		PaA		PeA	BMI	MT							
1	PBM	PaA	HF	PeA									
1		PaA											AA
1	PBM	PaA		PeA	BMI								
1	BPM		HF				AL	AH		PaE			
1	BPM							AH				ECG	
1	BPM							AH					
1	BPM		HF			MT	AL						
1	BPM									PaE			
1	BPM		HF				AL	AH				ECG	
1	BPM	PaA	HF	PeA		MT	AL	AH					
1	BPM				BMI	MT	AL	AH	TEy				
1							AL		TEy		Tea		
1							AL	AH					
1	BPM					MT							
1		PaA				MT							

Figure C.2: The relevant and record-able medical actions as found in the medical guidelines. The actions are listed as sequences, from most to least occurring sequences.

	BPM	PaA	HF	PeA	AL	AH	AA
Occurrences of medical actions that we chose							
11	BPM						
6		PaA		PeA			
5	BPM		HF		AL	AH	
4					AL		
4	BPM	PaA	HF	PeA			
2	BPM					AH	
2	BPM	PaA	HF				
2	BPM				AL	AH	
1		PaA					
1			HF				
1		PaA					AA
1	BPM	PaA		PeA			
1		PaA		PeA			AA
1	BPM	PaA	HF	PeA	AL	AH	

Figure C.3: The recorded medical actions, listed from most occurring sequence to least occurring sequence as found in the medical guidelines. These medical sequences cover 42 of the medical guidelines.

Removed actions		
6	PaE	
3	BMI	
2	TEa	TEy
4	TEy	
4	TEa	

Figure C.4: The eliminated medical actions, listed from most to least occurring sequences as found in the medical guidelines.

APPENDIX D: CMs for the DT, RF, and k-nn classifiers for various feature sets, trained and tested on medical actions

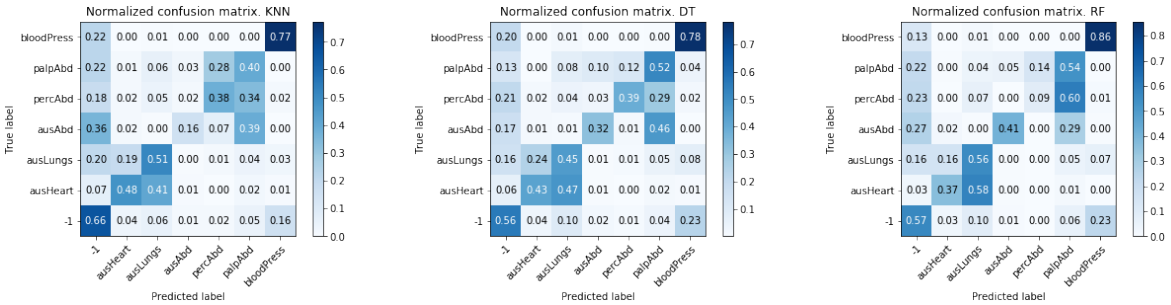


Figure D.1: CMs for the three classifiers on feature set 1.

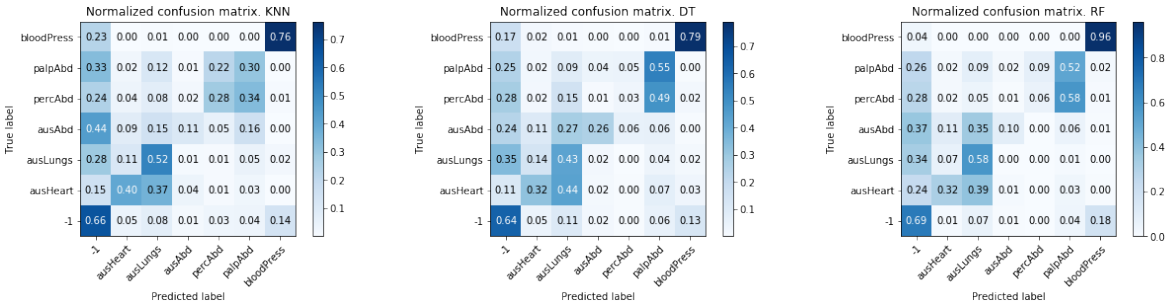


Figure D.2: CMs for the three classifiers on feature set 2.

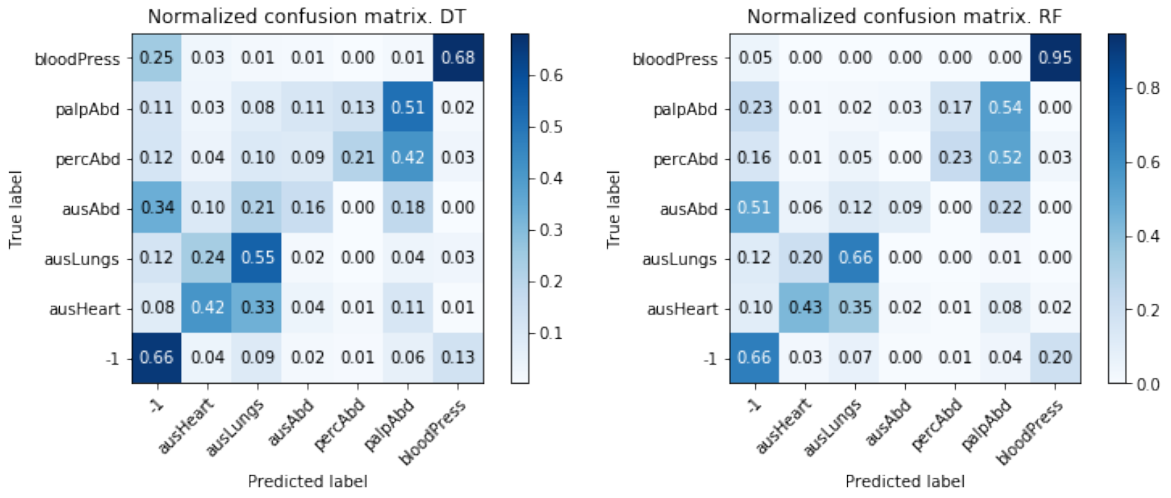


Figure D.3: CMs for the DT and RF classifiers on feature set 3.

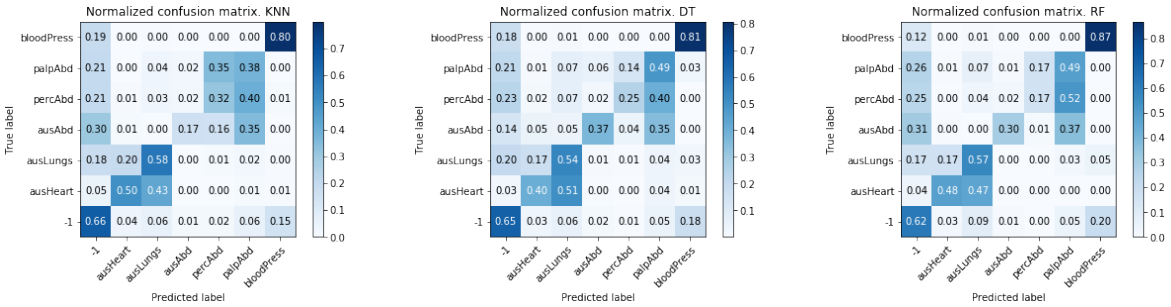


Figure D.4: CMs for the three classifiers on feature set 4.

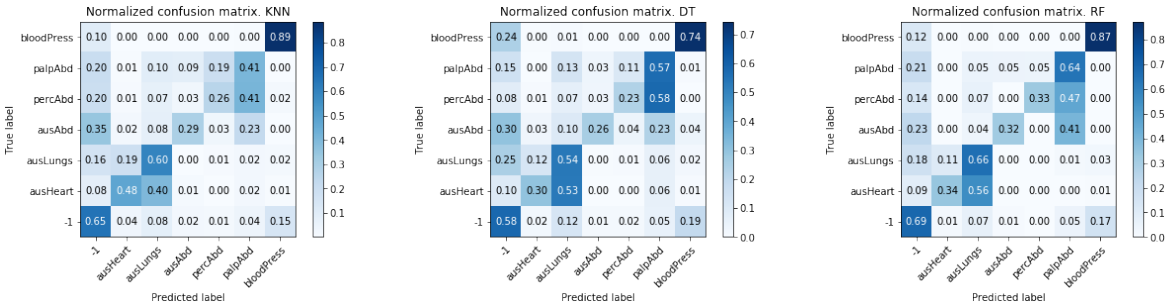


Figure D.5: CMs for the three classifiers on feature set 5.

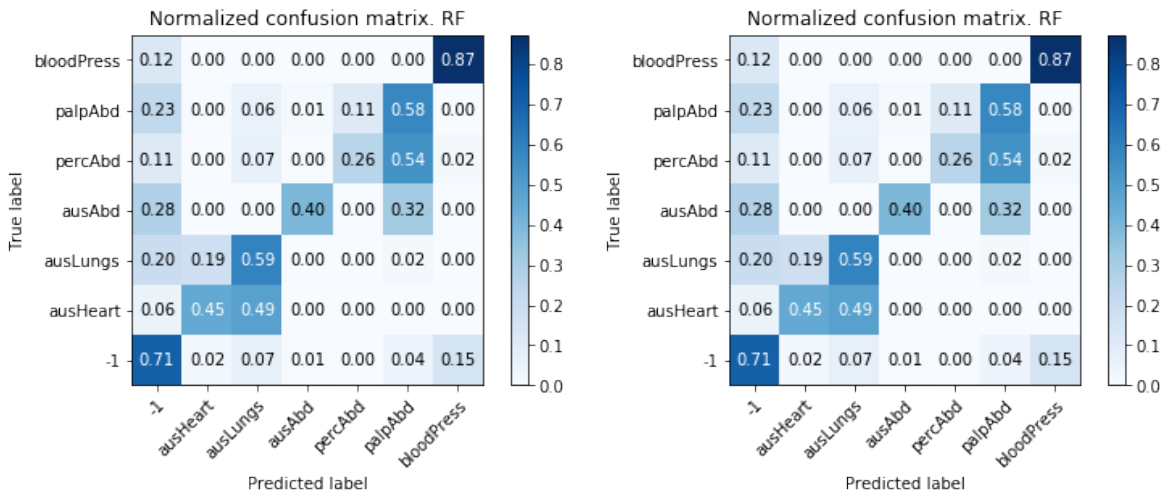


Figure D.6: CMs for the two classifiers on feature set 4 and 5.

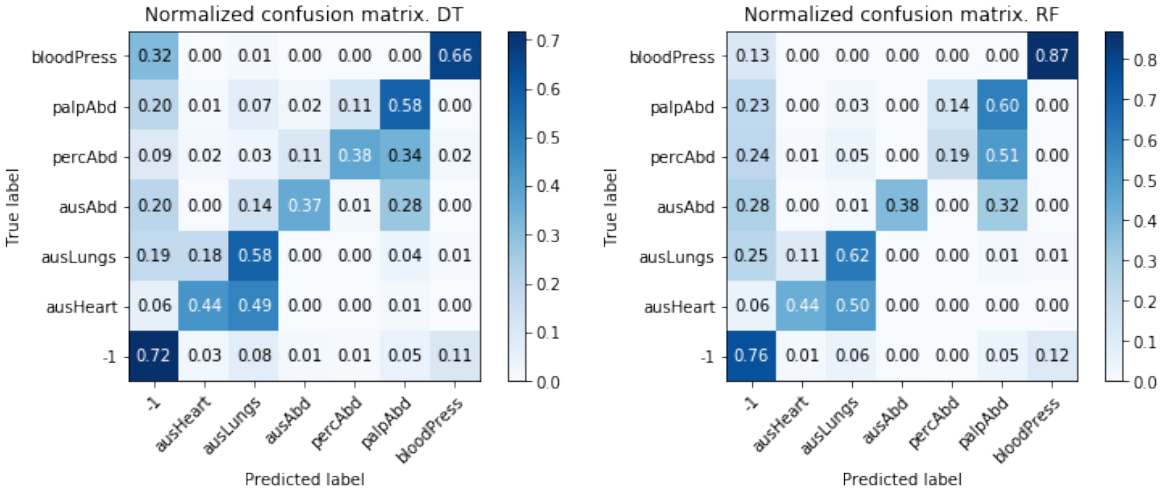


Figure D.7: CMs for the two classifiers on feature set 2, 4, and 5.

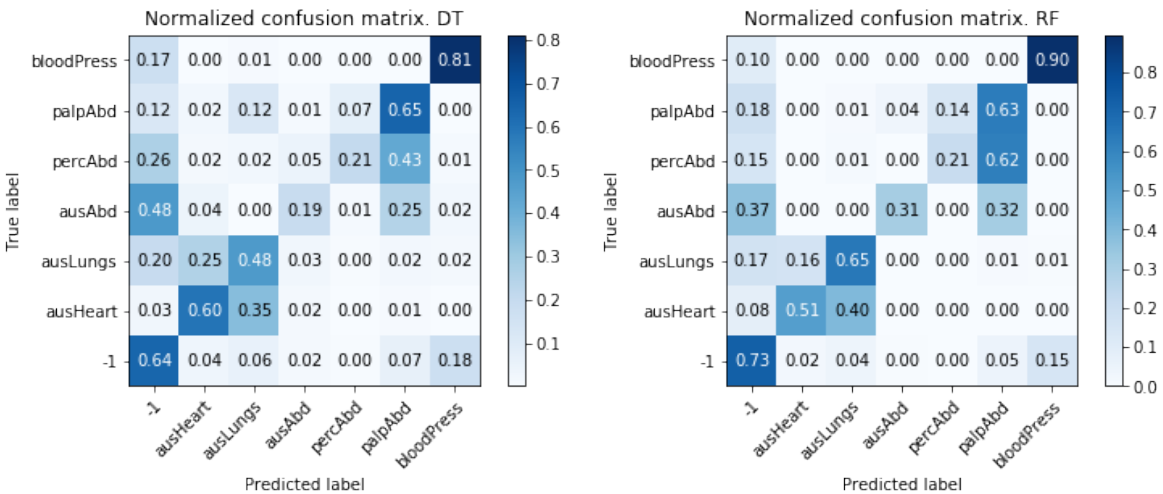


Figure D.8: CMs for the two classifiers on feature set 3, 4, and 5.

APPENDIX E: CMS for the DT, RF, and k-nn classifiers for various feature sets, tested on Area of Investigation

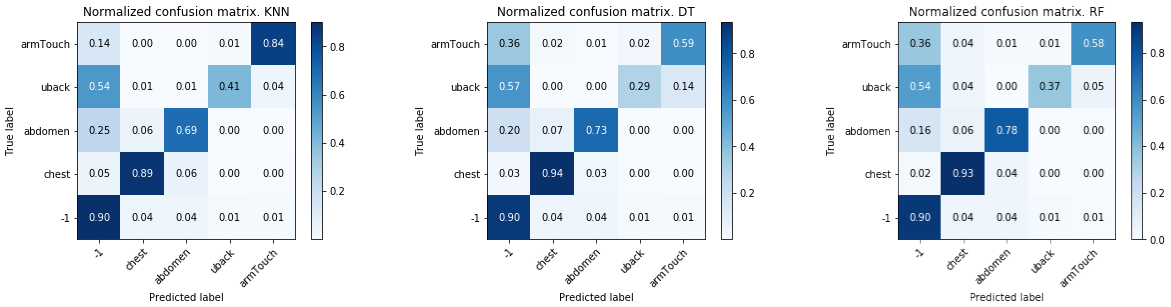


Figure E.1: CMs for the three classifiers on feature set 1.

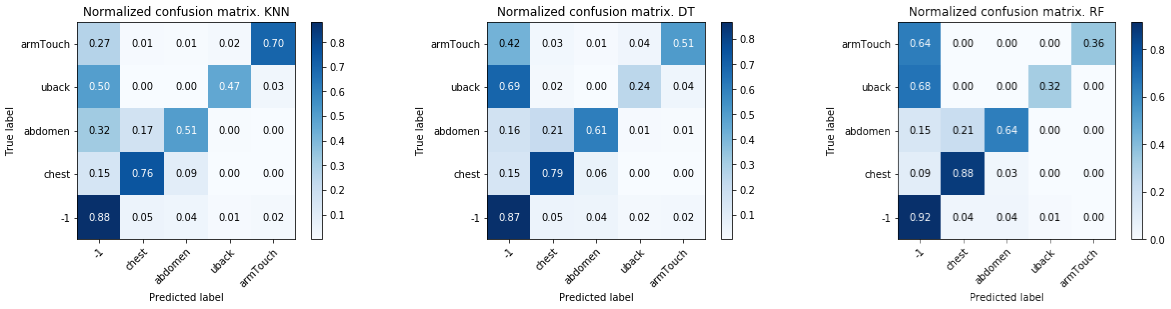


Figure E.2: CMs for the three classifiers on feature set 2.

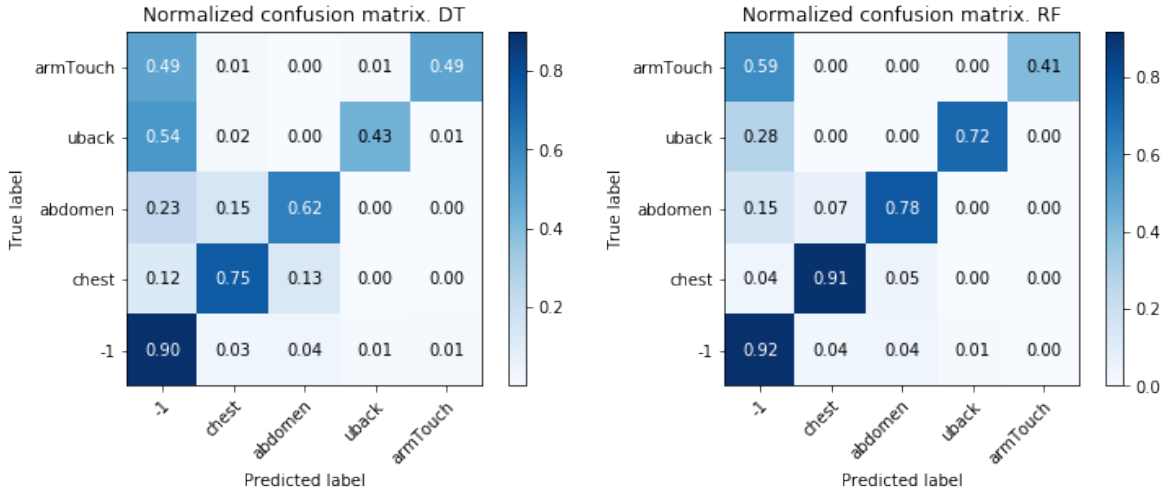


Figure E.3: CMs for the DT and RF classifiers on feature set 3.

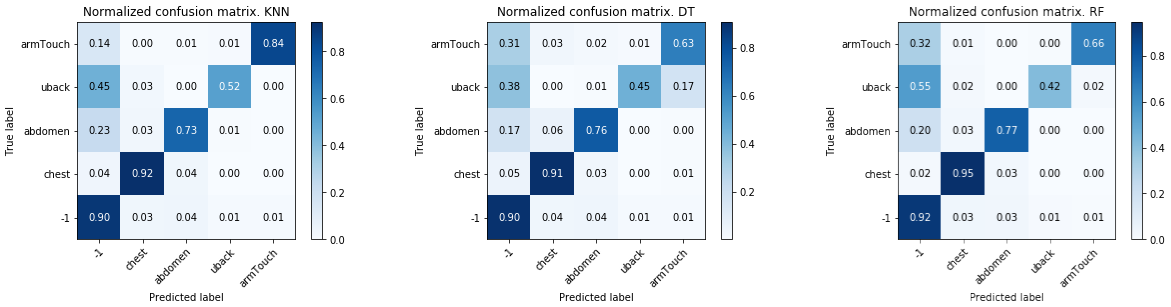


Figure E.4: CMs for the three classifiers on feature set 4.

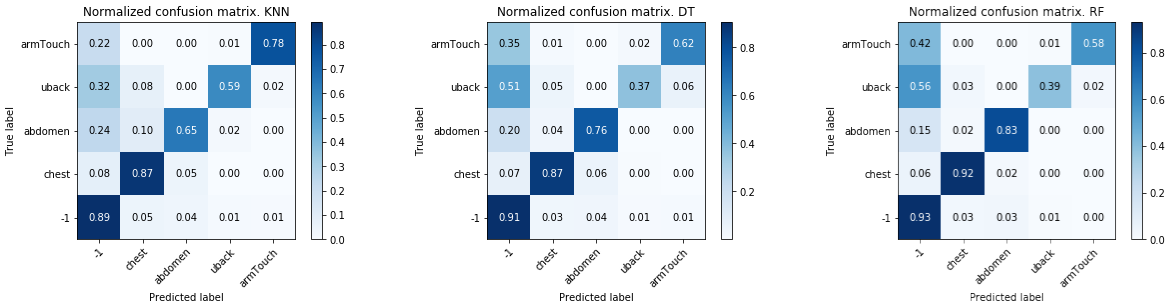


Figure E.5: CMs for the three classifiers on feature set 5.

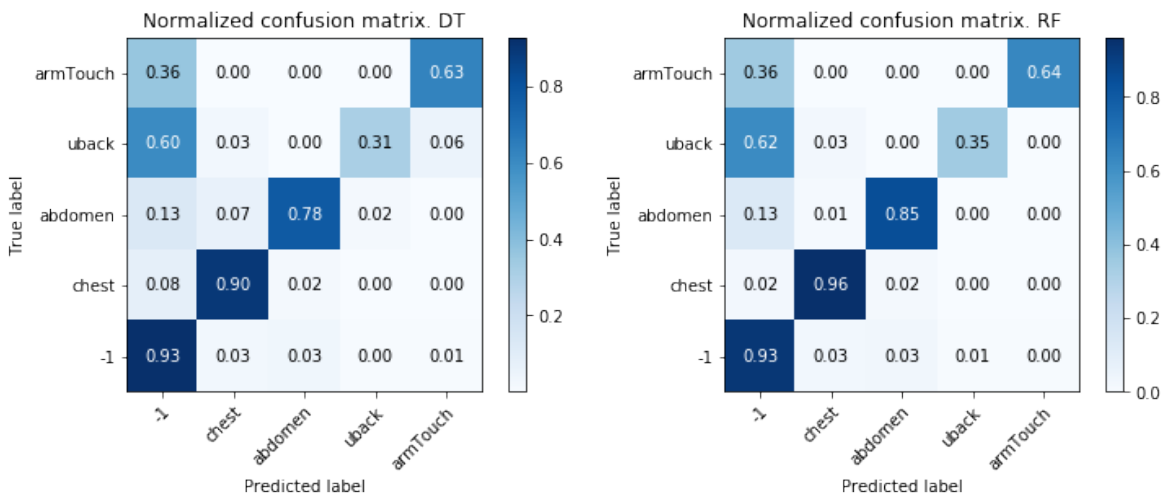


Figure E.6: CMs for the two classifiers on feature set 4 and 5.

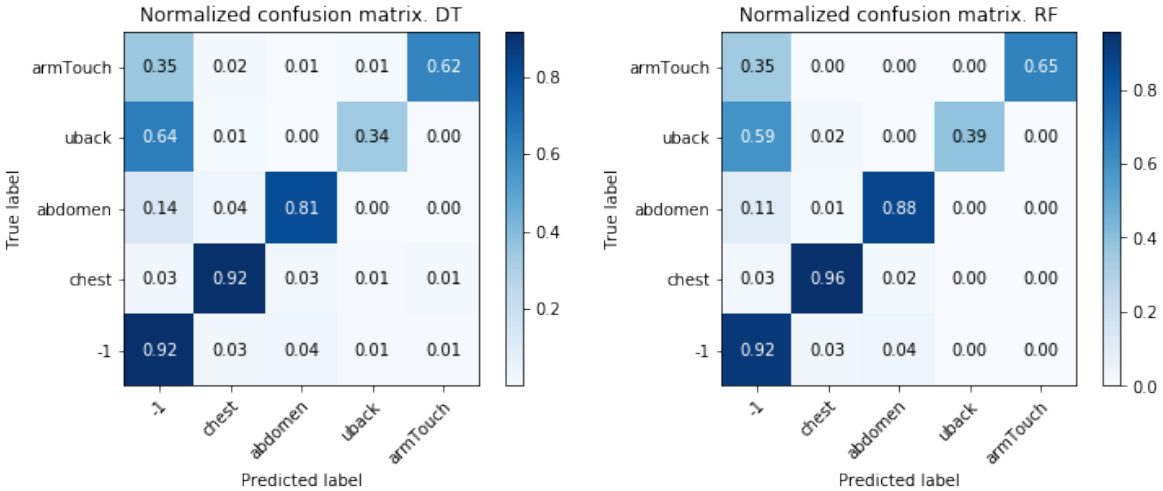


Figure E.7: CMs for the two classifiers on feature set 2, 4, and 5.

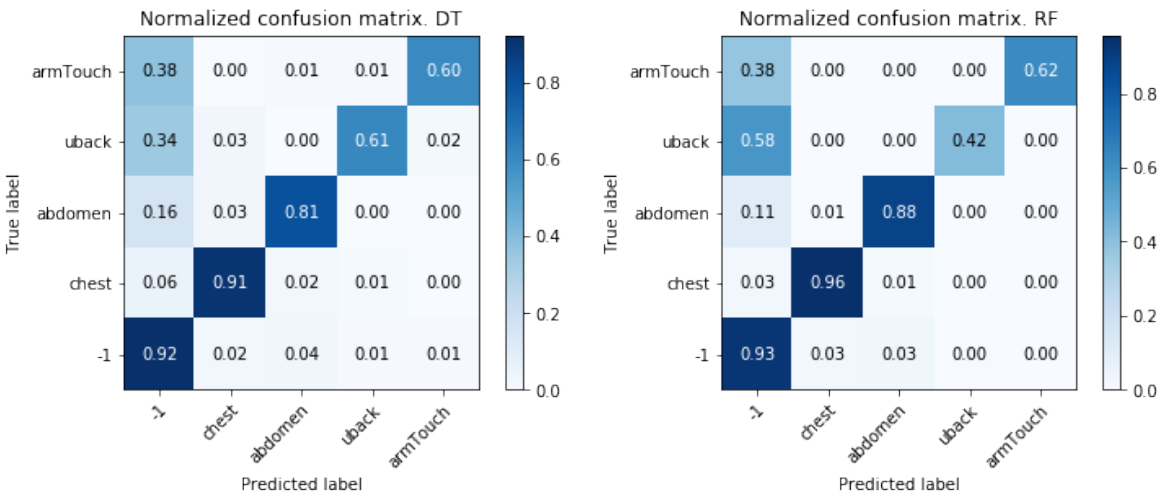


Figure E.8: CMs for the two classifiers on feature set 3, 4, and 5.