Hand pose estimation and movement analysis for occupational therapy

Luciano Walenty Xavier Cejnog

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Doctor of Science

Program: Computer Science
Advisor: Prof. Dr. Roberto Marcondes Cesar Jr.
Coadvisor: Prof. Dr. Teófilo Emidio de Campos

During this work, the author was supported by FAPESP.

São Paulo
December 14th, 2021

Hand pose estimation and movement analysis for occupational therapy

Luciano Walenty Xavier Cejnog

This version of the thesis includes the corrections and modifications suggested by the Examining Committee during the defense of the original version of the work, which took place on December 14th, 2021.

A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

- Prof. Dr. Roberto Marcondes Cesar Junior - IME/USP
- Prof. Dr. Janko Calic - BBC
- Prof. Dr. Paulo Andre Vechiatto de Miranda - IME/USP
- Profª. Drª. Valéria Meirelles Carril Elui - FMRP/USP
- Profª. Drª. Fátima Nelsizeuma Sombra de Medeiros - UFC

Ao amigo Victor Calixto de Souza (1989-2017).

Acknowledgements

I would like to thank my family, my mother **Eliana Xavier Cejnog**, my father **Walenty Cejnog**, my brothers **Pedro Walenty Xavier Cejnog** and **Bruno Walenty Xavier Cejnog**, for the unconditional support. All I do is for you and I love all of you so much.

I also would like to thank my advisor **Roberto Marcondes Cesar Jr.**, for always showing the paths, trusting me and giving the autonomy and providing always the support needed for the development of the work; to my co-advisor **Teófilo Emídio de Campos** for the closer orientation, support on the writing and on the development of the proposed method; to Professor **Valéria Meirelles Carril Elui** for the support on the planning and execution of the data acquisition experiments at FMRP-USP and for providing study references in the research area of hand occupational therapy.

I also thank CAPES and FAPESP [1] for the financial funding, University of São Paulo for the lessons and courses, and for the support in terms of infrastructure and tools needed for the execution of the project, and my company Data Machina for the total support in the finalization of the PhD thesis, allowing adaptation of the work time.

I am also very grateful to my friends for the continuous support ant encouragement, this was all very important. Each talk, each dialog, each moment was essential. There were a lot of times when I thought on giving up, and the strength I took on those moments was the reason I have not. Thank you very much and I hope I have not forgotten anyone.

Aderaldo Alexandre, Aida Camarini, Alisson do Rosário, André Muchon, Anita Marzolla, Arthur Ferguson, Beatriz Furtado, Bernardo Arêas, **Brune Coelho**, Caio Rodrigues, **Catarina Angeli**, Cely Freitas, **Clara Penz**, Daniel Hiroki Yamashita, **Daniela Carrini**, **Dedimar Dias**, Dharana Autran, Diego Ramalho, Douglas Oliveira, Eduardo Tita, **Emília França**, Eric Keiji, Erick Mendes, Estephan Dazzi, Evelyn Cervantes, Fernando Akio, Fernando Marques, Flávio Silva, Fábio Marinho, **Gabriel Bonz**, Gabriel Xavier, **Guilherme de Camargo**, Guilherme Martins, Guilherme Xavier, **Heitor Fernandes**, Helena Maia,

[1] FAPESP processes #16/13791-4, #15/22308-2, #14/50769-1

Henrique Vitoi, Henrique Xavier, Iago Araújo, **Igor Cataneo Silveira**, Iuri Carvalho, **Ivo Pons**, **Jean Alves**, **Jefferson Miranda**, **Jessica Dias**, **Jéssica Gasparoni**, Joseph Hans Murrugara, **Juliana Bellini**, Juliana Maruyama, Júlia Angeli, Larissa Penteado, Laura Lima, Leissi Castañeda, Leonardo Menezes, Leonardo Oliveira, Liliane Almeida, Lucas Dias, **Lucas Ferraz**, Lucas Nunes, **Luiza Helena Zancanella**, **Marcela Okuyama**, **Marcelo Fernandes**, Marcelo Machado, **Márcio Cabral**, Marcos Silva, Marcos Xavier, Marcus Vinícius da Silva, Marcus Vinícius Vasconcelos, **Mateus Riva**, Melisa Paiba, Michele Oliveira, Mitsuo Koza, Natália Lopes, Nury Arequispa, Patrick Bono, Paulo Vitor Freitas, Pedro Ivo Lancelotta, Pedro Mendonça, Phillipe Israel, Rafael Alves, Rafael Barreto, Rafael Pocai, **Ramon Nogueira**, **Raphael Ferreira**, Raúl Rincón, Rita Xavier, Roberto Bodo, Rodrigo de Almeida, **Rodrigo Masaru Ohashi**, Rodrigo Pontes, Rosa Clara Alves, Samuel Ferreira, Sandra Ferreira, Sandra Silva, Stephania Falcão, Tacila Rocha, Tércio Garcia, **Vanessa Salmazo**, **Vitor Rodrigues Costa**, **Vitor Stipp**, Wilken Sobral, Yan Mendes, Yuri Monteiro.

Agradecimentos

Agradeço à minha família, minha mãe **Eliana Xavier Cejnog**, meu pai **Walenty Cejnog**, meus irmãos **Pedro Walenty Xavier Cejnog** e **Bruno Walenty Xavier Cejnog**, pelo apoio incondicional. Tudo que eu faço é por vocês e amo vocês demais.

Agradeço ao meu orientador **Roberto Marcondes Cesar Jr.** por sempre mostrar caminhos, por confiar em mim e me dar autonomia e totais condições para o desenvolvimento do trabalho; ao meu co-orientador **Teófilo Emídio de Campos** pela orientação próxima, suporte na escrita e no desenvolvimento do método proposto. Agradeço à professora **Valéria Meirelles Carril Elui** pelo apoio total no planejamento e execução da aquisição de dados na FMRP-USP e fornecimento de referências de estudo na área de terapia ocupacional de mão.

Agradeço à CAPES e à FAPESP [2] por proverem fomento ao trabalho. À USP por prover o aprendizado necessário nas disciplinas e as ferramentas necessárias para execução do projeto. À Data Machina pelo total apoio na parte final desse doutorado, permitindo adaptação no horário de trabalho.

Agradeço também aos amigos que sempre me apoiaram, encorajaram e foram muito importantes. Cada conversa, palavra trocada e momento passado juntos foram essenciais. Muitas vezes pensei em desistir e todos vocês foram importantes e me deram forças nesses momentos. Muito obrigado e espero não ter esquecido de ninguém!

Aderaldo Alexandre, Aida Camarini, Alisson do Rosário, André Muchon, Anita Marzolla, Arthur Ferguson, Beatriz Furtado, Bernardo Arêas, **Brune Coelho**, Caio Rodrigues, **Catarina Angeli**, Cely Freitas, **Clara Penz**, Daniel Hiroki Yamashita, **Daniela Carrini**, **Dedimar Dias**, Dharana Autran, Diego Ramalho, Douglas Oliveira, Eduardo Tita, **Emília França**, Eric Keiji, Erick Mendes, Estephan Dazzi, Evelyn Cervantes, Fernando Akio, Fernando Marques, Flávio Silva, Fábio Marinho, **Gabriel Bonz**, Gabriel Xavier, **Guilherme de Camargo**, Guilherme Martins, Guilherme Xavier, **Heitor Fernandes**, Helena Maia, Henrique Vitoi, Henrique Xavier, Iago Araújo, **Igor Cataneo Silveira**, Iuri Carvalho, **Ivo**

[2] Processos FAPESP #16/13791-4, #15/22308-2, #14/50769-1

Pons, **Jean Alves**, **Jefferson Miranda**, **Jessica Dias**, **Jéssica Gasparoni**, Joseph Hans Murrugara, **Juliana Bellini**, Juliana Maruyama, Júlia Angeli, Larissa Penteado, Laura Lima, Leissi Castañeda, Leonardo Menezes, Leonardo Oliveira, Liliane Almeida, Lucas Dias, **Lucas Ferraz**, Lucas Nunes, **Luiza Helena Zancanella**, **Marcela Okuyama**, **Marcelo Fernandes**, Marcelo Machado, **Márcio Cabral**, Marcos Silva, Marcos Xavier, Marcus Vinícius da Silva, Marcus Vinícius Vasconcelos, **Mateus Riva**, Melisa Paiba, Michele Oliveira, Mitsuo Koza, Natália Lopes, Nury Arequispa, Patrick Bono, Paulo Vitor Freitas, Pedro Ivo Lancelotta, Pedro Mendonça, Phillipe Israel, Rafael Alves, Rafael Barreto, Rafael Pocai, **Ramon Nogueira**, **Raphael Ferreira**, Raúl Rincón, Rita Xavier, Roberto Bodo, Rodrigo de Almeida, **Rodrigo Masaru Ohashi**, Rodrigo Pontes, Rosa Clara Alves, Samuel Ferreira, Sandra Ferreira, Sandra Silva, Stephania Falcão, Tacila Rocha, Tércio Garcia, **Vanessa Salmazo**, **Vitor Rodrigues Costa**, **Vitor Stipp**, Wilken Sobral, Yan Mendes, Yuri Monteiro.

Resumo

Luciano Walenty Xavier Cejnog. **Estimativa de pose de mão e análise de movimento no contexto de terapia ocupacional**. Tese (Doutorado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Estimativa de pose de mão é um problema considerado complexo dentro da área de visão computacional com uma vasta gama de aplicações, especialmente na área de interface humano-computador. Com a evolução do estado-da-arte em técnicas de aprendizado profundo e com a popularização de sensores $3D$ de baixo custo, o estado-da-arte atual do problema vem se atualizando continuamente e muitos métodos novos têm sido propostos nos últimos anos. Esses métodos em sua maioria são baseados no uso de grandes volumes de dados para treinamento, e alcançam resultados cada vez melhores nas bases de dados padronizadas, como NYU, ICVL e HANDS17. Uma das aplicações que se beneficiaria do uso de visão computacional é a terapia ocupacional de mão. Por exemplo, em doenças crônicas como a artrite reumatoide (AR), a avaliação do estado funcional do paciente é fundamental para o tratamento bem como para a prevenção de deformidade dos dedos. Um dos procedimentos para o diagnóstico das deformidades dos dedos é a medição dos ângulos de movimento, por exemplo a flexão/extensão e abdução/adução dos dedos, feita por um goniômetro em um processo simples, mas que pode ser invasivo e demorado para o paciente. Esta tese busca preencher uma lacuna do estado-da-arte ao propor e avaliar a viabilidade da utilização de um arcabouço composto de um sensor $3D$ de baixo custo e uma técnica estado-da-arte em estimativa de pose de mão $3D$ para aquisição automática dos ângulos da mão em pacientes de artrite reumatoide. O algoritmo proposto é aplicado em um conjunto de imagens de profundidade, retornando a posição das juntas da mão estimadas a partir de uma rede neural convolucional profunda. O algoritmo utilizado pode ser executado em tempo real, permitindo a visualização dos esqueletos resultantes ao mesmo tempo em que as imagens são adquiridas. A partir dessa estimativa, os ângulos de flexão/extensão e de abdução/adução da mão são calculados aplicando operações de geometria computacional. A dificuldade em se encontrar bases de dados relativas a pessoas com AR torna a estimativa de poses de mão dos pacientes um desafio ainda maior para os métodos de visão computacional baseados em dados. Dessa forma, foi proposto um protocolo de aquisição de dados para grupos de pacientes e controle. Foram feitos experimentos de comparação com os dados do goniômetro dos acometidos pela AR. Os resultados mostram que é possível distinguir automaticamente os conjuntos de acometidos e controle usando descritores de Fourier. Os ângulos mensurados pelo sensor podem ser usados como indicativo das capacidades de movimento dos pacientes. O procedimento é simples, não invasivo e mais amigável para os

acometidos pela AR, reduzindo o tempo de avaliação além de oferecer dados em tempo real do movimento dinâmico.

Palavras-chave: Estimativa de pose de mão. Visão computacional. Terapia ocupacional.

Abstract

Luciano Walenty Xavier Cejnog. **Hand pose estimation and movement analysis for occupational therapy**. Thesis (Doctorate). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Hand pose estimation is a challenging problem in computer vision with a wide range of applications, especially in human-computer interface. With the development of inexpensive consumer-level depth cameras and the evolution on deep learning techniques, the current state-of-art in the problem is continuously developing and several new methods have been proposed in recent years. Those methods are mostly data-driven and reach good results in standard datasets such as NYU, ICVL and HANDS17. An application that would benefit from the use of computer vision techniques is hand occupational therapy. In chronic diseases like rheumatoid arthritis (RA), the evaluation of the hand functional state is fundamental for the treatment and prevention of finger deformities. One of the procedures for deformity diagnosis is the measurement of movement angles i.e. flexion/extension and abduction/addution, made using goniometers in a process that can be time-consuming and invasive for the patient. The main proposal of this PhD is to fill a gap in the literature by proposing and evaluating the viability of using a framework composed of a $3D$ low-cost sensor and a $3D$ hand pose estimation state-of-art method for automatic assessment of rheumatoid arthritis patients. Given depth maps as input, our framework estimates 3D hand joint positions using a deep convolutional neural network. The proposed pose estimation algorithm can be executed in real-time, allowing users to visualise 3D skeleton tracking results at the same time as the depth images are acquired. Once 3D joint poses are obtained, our framework estimates flexion/extension and abduction/adduction angles by applying computational geometry operations. The absence of public datasets with RA patients in the literature makes the estimation of hand poses of patients a challenge for computer vision data-driven methods. We therefore proposed a protocol to acquire new data from groups of patients and control. We performed experiments of identification of RA patients and control sets and also performed comparison with goniometer data. Results show that a method based on Fourier descriptors is able to perform automatic discrimination of hands with Rheumatoid Arthritis (RA) and healthy patients. The angle between joints can be used as an indicative of current movement capabilities and function. The acquisition is much easier, non-invasive and patient-friendly, significantly reducing the evaluation time and offering real-time data for the dynamic movement.

Keywords: Hand pose estimation. Computer vision. Occupational therapy.

List of acronyms

RA	Rheumatoid arthritis
DASH	Disabilities of the Arm, Shoulder and Hand
ROM	Range of Motion
RGB	Red Green Blue color system
PSO	Particle Swarm Optimization
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
VAE	Variational AutoEncoder
R200	Intel RealSense® R200
SR300	Intel RealSense® SR300
GPU	Graphics Processing Unit
CUDA	Compute Unified Device Architecture
SVM	Support Vector Machine
RBF	Radial Basis Function Kernel
QDA	Quadratic Discriminant Analysis
MCP	Metacarpophalangeal hand joint
PIP	Proximal interphalangeal joint
DIP	Distal interphalangeal joint
F-II-PIP	Flexion angle associated with the PIP joint on finger II
FFT	Fast Fourier Transform
TAM	Total Active Motion
LOO	Leave-one-person-out experiment
SS	Sample synthesis
Fourier	Descriptor based on Fourier Transform
Baseline	Descriptor based on maximum and minimum values of each angle
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

List of symbols

$\mathcal{D}$	Depth image
$\vec{S}(t)$	Skeleton at frame t
MCP_f	Metacarpophalangeal joint for finger f (illustrated in Figure 1.5)
PIP_f	Proximal interphalangeal joint for finger f (illustrated in Figure 1.5)
DIP_f	Distal interphalangeal joint for finger f (illustrated in Figure 1.5)
fingertip_f	Tip joint for finger f (illustrated in Figure 1.5)
W	Wrist joint (illustrated in Figure 1.5)
CMC	Carpometacarpal thumb joint (as shown in Figure 1.5)
$\widehat{\text{F-x-MCP}}$	Flexion angle for joint MCP_x
$\widehat{\text{F-x-PIP}}$	Flexion angle for joint PIP_x
$\widehat{\text{F-x-DIP}}$	Flexion angle for joint DIP_x
A-x-tip	Abduction distance between tips of fingers x and $x-1$
$\vec{A} = a_i(t)$	Angle representation of a clip as a set of functions a_i represents the ith. angle in frame t.
$\mathcal{F}_{a_i}$	Fourier transform of the angle representation a_i.
$\mathcal{F}$	Concatenation of Fourier transforms of all angles a_i that compose the angle representation $\vec{A}$ of the clip.
$\mathcal{N}(\mu, \sigma)$	Gaussian distribution with mean μ and standard deviation σ.
$ROM(j)$	Range of Motion of a joint j
$TAM(f)$	Total Active Motion of a finger f.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

This chapter defines the problem addressed and the goals of this thesis. Section 1.1 contextualizes the problem of using computer vision to help the evaluation of hand range of motion in patients with rheumatoid arthritis. Section 1.2 formalizes the problem definition as an investigation of computer vision techniques for hand range of motion evaluation. Section 1.3 describes the main goal, Section 1.4 details the main contributions of the thesis, and Section 1.5 provides an outline of how the rest of the thesis is organized.

1.1 Motivation

Hand pose estimation is an important task in the computer vision field, with several applications in areas such as human-computer interface, augmented reality, sign language recognition and robotics. It is a very challenging problem due to the high dimensionality of the hand structure, self-occlusions and ambiguities on the model and the similarity between the fingers. With the recent development of consumer-level 3D depth cameras and advances in computer vision and deep learning, some of those problems are mitigated and more robust and accurate methods are being presented in each major conference and journal of the area. Different fields of application could benefit from the latest advances on hand pose estimation.

An important field that was not explored in details by the computer vision community (Meals *et al.*, 2018) is hand surgery recovery and occupational therapy, in particular in the treatment of Rheumatoid Arthritis (RA). RA is an autoimmune chronic disease that leads to joint deformities due to an inflammation that causes the erosion of tissues, including bones. This inflammatory mechanism was discovered very recently (Donate *et al.*, 2021). Findings of population-based studies show RA affects 5 to 10% of adults in developed countries. The disease is three times more frequent in women than men, and 50% of risk of developing RA is attributable to genetic factors (Scott *et al.*, 2010). The clinical complaints include pain, swelling and motion limitations of the affected joints. A physical examination will reveal the presence of pain, increased joint volume, intra-articular effusion (presence of intra-articular fluid), heat and eventual redness. Recent advances in occupational therapy research indicate that the first 12 months with RA symptoms stand out as an acknowledged

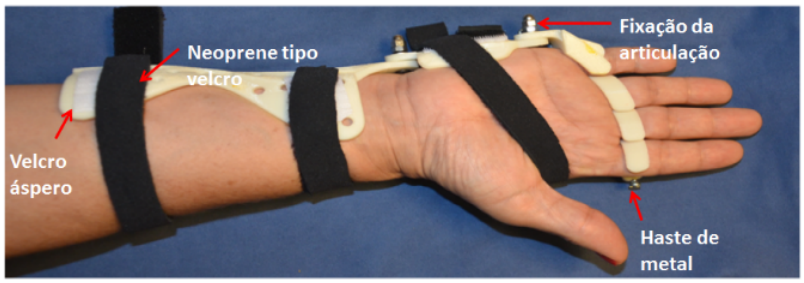
"window of therapeutic opportunity" (Mota *et al.*, 2013). Therefore, identifying the disease in its early stages is fundamental in preventing its progression.

Figure 1.1 shows an example of hand with rheumatoid arthriris and ulnar deviation in contrast with a normal hand.

Figure 1.1: *Example of hand with ulnar deviation from a patient on hand flexor tendon surgery recovery (on the right), in contrast with a normal hand of the same patient (on the left). Courtesy of Prof. Valeria Elui.*

One common step in the treatment of rheumatoid arthritis is the design of orthoses for injured hands. Orthoses are external devices applied to any part of the body to stabilize it or immobilize it, prevent or correct deformities, protect against injury, maximize function and reduce the pain caused by deformity (Goia *et al.*, 2017). Orthoses are tailor-made by therapists and for the hand case act like a lever system distributing the force applied to the ulnar deviation. Figure 1.2 illustrates and presents details about hand orthoses used on the patients that participated of the dataset formation.

The evaluation of hand function is fundamental for the therapist to plan the treatment as well as record the results. Literature in hand therapy define metrics and guidelines in order to extract those metrics with precision (Marques, 1997). A widely used metric for measuring the joint angles is range of motion (ROM). The range of motion is defined as the quantity of movement of an articulation. Active range of motion refers to movement without interference of external factors, providing information about the capacity, coordination and muscular power of the patient. Passive range of motion refers to movement only by external factors, and it is used to verify the integrity of articular surface and the extensibility of the articular capsule (Norkin and White, 1997). It is conventioned by



(a) *Parts of an orthosis. Courtesy of Prof. Valéria Elui.*

Figure 1.2: *Example of an orthosis used on the hand, tailor-made devices made to distribute the force and leverage the effects of rheumatoid arthritis.*

occupational therapists that the erect anatomical posture corresponds to 0° of movement. Thus, the maximal amplitude value for each articulation is 180°. In this thesis, we will focus on measuring *active range of motion for hand articulations*, using the movement patterns of flexion/extension and abduction/adduction. Figures 1.3 and 1.4 show examples of flexion and abduction movements recorded from control and patients.

Table 1.1 shows the minimal and maximal amplitude considered normal for the hand angles, according to Marques (1997). Figure 1.5 shows the corresponding joints for a hand model.

Joint	Movement	Min Amplitude	Max Amplitude
Metacarpophalangeal (MCP)	Flexion	0°	90°
	Extension	0°	30°
	Abduction	0°	30°
	Adduction	0°	20°
Interphalangeal (PIP and DIP)	Flexion	0°	110°
	Extension	0°	10°
Thumb CMC	Flexion	0°	15°
	Extension	0°	70°
	Abduction	0°	70°

Table 1.1: *Standard measurements for joint angles in goniometry (Marques, 1997).*

Typically, *Disabilities of the Arm, Shoulder and Hand* (*DASH*) questionnaires (Orfale *et al.*, 2005) are used to assess hand function during the recovery process. This evaluation method is based on the patient qualitative self-evaluation of difficulty in the execution of daily activities, such as writing, preparing a meal or making a bed. Quantitative evaluation of the hand function is usually assessed by active range of motion measurements. The standard procedure for this evaluation is the use of a goniometer. With a specific hand/finger goniometer, as exemplified in Figure 1.6, the therapist can access objectively and reliably the range of motion measurements. Such devices are widely used due to their simplicity and low cost. The procedure, however, requires a trained therapist that follows the protocols, is time consuming and requires a careful setup and patient positioning.

Although the manual goniometer is a widely used device for assessing hand angles, the literature has explored alternatives for automatizing the measurement procedure.

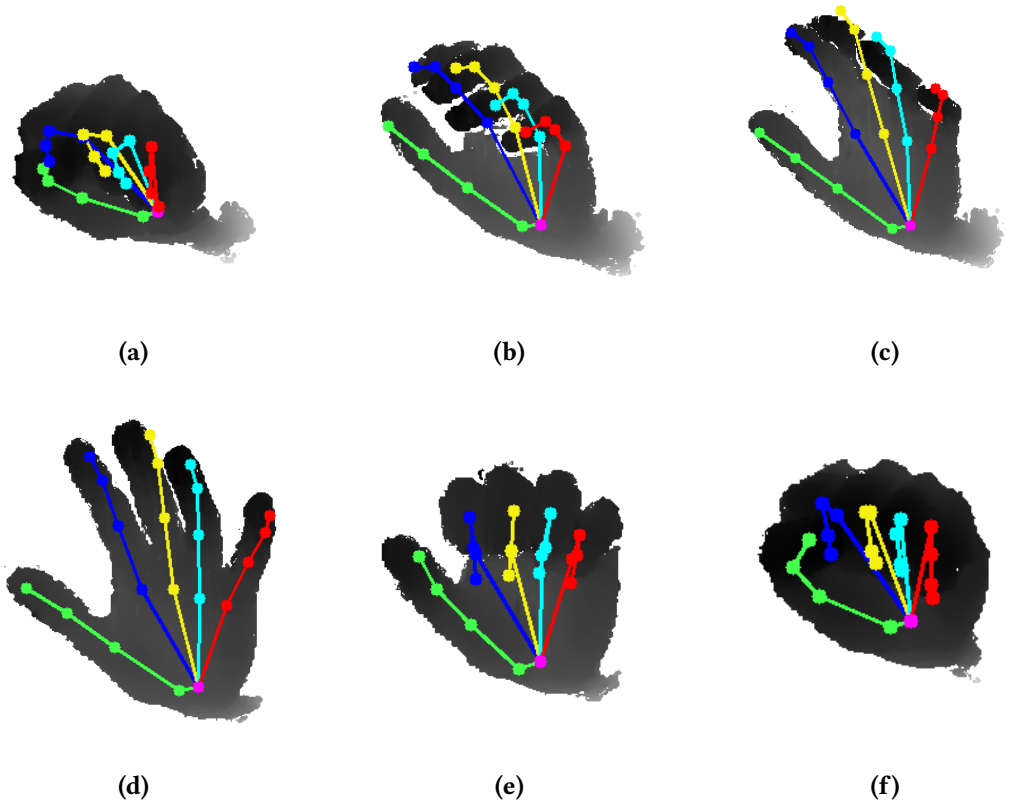


Figure 1.3: *Example of extension/flexion movement. Frames were recorded using the* Intel RealSense® SR300 *sensor, and a hand pose estimation algorithm was applied in order to provide the hand joints.*

One possibility for automatizing range of motion measurements is the use of electric sensors (Tajali *et al.*, 2016; Gutiérrez-Martínez *et al.*, 2014). The results presented show that those devices are reliable and obtain measurements highly correlated to the ones obtained by manual goniometry, but the technology is still expensive and of limited distribution.

Another idea that has been explored is digital photogrammetry, which consists in the determination of angles in hand images. This approach was mostly used by surgeons, and some recent works indicate that the reliability of this method has increased over the years (Carvalho *et al.*, 2012). However, viability studies in the literature (Ellis *et al.*, 1997; Bruton *et al.*, 1999; Meals *et al.*, 2018) show that the use of digital photogrammetry has limited reliability and precision for measuring hand joint angles in comparison to the manual goniometry. According to Meals *et al.* (2018), one of the main limitations of this approach is that the result is not immediately assessed: joints must be photographed and then measured. This work indicates future possibilities of using $3D$ scanning and video capture technology to the development of an automatic goniometer for the hand. Most state-of-the-art hand pose estimation methods, including the Pose-REN method used in our pipeline (Chen *et al.*, 2019), present the possibility of assessing results in real-time, simplifying the automation of the process.

Among recent works that proposes solutions based on computer vision for occupational

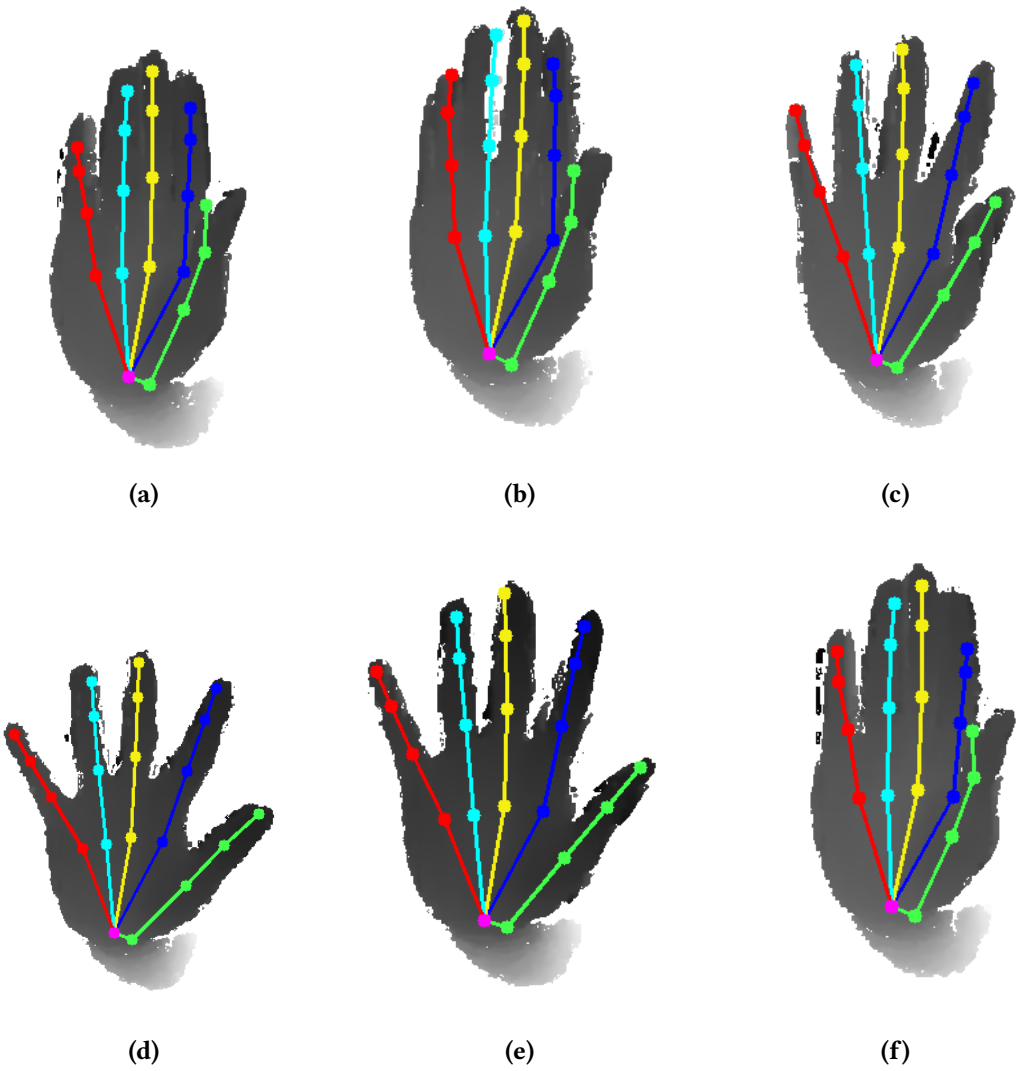


Figure 1.4: *Example of abduction/adduction movement. Frames were recorded using the* Intel RealSense® SR300 *sensor, and a hand pose estimation algorithm was applied in order to provide the hand joints.*

therapy in general, Pereira *et al.* (2017) proposes a smartphone accelerometer-based app to measure active and passive knee ROM in a clinical setting. The hand problem, however, is arguably more challenging than the knee, and 2D hand pose estimation results are still not reliable (see Section 2.4). An alternative is the use of depth sensors, and despite its recent rise of popularity few works to date make use of such devices for this task. We highlight the work of Lima *et al.* (2016), that uses information obtained by a Leap Motion sensor to estimate hand angles. Leap Motion is a sensor developed for hand tracking in the context of Human-Computer interaction and is composed by three $2D$ cameras and two monochromatic infra-red sensors. This sensor was tested in our pipeline as input for hand analysis. However, we chose not to use it because the black-box hand pose estimation algorithm presented in the SDK is not suitable for hands with ulnar deviation and did not yield feasible results to practical use in our preliminary setup tests (see more in Section 3.1).

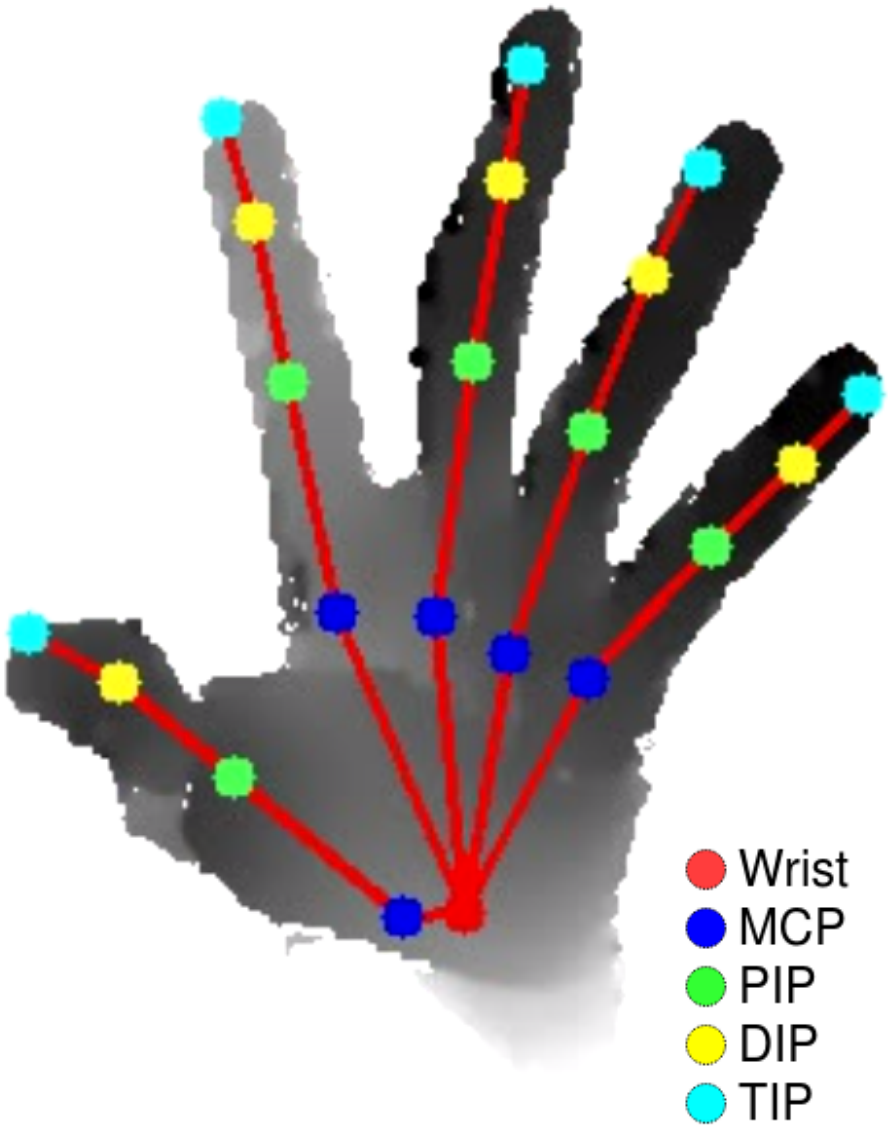
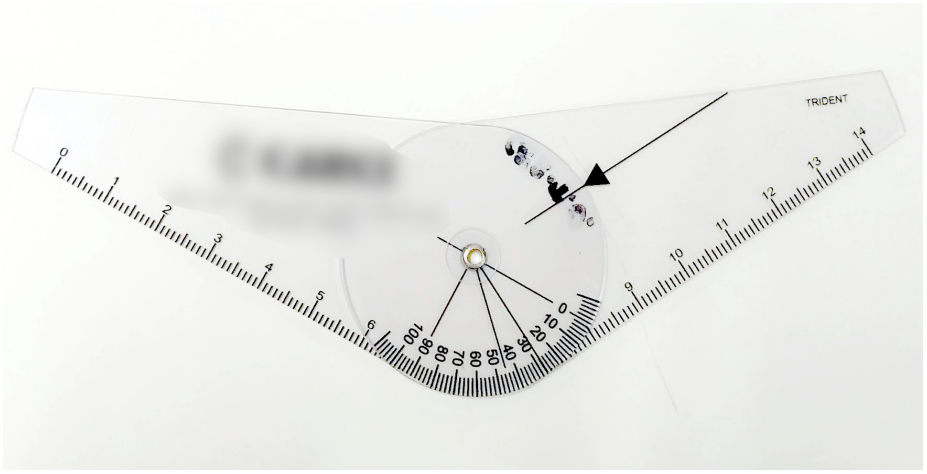
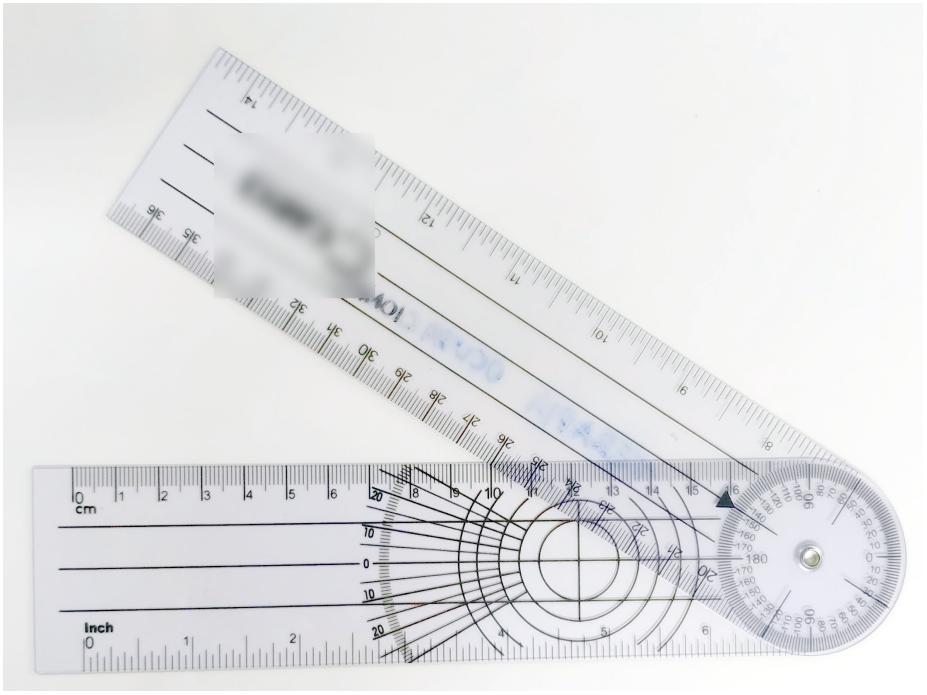


Figure 1.5: *Identification of hand joints. This figure was produced using the Intel Realsense® SR300 sensor, with real data from a hand with Rheumatoid Arthritis and an orthosis. The joints follow the hand model used in the HANDS17 dataset.*

This thesis seeks to evaluate the possibilities of using state-of-art computer vision-based systems to assist the therapists in diagnosing rheumathoid arthritis. This PhD research is part of the project "Hand tracking for occupational therapy" (proc. FAPESP 14/50769-1), that aims to study computer vision techniques capable of providing support to hand flexor tendon surgery recovery. The project is a collaboration with Professor Teófilo E. Campos, from the Computer Science Department at the Universidade de Brasília (UnB), Professor Adrian Hilton, from the Centre for Vision, Speech and Signal Processing (CVSSP) of the University of Surrey, Professor Janko Calic, who was also at the CVSSP but moved to the BBC in 2015, Professor Maria da Graça Campos Pimentel, from Instituto de



(a)



(b)

Figure 1.6: *Examples of goniometers used for hand range of motion measurements (provided courtesy by Prof. Valéria Elui).*

Ciências Matemáticas e de Computação from USP, and Professor Valeria Meirelles Carril Elui, Faculdade de Medicina de Ribeirão Preto, USP.

1.2 Problem definition

The project aims to investigate the use of computer vision techniques to provide objective feedback to the patient and to produce quantitative evaluations about their hand movement function and evolution. Ideally, the framework should handle patients with rheumatoid arthritis and also healthy hands.

In particular, we detail the development of a new framework for hand data acquisition, pose estimation and analysis, that can be applied to Rheumatoid Arthritis patients. The camera-based framework can enhance the comfort of the patient and the efficiency during the range of motion angle assessments. The procedure is markerless, does not require setting up an environment or external structure to capture those measurements and uses state-of-art computer vision techniques.

For data acquisition, the objective is to create a dataset containing hand poses from a group of occupational therapy patients with hand problems and from a control group making the same set of movements (flexion and abduction). In hand pose estimation, the objective is to estimate 3D joint positions from a raw depth image, obtained from a depth sensor. For hand analysis, we seek to estimate flexion/extension and abduction/adduction measurements from the skeletons estimated with the algorithm, and compare it with the measurements obtained with a goniometer, which is a standard evaluation method for occupational therapy.

1.3 Goal

The main goal of the thesis is to contribute to the development of a computer vision-based framework for automatic hand range of motion measurements aiming to help therapists and patients with rheumatoid arthritis. This framework should use a setup for data acquisition that is simple enough for use in a therapeutic setting of an hospital or clinic. In this context, the specific goal of this work is to develop computer vision methods to estimate hand joint angles in sequences of depth images, evaluating finger movement patterns of flexion/extension and abduction/adduction.

1.4 Contributions

The main contributions of the thesis are listed below:

- We propose an original end-to-end hand pose estimation and finger movement analysis approach for occupational therapy;
- We apply a state-of-the-art hand pose estimation method (Chen *et al.*, 2019) for automatic finger range of motion evaluation of patients in hand occupational therapy;
- We propose hand movement analysis tools based on the estimated angles and range-of-motion measurements from skeletons - we describe these results in the publication "Hand range of motion evaluation for Rheumatoid Arthritis patients", presented in the *14th IEEE International Conference on Automatic Face and Gesture Recognition* (Cejnog, de Campos, Elui, and R. M. Cesar Jr., 2019);
- We propose a dataset acquisition protocol and report the main decisions, difficulties of the process and the final acquisition protocol;
- We present a new dataset of depth maps and hand tracking results obtained using from patients of Rheumatoid Arthritis being treated at the Hospital das Clínicas in the Faculdade de Medicina Ribeirão Preto / University of São Paulo;

- We perform experiments with the dataset, comparing with measurements obtained by goniometers. Although the accuracy of the comparison is limited, a simple Fourier Descriptor in the time series of angle measurements is capable of discriminating between patients and healthy subjects - these results were presented in the publication "A framework for automatic hand range of motion evaluation of rheumatoid arthritis patients", published in *Informatics and Medicine Unlocked* (Cejnog, de Campos, Elui, and Roberto Marcondes Cesar Jr., 2021).

1.5 Organization

The rest of this document is organized as follows: Chapter 2 presents a literature review on hand pose estimation and details the state-of-art method used in our framework. Chapter 3 details the data acquisition and dataset formation, while Chapter 4 describes the methods used for hand pose estimation and hand analysis. Chapter 5 presents experimental results, identifying patterns that define control and patient sets. That chapter also presents a comparison of measurements obtained by sensors with goniometer measurements. Chapter 6 presents concluding remarks.

Chapter 2

Bibliographical review on hand pose estimation

The literature on hand function evaluation for therapists, together with some related traditional techniques was discussed in Chapter 1. In this chapter we present a bibliographical review for hand pose estimation. We describe the methods subdivided in a historical cut: early methods (1997-2007), depth sensors (2011 - 2016), deep learning on depth maps (2016 to state-of-the-art) and deep learning on RGB images (2017 to state-of-the-art). We decided to treat image-based deep learning methods in a separated section because the nature of the solutions and the datasets used for comparison are specific for methods in this category.

2.1 Early methods

The hand is a natural interface to input data to machines, hence the high interest on gesture recognition for Human-Computer Interaction. The first applications on hand gesture recognition were based on sensors and gloves, such as the CyberGlove (Kessler *et al.*, 1995), illustrated in Figure 2.1, and MIT Data Glove (Zimmerman *et al.*, 1987). However, the use of such devices limit the applicability of hand tracking to unnatural interactions.

Another version of the problem is markerless image-based hand pose estimation, that has immediate applications for which many solutions have been proposed with single-viewpoint and multi-viewpoint input devices. Image-based methods for hand pose estimation are strongly based on machine learning, using characteristics such as hand shape, orientation, finger's flexion angles, among others.

Some early works on hand pose estimation were based on multiple RGB views, as a solution to reduce data ambiguity. In this context, the works of Erol *et al.* (2007) and Campos (2006) present an overview of multi-view hand tracking methods. The thesis of Campos (2006) present methods for skin segmentation, articulated object tracking and multiple view hand pose estimation. Figure 2.2 presents the pipeline introduced in Campos and Murray (2006), which uses multiple views, skin colour segmentation, contour

Figure 2.1: *CyberGlove II, reproduced from http://www.cyberglovesystems.com/cyberglove-ii/, accessed in 16/11/2017.*

extraction and computation of global image descriptors. That pipeline follows a framework based on bags of visual words. The global descriptors of all views are concatenated and used as input to RVM, a sparse regressor for pose estimation.

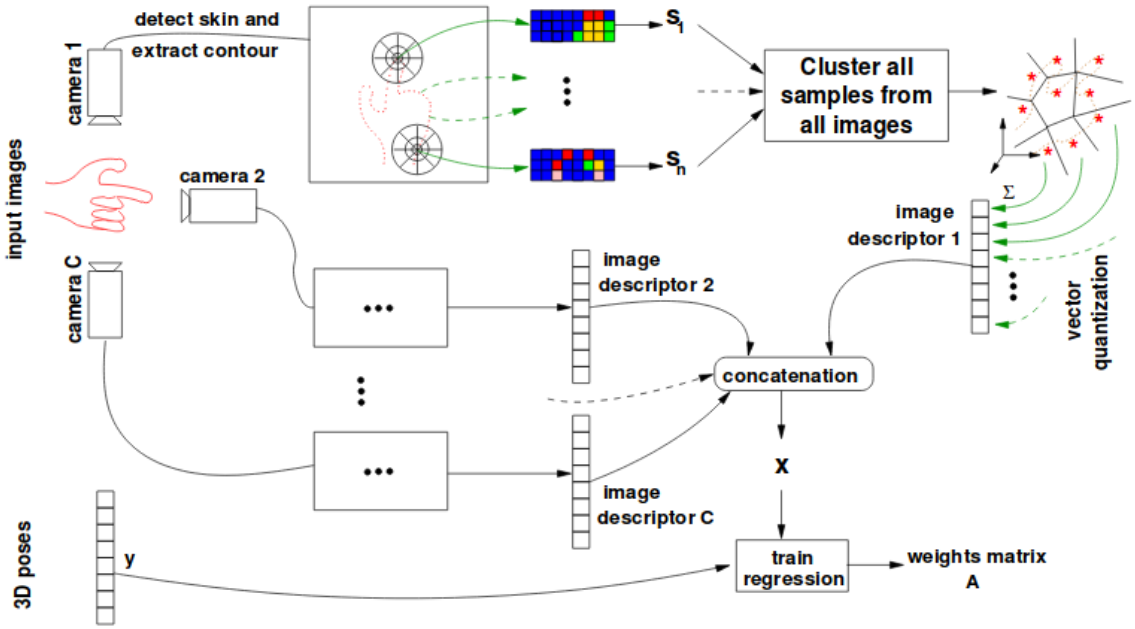


Figure 2.2: *Pipeline proposed by* Campos *(2006) for multiple view hand pose estimation (reproduced with permission from the author).*

Another line of works is based on tracking of single RGB sequences, applying general object tracking models. In this context, we highlight the work of Stenger *et al.* (2006), that combines a hierarchical-based articulated object detection with a probabilistic particle filtering tracking model. Detection is based on a tree-based detection and filtering that uses training data for pose clustering, modeling edge and color likelihoods. Tracking uses a Markov transition matrix, model that estimates the probabilities of state transitions to infer future states in the hierarchical tree search space. The main limitation of this method is that the angular resolution and allowed appearance of the results are dependent of the number of poses on the training set. This work was the basis for many hand gesture

recognition and tracking systems. For more references on the hand gesture recognition problem, refer to the survey of Pisharady and Saerbeck (2015).

2.2 Methods based on depth sensors

With the development of low-cost depth sensors, most methods started to use depth maps as input. The use of depth sensors reduce the data ambiguity without the necessity of configuring and calibrating a multiple view setup, making easier the hand pose estimation task. Oikonomidis *et al.* (2011) pioneered the use of depth sensor data for hand tracking. The basic idea is the minimization of an energy function using a Particle Swarm Optimization (PSO) system. Later, the project focused on tracking the articulated motion of two hands (Oikonomidis *et al.*, 2012).

Approaches based on minimizing an energy term over a model are called generative (or model-driven) methods, being developed in contrast to discriminative (or data-driven) methods, which are based on learning over a dataset. Hybrid approaches combine discriminative and generative subtasks in their pipeline.

Among generative and hybrid methods, a project from Microsoft[1] has presented several new solutions (Sun *et al.*, 2015; Sharp *et al.*, 2015; Tang *et al.*, 2015) for this problem. Sun *et al.* (2015) present a coarse-to-fine hierarchical approach, based on the degrees of freedom of each joint of the model. The pose of different parts of the hand is recovered sequentially, following the order of complexity of each point. Two metrics are used: the per-joint error averaged on all images and the percentage of successful frames (success rate). A new dataset was also made available by the authors.

Sharp *et al.* (2015) propose an *analysis by synthesis* approach to hand tracking, infering the parameters that allow the generation of the input image. The method is based on hand RoI extraction, per-frame reinitialization and model fitting (based on PSO). The function minimized in the model fitting process is named golden energy and represents the difference between joints of a rendered model over the depth image and the position of the points in raw depth maps.

Tang *et al.* (2015) propose a sequence of predictors organized into a kinematic hierarchy, following the basic idea of Sun *et al.* (2015)'s approach. For each step of the predictor, the method produces random samples, minimizes the energy and estimates a partial pose. The generation of random samples is made through a regression forest. The method is evaluated with respect to the variation of the number of particles.

Still among the model-based approaches, Tagliasacchi *et al.* (2015) and Tkach *et al.* (2016) adapt the golden energy term proposed at Microsoft by adding terms based on normal compatibility and closest-point correspondences. Those works were important in showing that classical solutions like the Iterative Closest Point (ICP) algorithm could be adapted in order to help solving harder problems.

Many databases were developed in order to measure and compare model errors. Initially databases like *Dexter1* (Sridhar *et al.*, 2013), *Synthetic* (later called MSHD, designed for

[1] https://www.microsoft.com/en-us/research/project/fully-articulated-hand-tracking/

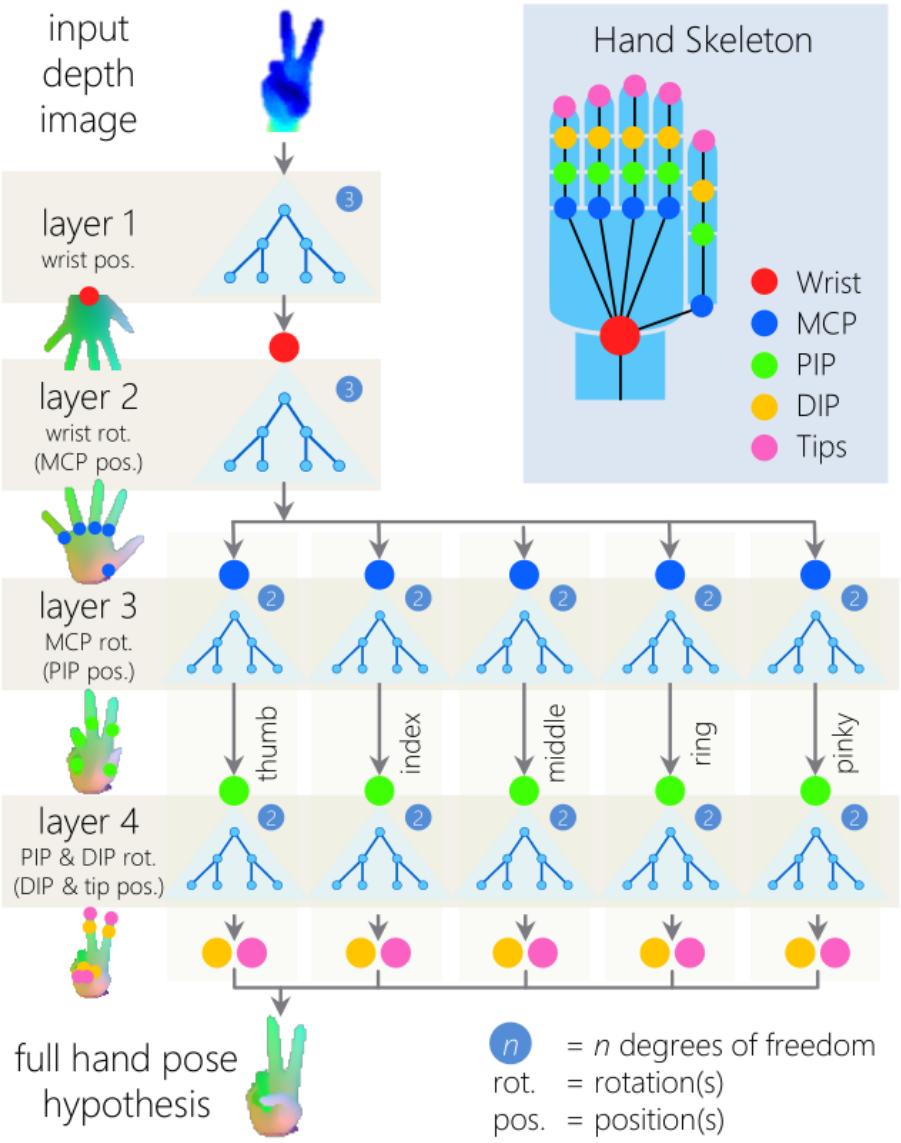


Figure 2.3: *Hierarchical hand pose detection pipeline, extracted from Tang et al. (2015). Copyright ©2015 IEEE.*

stress-test robustness) and *Fingerpaint* (video and depth captured from painted hands, providing a semi-automatic ground-truth). More recently, three main datasets have been used - ICVL (Tang *et al.*, 2015), NYU (Tompson *et al.*, 2014) and MSRA (Sun *et al.*, 2015). Those datasets provide hand data with annotated joint poses, and are described with more details in Section 2.5.

2.3 Methods based on deep learning

In recent years, the development of deep learning algorithms led to significant advances in machine learning and its applications, particularly in Computer Vision (LeCun *et al.*, 2015).

The advent of those new machine learning algorithms combined with the development of accurate solutions for 2D joint detection based on CNNs (S.-E. Wei *et al.*, 2016; Newell *et al.*, 2016), like Convolutional Pose Machines (illustrated in Figure 2.4), led the community of hand pose estimation to design methods based on convolutional neural networks (Oberweger and Lepetit, 2017; Oberweger, Wohlhart, *et al.*, 2015; Zhou *et al.*, 2016; Ge *et al.*, 2017; Guo *et al.*, 2017).

Figure 2.4: *Advances in machine learning allowed significant progress in 2D joint detection, with new methods like Convolutional Pose Machines, reproduced from S.-E. Wei et al. (2016). This method uses a sequential architecture composed of CNNs, producing increasingly accurate estimates for joint locations, illustrated in parts (a) predicting from local evidence, (b) multi-part context and (c) convergence from additional iterations. Those advances also impacted on new solutions for hand pose estimation. Copyright ©2016, IEEE.*

Those methods differ among themselves in the neural network architecture and type, the input image type, the hand representation used and the use of prior constraints. As an example, the *DeepPrior++* (Oberweger and Lepetit, 2017) uses a Residual Neural Network, which is a deep network whose training is based on minimizing residual weights in each layer. This work uses data augmentation in the training, such that realistic samples can be generated from simple geometric transformations over the original training samples. Guo *et al.* (2017) use an ensemble-based neural network, which integrates the results of different regressors in different regions of the image. Chen *et al.* (2019) compute a feature map for each joint and fuse those maps using a structured region ensemble network (named Pose-REN), reaching competitive results. Wan *et al.* (2017) propose the use of Generative Adversarial Networks (GAN) and Variational Autoencoder (VAE), two strong ideas in the recent wave of advances in machine learning. The VAE is used in order to learn and model the distribution of hand poses, and the GAN is used to model the distributions of depth images. In the training process, the method learns a mapping between the latent spaces of both networks in a multitask optimization framework. This method allows training and learning from unlabeled data.

Another line of work includes methods based on volumetric information, which use context features of the 3D point sets in order to more accurately locate the joints. Methods such as *Anchor-to-Joint* (A2J) (Xiong *et al.*, 2019), V2V-PoseNet (Moon, Chang, *et al.*, 2018) and DenseNet (Wan *et al.*, 2018) currently reach very competitive results in all state-of-art datasets for hand pose estimation. DenseNet obtains the hand pose by fusing 2D and 3D heatmaps.

Moon, Chang, *et al.* (2018) uses an encoder-decoder architecture to convert the 2D depth image in a 3D voxel grid, and then estimates the per-voxel likelihood of each keypoint, identifying the positions of highest likelihoods in the V2V-PoseNet method. Those positions are then warped back to real world coordinates. This approach has the drawback of the high computational cost of the voxelization procedure, increasing the difficulty of the training process.

A2J (Xiong *et al.*, 2019) uses anchor points in the depth image which capture the global-local context information. The joint position is regressed by weighting the influence of each anchor point. The neural network used is a 2D-CNN, which lowers the computational cost of training.

Fang *et al.* (2020) recently proposed JGR-P2O, a system for pixel-to-offset prediction based on joint graph reasoning. This system explicitly models the dependencies among joints and the relations between pixels and the joints for better local feature representation learning. This method unifies pixel-wise offset predictions and direct joint regression for end-to-end training, leading to state-of-the-art results with a relatively low computational cost.

A recent approach also worth-citing is that of Poier *et al.* (2018), that uses multiple views to learn implicit pose representations of the hand. An important aspect of this approach is that the training is semi-supervised, using labeled and unlabeled data.

The development of deep learning methods brought the necessity of larger datasets. As a consequence, new million-scale datasets have been made available in 2017: the BigHand2.2M (Yuan, Ye, *et al.*, 2017) and First-Person Action dataset (Garcia-Hernando *et al.*, 2018). With these datasets, deep learning methods can use a much larger training set and reach better results. To consolidate the trend of using CNNs, the International Conference on Computer Vision board organized the HANDS in the million 2017 challenge on 3D pose estimation (Yuan, Garcia-Hernando, *et al.*, 2018a)[2], a competition on a benchmark using the BigHand2.2M dataset.

The results of this challenge were presented by Yuan, Ye, *et al.* (2017) in the form of a survey that discusses design choices as well as the corresponding evaluation results. Aspects evaluated and taken into account were:

- The nature of the input images (2D or 3D): while depth images can be seen as $2D$ points with depth, some methods perform joint detection in a $3D$ voxel grid: result shows that $3D$ volumetric representation presents higher performance;
- If the method uses probability density maps (detection-based) or regresses the parameters directly from the depth image (regression-based): results point that detection-based methods tend to outperform regression-based methods, but regression methods can reach good results using explicit spatial constraints;
- Whether the regression is hierarchical (regression is made by subtasks, usually branches of joints are detected separately and concatenated) or holistic (the whole hand pose is regressed directly in one optimization step), and whether structural constraints and priors are incorporated in the network: it was found that the error

[2] http://icvl.ee.ic.ac.uk/hands17/challenge/

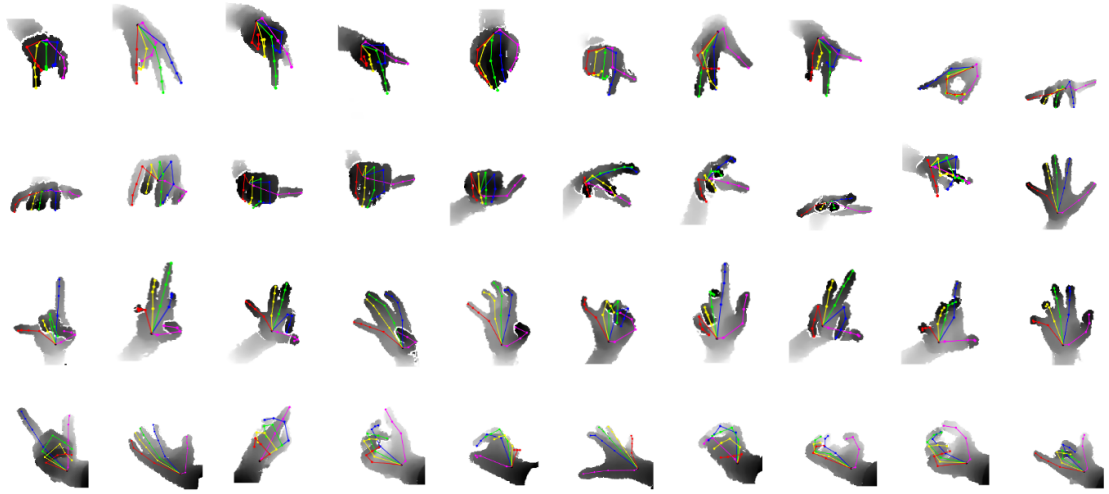


Figure 2.5: *Hand pose samples from the BigHand2.2M dataset, reproduced from Yuan, Ye, et al. (2017). Copyright ©2017, IEEE.*

on occluded joints is narrowed in methods with explicit modeling of structure constraints and hierarchical joints;

- Whether the training is divided in stages and one stage is used to enhance the result of the subsequent stages: cascaded methods performed better in general;

- In general, discriminative methods still generalize poorly to unseen hand shapes, and the use of models with better generative capacity can be a promising choice.

The method with the best result on the challenge is A2J (Xiong *et al.*, 2019), while JGR-P2O (Fang *et al.*, 2020) is the best performing method on NYU and ICVL datasets.

The current panorama of the area indicates the continuous improvement of methods based on deep CNNs for depth images and that there are efforts of many research groups around the world in this direction.

Although a lot of improvement has been observed since the first commodity depth sensors became available, all methods have their limitations and depth maps are still far from perfect. One potential direction for future work is to exploit a pre-processing step to denoise depth maps using a method such as that of Yan et al. (Chenggang Yan, Li, *et al.*, 2020).

Another potential approach for 3D hand pose estimation is the use of a tracking-as-detection method (or indeed, tracking-as-retrieval), similar to what was done by Stenger *et al.* (2006). Despite the relative success back then, such an approach does not seem to have been explored again since the deep learning revolution. The deep multi-view retrieval method (Chenggang Yan, Gong, *et al.*, 2020) has certainly a high potential of success in a tracking-as-retrieval framework.

2.4 Deep Learning image-based methods

With the development of new solutions for 2D joint detection based on CNNs (as mentioned in Section 2.3), new methods that use learning-based 2D joint detection and Inverse Kinematics have been proposed to estimate hand pose based exclusively on RGB image (Zimmermann and Brox, 2017; Panteleris and Argyros, 2017; Mueller *et al.*, 2018; Panteleris, Oikonomidis, *et al.*, 2018). The development of monocular image-based pose estimation methods is important for generalization and ease of use, but the absence of the depth dimension makes the problem much harder. Data-driven methods need much larger datasets to train in order to obtain a good generalization capacity.

As far as we are aware, the first method to perform 3D hand pose estimation from 2D input using deep learning was the approach of Zimmermann and Brox (2017). This method uses three networks in order to compute probability maps. The first network (HandSegNet) is based on the person detector provided by S.-E. Wei *et al.* (2016), casting the hand localization as a segmentation problem. The training process is done with the synthetic dataset presented in the paper (RHD dataset). With the mask provided by this network, the image is cropped and normalized. The pipeline follows with the identification of 2D keypoints on the segmented region, using an architecture similar to the Pose Network (PoseNet) also presented in S.-E. Wei *et al.* (2016). The following step is the application of the PosePrior network, in order to estimate the most likely 3D configuration given the 2D keypoints. This network is trained with respect to a canonical frame, and this makes the training more efficient.

Mueller *et al.* (2018) identified that the approach of Zimmermann and Brox (2017) generalizes poorly to real world images due to the use of synthetic images in the training process. To minimize this problem, Mueller *et al.* (2018) propose the use a Cycle-GAN to enhance synthetic data, such that its statistical distribution resambles real-world hand images. After the training, the method applies a CNN (*RegNet*) to predict 2D heatmaps and 3D joint positions. The final step is a kinematic skeleton model fitting, through energy minimization. Boukhayma *et al.* (2019) incorporate the use of the MANO hand model (Romero *et al.*, 2017). This is a differentiable model that encodes parameters of the view, shape and pose for a hand image, and the method uses a Residual Network (He *et al.*, 2016) to train and estimate parameters of the model.

Recently, Santavas *et al.* (2020) presented a single-stage method, based on a lightweight architecture. This method uses DenseNets as a backbone - each layer propagates its own features to subsequent layers through a channel-wise concatenation. The method relies on a new neural network block (Attention Augmented Inverted Bottleneck Block) and modified pooling (*blur pooling*) and activation (Mish) functions. The method of Santavas *et al.* (2020) is currently the best method in all datasets (Simon *et al.*, 2017; Hampali *et al.*, 2019) used in the literature for 2D hand pose estimation, and it is expected that 2D methods continue to evolve and reach better results.

Moon, Yu, *et al.* (2020) presented a dataset with 2.6 million images of 2D annotations on hand interactions. Since the publication is very recent, the only method tested was the baseline Inter-Net proposed in the publication, but it is expected that this dataset impacts the quality and development of new solutions, with a much larger set of samples.

There are other problems that relate to inferring 3D from 2D RGB images, such as room layout estimation (C. Yan *et al.*, 2020) can certainly inspire methods for hand pose estimation without depth information. More closely related are the methods of 2D joint detection based on CNNs and their successful application to problems such as human pose estimation (Cao *et al.*, 2021).

2.5 Summary of datasets

Table 2.1 presents a summary with the datasets cited through this chapter.

More information about the datasets proposed for hand pose estimation, as well as the main papers published on conferences, theses, workshops and challenges are available in the repository Awesome Hand Pose Estimation[3], which is frequently updated with new data, composing a snapshot of the state-of-the-art on hand pose estimation.

Dataset	Input	Year	Synthetic / Real	#frames (train/test)	#subjects	#joints	Reference
Dexter1	Depth + RGB	2013	Real	2137	1	6	Sridhar *et al.* (2013)
NYU	Depth	2014	Real	72k/8k	2	36	Tompson *et al.* (2014)
ICVL	Depth	2014	Real	331k/1.5k	10	16	Tang *et al.* (2015)
Fingerpaint	Depth	2015	Synthetic	100k	1	21	Sharp *et al.* (2015)
MSRA	Depth	2015	Real	76375	9	21	Sun *et al.* (2015)
HANDS17	Depth	2017	Real	2.2M	10	21	Yuan, Ye, *et al.* (2017)
RHD	RGB	2017	Synthetic	41258 / 2728	20	21	Zimmermann and Brox (2017)
FreiHAND	RGB	2019	Real	130K/3960	-	21	Mueller *et al.* (2018)
InterHand2.6M	RGB	2020	Real	2.6M	27	21	Moon, Yu, *et al.* (2020)

Table 2.1: *Summary of the main datasets used in the literature in hand pose estimation.*

2.6 Discussion

This chapter presented the main methods in the literature for hand pose estimation. The popularization of depth sensors and the development of data-driven deep learning methods allowed new solutions to arise, with deep learning approaches reaching the best results to date in the standard datasets (ICVL, MSRA, NYU and HANDS17).

The RGB variant of hand pose estimation is a much more difficult problem, in an earlier state of development with deep learning solutions. The MANO hand model is an important element in this pipeline, making easier the encoding of hand shape parameters in neural network architectures.

In the context of our work, the goal was to find a method suitable for hands with rheumatoid arthritis, as well as healthy hands. Ideally, the application would benefit if the method could handle 2D inputs, with the therapists being able to evaluate video RGB inputs recorded from the patients remotely. Therefore, we made preliminary experiments with the image-based method of Zimmermann and Brox (2017), whose source code was made available by the authors[4]. The implementation of the method was publicly available

[3] https://github.com/xinghaochen/awesome-hand-pose-estimation#datasets, committed on 28/03/2021
[4] https://github.com/lmb-freiburg/hand3d

and we were able to perform qualitative evaluations. We evaluated qualitatively that the results were inconsistent and very sensitive to skin tones and the presence of the orthosis. However, since this method is a baseline for image-based deep learning methods, the use of more refined approaches can deal with some of the drawbacks. Santavas *et al.* (2020) show robust qualitative results for hand-object interaction, but skin tone diversity is an issue that is still not addressed by any recent method or dataset.

In the project we opted to use a method based on depth images, since all the drawbacks noticed were inherent to any method that uses RGB images as input. Depth image processing considers only the geometry of the hand, and therefore such methods are not affected by skin colour variations. The presence of the orthosis is also shown to be a minor drawback in our evaluation. Therefore, depth-based methods have been considered a better choice for data acquisition from patients with rheumatoid arthritis, considering the current state-of-the-art the literature. We evaluated some methods in our pipeline, and chose the Pose-REN method (Chen *et al.*, 2019). Although more recent methods present better results, this method is competitive in all datasets and can be executed in real-time. Furthermore, the authors provide a demo code to run the method for any depth image input using pre-trained models[5], compatible with the Realsense SR300 sensor used in our acquisitions. Our experiments showed that this method is robust in most situations, even with the orthosis. More details on the dataset acquisition process and on the Pose-REN method are described in Chapter 3 and Section 4.1, respectively.

As described in Section 1.1, the use of computer vision for joint estimation in a "real-world" clinical setting is still under development and relies on the development of precise methods to deal with the challenges inherent to the pose estimation problem. In particular, for the hand/finger problem, we were not able to identify computer vision approaches in the literature that are trained specifically for precise joint identification in health applications.

However, we expect that the significant advances in computer vision and hand pose estimation can lead to a series of advances in the inherent applications. For hand occupational therapy, the possibility of acquiring 3D frames and skeletons reduces most of ambiguities found in 2D visual estimation, and its use in the treatment of patients can be far less intrusive than the goniometers. It is worth mentioning that according to Meals *et al.* (2018), one of the main weaknesses of digital photogrammetry is that joints must be photographed and then measured. Most hand pose estimation methods, including Pose-REN (Chen *et al.*, 2019), present the possibility of acquiring results in real-time, simplifying the method and favouring the automation of the process.

[5] https://github.com/xinghaochen/Pose-REN/tree/master/src/demo

Chapter 3

Data acquisition protocol and dataset formation

In this thesis, we propose a complete framework for hand range of motion estimation. This chapter details the dataset formation process and the framework proposed for hand range of motion estimation, presented in Figure 3.1. This Chapter focuses on data acquisition, detailing and illustrating project decisions about sensors, methods, patient positioning and other issues found, which was the first step of the project implementation. The proposed pipeline, illustrated in Figure 3.1 is able to extract and analyze hand joint angle measurements from RGBD images acquired in real-time.

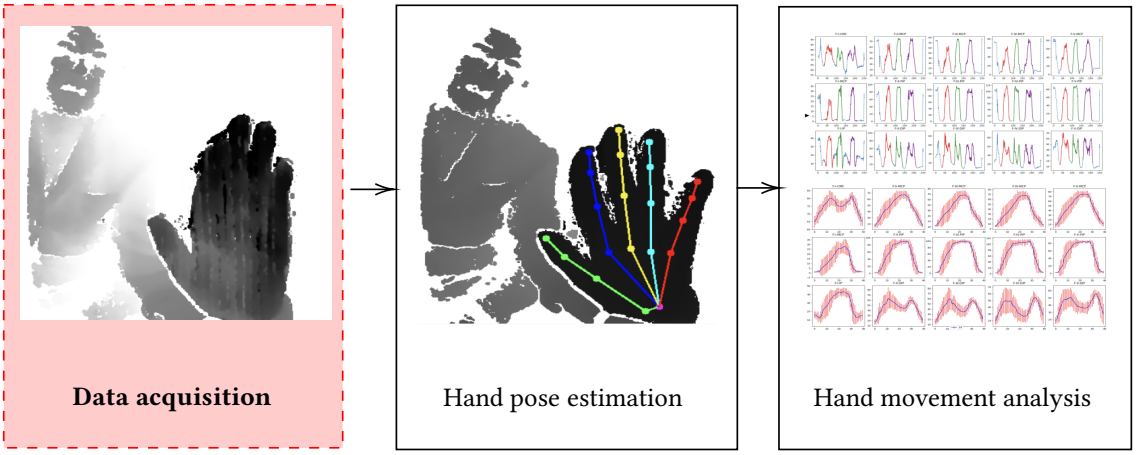


Figure 3.1: *Proposed pipeline, highlighting the developments of current Chapter. The next steps are discussed in Chapters 4 and 5.*

This Chapter presents a narrative of the process of iteratively designing acquisition setups and evaluating recent methods, with data from real patients.

The development of the data acquisition module was made in collaboration with Professors Valeria Elui and Daniela Goia at the *Hospital das Clínicas at Faculdade de Medicina de Ribeirão Preto (FMRP-USP)*, located at the University of São Paulo's Ribeirão Preto campus. We studied different sensors and designed an acquisition protocol. As first

step of the project, our goal was to design a baseline setup for data acquisition, study depth sensors and acquire data from patients in recovery of flexor tendon surgery.

3.1 Data acquisition hardware and protocols evaluation: first experiments

Initially the plan was to use three different sensors in the acquisition, so that the hand can be captured from multiple views. The evaluated sensors were: the *Intel RealSense® R200*, suitable for acquisitions in medium range; the *Intel RealSense® SR300*, that can capture points at a closer range, and the *Leap Motion®*, which generates a black-box coarse hand tracking result and is designed as an interface device for gesture recognition. The libraries used in the project were *librealsense* and the *Leap Motion Orion SDK*.

(a) *R200* **(b)** *SR300* **(c)** *Leap Motion*

Figure 3.2: *Sensors used on the initial setup.*

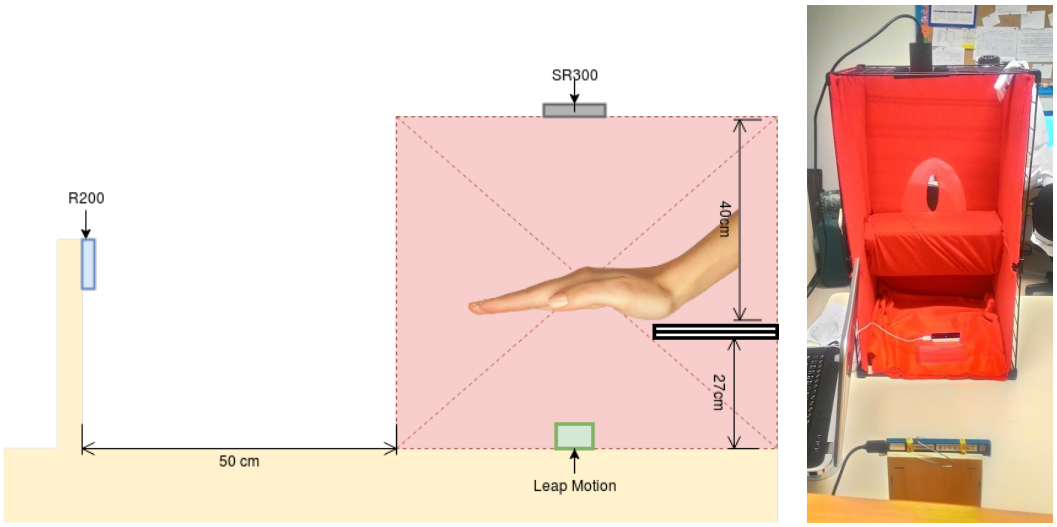
The setup was built in a way to maximize the amount of relevant information extracted from the three sensors, which are positioned at their minimal depth range that produces stable results. This was a concern especially in the R200, since it is a medium range sensor. It was positioned to capture the hand from a frontal view with a larger distance. The SR300 captures the hand from top viewpoint, and the Leap Motion in an even shorter distance, from a bottom view. In some of the captured sequences the patient used an orthosis, a mechanical device used on the treatment process in order to enhance the movement capability.

Sensor	Range (m)	Depth resolution	Num Cameras
Realsense R200	0.5m - 3.5m	480p	2 IR, 1 RGB
Realsense SR300	0.2m - 1.5m	480p	1 IR, 1 RGB
Leap Motion	up to 0.8m	640x240	2 IR

Table 3.1: *Attributes of the sensors used in the initial setup*

The setup was mounted in an uniform background environment due to the expectations of using RGB methods. The patient also wore a blue wristband in the first experiments in order to facilitate finding the wrist and identifying the hand. The sensors were disposed

in such a way that the hand is positioned near to the minimal range of each sensor, maximizing its capture resolution and details. Figure 3.3a shows a representation of the setup from a side, with the measurements of the distances used in the sensor positioning. Figure 3.3b shows the back view of the setup, after mounted.



(a) *Diagram of the designed setup. The sensors were positioned such that the distance to the hand is closer to the minimum range.*

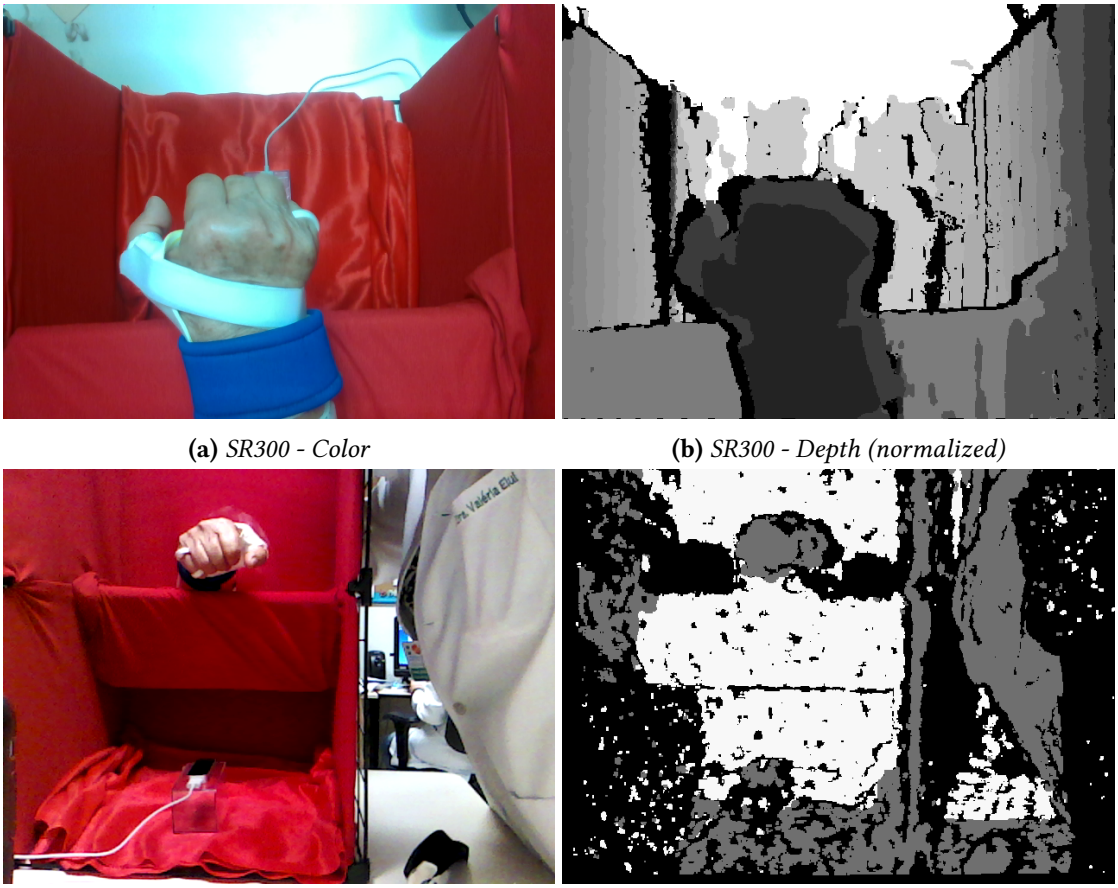
(b) *Setup after mounted.*

Figure 3.3: *Setup used on the acquisition process. All sensors were positioned to maximize the capture resolution - the hand is positioned near to the minimal range of each sensor (40cm for the SR300 and 50cm for the R200).*

Initially, we acquired data from two different patients in different combinations of the sensors. With the setup mounted, we were able to simultaneously obtain data from all sensors. For one of the patients, three sequences were gathered with different sensor combinations. For the other patient, the dataset contains the same sequences with and without an orthosis. In all sequences, the patient wore a blue wristband, to facilitate the localization of the wrist. We observed an interference between the infrared lights emitted by the SR300 and the Leap Motion, and took the decision of doing the acquisition separatedly for the sensors. Figure 3.4 show samples of this acquisition.

We have then decided to acquire data using only the sensor SR300, since our objective was to form a dataset from hands with unusual movement patterns and with orthosis. We took the decision of withdrawing the R200 since the medium-range depth images captured by this sensor do not capture enough detail of the hands (see Figure 3.4d). The Leap Motion also had issues, and could rarely locate and track the hand. The hand tracking algorithm of this sensor is a black-box algorithm, so it was impossible to assess depth maps and evaluate results or perform fine-tuning operations. Our assumption in this issue is that the joints of the patients could not be recovered by the Leap Motion since they are different from the usual hand joint pattern. Figure 3.5 presents an example of data acquired in this session.

We gathered sequences from five patients using the SR300, with and without the orthosis. At this point, a new version of *librealsense* was released, with a record/playback tool which was used for the acquisition. Besides, we evaluated an implementation of Guo



(a) *SR300 - Color* **(b)** *SR300 - Depth (normalized)*

(c) *R200 - Color* **(d)** *R200 - Depth (normalized)*

Figure 3.4: *Example of an acquisition from a patient with orthosis, from sensors SR300 and R200.*

et al. (2017) and used it to qualitatively evaluate some of the results. Some sequences from different camera positions were obtained as well. Figure 3.6 presents a frame acquired in this session.

Acquisition 1 - July 12th 2017	
Patients	3
Sequences	12
Frames	2889
Size (MB)	2827.8
Sequences with orthosis	3
Experiments	Simultaneous captures with Leap Motion, SR300 and R200.
Conclusions	Leap Motion uses the same frequencies as R200 and SR300, and an interference pattern can be seen in the depth images. Since it is a medium/long range sensor, R200 does not generate reliable data for hand pose estimation.
Decisions	Acquire data from Leap Motion and SR300 separately. Stop using the R200 sensor.

Table 3.2: *Summary of the data acquisition experiment - July 12th. 2017*

Acquisition 2 - October 11th 2017	
Patients	1
Sequences	4
Frames	1353
Size (MB)	501.9
Sequences with orthosis	2
Experiments	Only separated captures obtained from SR300 and Leap Motion. Only one of the scheduled patients attended. Some control sequences were recorded.
Conclusions	Leap Motion does not recognize hands with orthoses and unusual shapes.
Decision	Use only the SR300 sensor.

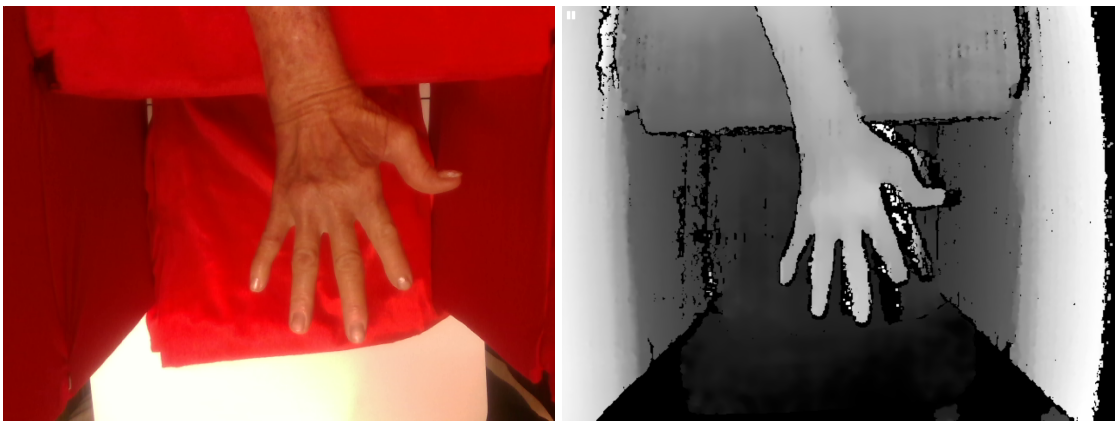
Table 3.3: *Summary of the data acquisition experiment - October 11th. 2017*

Acquisition 3 - November 23rd 2017	
Patients	5
Sequences	10
Frames	6076
Size (MB)	1976.7
Sequences with orthosis	4
Experiments	All sequences were recorded with the SR300 sensor. As *librealsense2* was released, we used the new record/playback tool to capture data. Some experiments were made with a hand pose estimation algorithm.
Conclusions	Hand tracking method had some difficulties to segment the hand from the rest of the image.

Table 3.4: *Summary of the data acquisition experiment - November 23rd. 2017*

(a) *Color* **(b)** *Depth (normalized)*

Figure 3.5: *Example of an acquisition from the sensor SR300, made in October 11th.*



(a) *Color* **(b)** *Depth (normalized)*

Figure 3.6: *Example of an acquisition from the sensor SR300, made in November 23rd.*

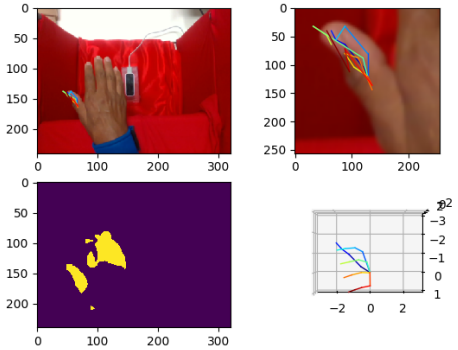
3.2 Hand pose estimation and acquisition protocol improvements

The acquired data was used to test hand pose estimation methods, and choose an adequate method for our framework. At a first moment, it was hard to find a method suitable for use "in the wild", with most methods being reproductible only in the evaluation datasets, but we managed to test the methods of Guo *et al.* (2017) and Zimmermann and Brox (2017) with the patient data. The implementation of both methods was available[1] [2], as well as an 'in-the-wild' hand pose estimation applications that generated real-time results for each frame. Those results, however, were shown to be inadequate, evidencing issues in the setup.

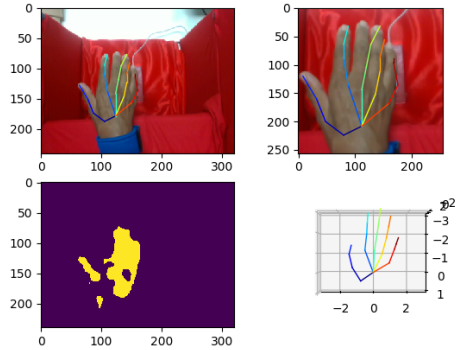
Figure 3.7 shows the results of Zimmermann and Brox (2017) method. Since this two-step method that uses segmentation is image-based, the presence of the orthosis and

[1] https://github.com/lmb-freiburg/hand3d
[2] https://github.com/guohengkai/region-ensemble-network

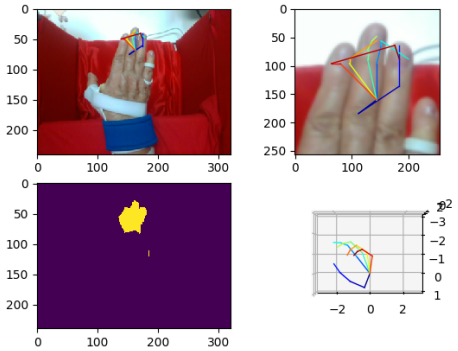
the poor skin segmentation obtained were issues that affected the quality of the results. Figure 3.8 shows result samples of Guo *et al.* (2017) method. Since it is depth-based, the demo works well when the hand is the closest element to the camera, but a background clutter removal method is necessary for result enhancement. However, the level of details in the results obtained were far from the desired level, even when the segmentation was correctly made.



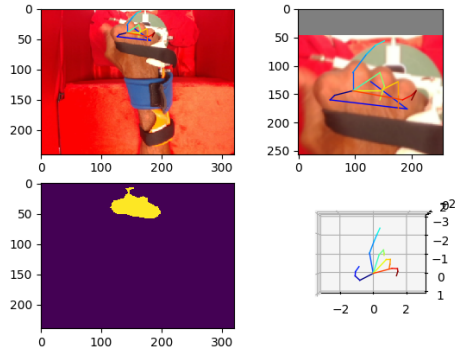
(a) *In this case the failure on the hand segmentation (bottom left) yields a bad pose recovery.*



(b) *The HandSegNet worked better in this image, leading to a better result if compared to (a), that shows a similar hand pose.*



(c) *This is a case with the orthosis, and it affects strongly the color-based hand segmentation and consequently the hand pose estimation.*



(d) *This case is affected by the occlusion caused by the orthosis and by the poor result of the segmentation.*

Figure 3.7: *Sample results from Zimmermann and Brox (2017) method in one frame of our dataset.*

Following these experiments, it was decided that the hand pose estimation should be executed in real-time during the capture, allowing the repositioning of the hand by the therapist. The real-time execution of the code demanded an equipment with GPU processing and CUDA library.

In this meantime, the Pose-REN method (Chen *et al.*, 2019) was published as a preprint and its implementation was available in the repository of the authors. With the availability of a pre-trained model on HANDS17 dataset, we were able to get much more accurate results. The main constraint of the method was the need of a background clutter removal algorithm, such that the hand needs to be the closest object to the camera. This made

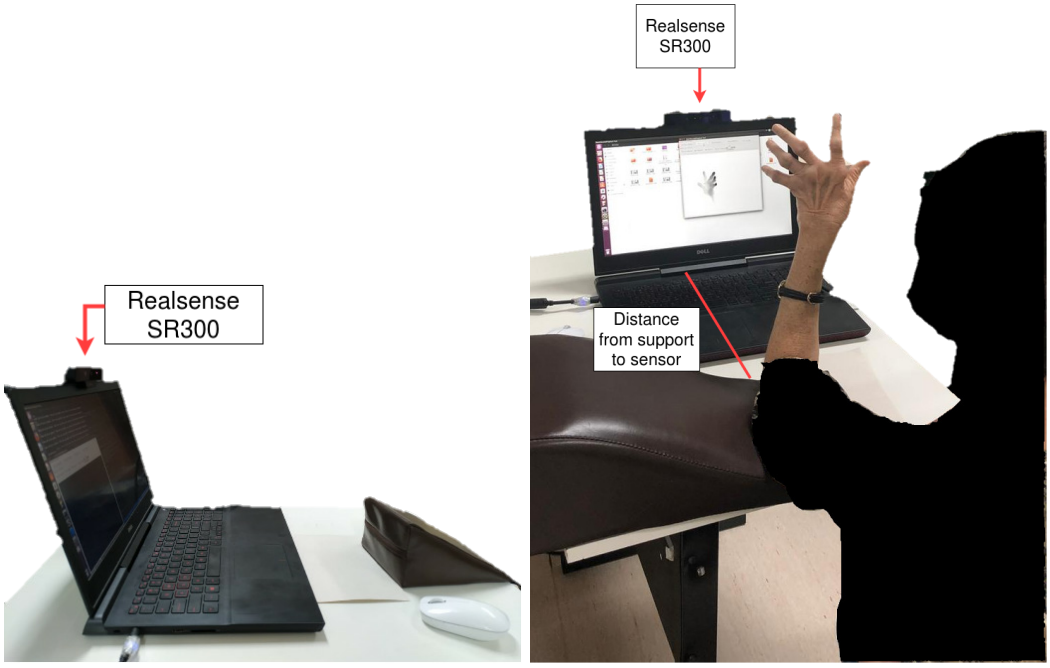
(a) *In this example, the hand is the closer object to the camera, so the method returns a coarse pose.*

(b) *In this case, the wrist was in the same depth as the hand, so the result was inaccurate.*

Figure 3.8: *Sample results from Guo et al. (2017) method in one frame of our dataset.*

the results recorded from previous acquisition setups unfeasible, because patients were resting their wrist on a support which perturbs the depth map. It was not possible to segment the hand with a simple depth threshold (*e.g.* Figures 3.4 and 3.6). Therefore, with the resources available and the expertise gained in previous acquisitions, we decided to make new data acquisition sessions. The new acquisition protocol took this into account, simplifying the setup and positioning the sensor in front of the patient, and with an arm rest for the patient be more comfortable in the process. During the acquisition, the patient was oriented to flexion the arm such that only the hand is visible. With the algorithm executing in real-time, the position could also be adjusted during the acquisition process,

because the patient and therapist have online feedback. This new setup positioning is shown in Figure 3.9.



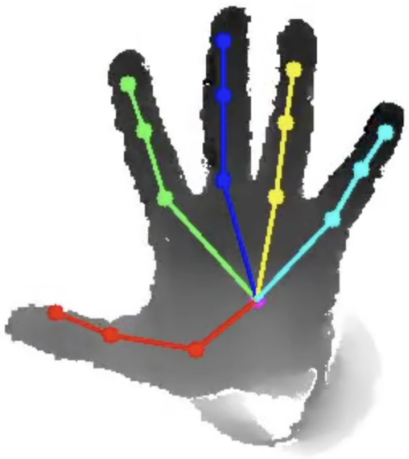
(a) *a* **(b)** *b*

Figure 3.9: *Setup used for data acquisition, with the* Intel RealSense® SR300

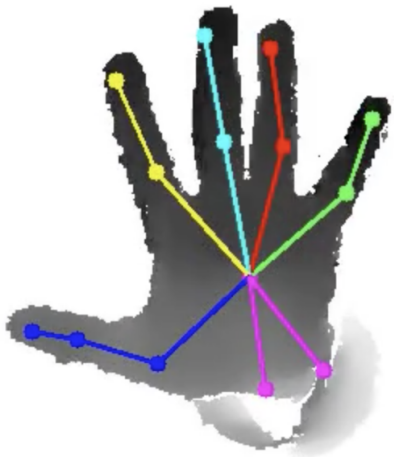
Control data was recorded during September 2018, in São Paulo. Eight people were recorded performing movements of flexion/extension and abduction/adution with both hands, totalling 100 sequences. None of those people had rheumatoid arthritis or ulnar deviation on fingers.

We tested the four pre-trained models made available by the authors of Pose-REN, as exemplified in Figure 3.10, and after a qualitative evaluation on a number of video sequences, and considering previous works on the literature, we chose to use the HANDS17 model. The availability of a pre-trained model in the HANDS17 dataset enhanced greatly the precision and robustness of the results, since this dataset presents the possibility of training with a much larger training set (see Table 2.1 for comparison).

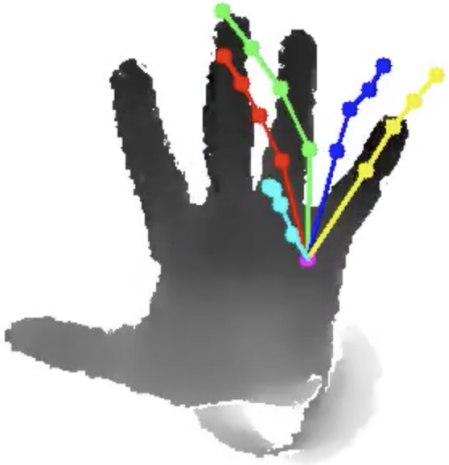
A new acquisition with rheumatoid arthritis patients was set up in October 5th 2018 in Ribeirão Preto. In this acquisition session, we obtained data from three patients with rheumatoid arthritis. For each patient, hand and movement type we acquired data with and without the orthosis. Each movement sequence is saved in a file and can be reproduced as input for a virtual RealSense sensor. Table 3.5 presents a summary of the current state of our dataset after the new acquisition sessions. The quality of the results was improved with the use of Pose-REN, and even the orthosis and poses with self-occlusion were handled with relative success by the method. The main drawback observed was in the abduction acquisitions, in hand poses where the fingers are closed. Figure 3.11 exemplifies some of the sequences recorded with this setup.



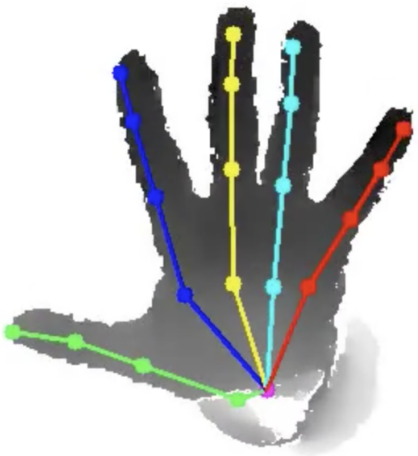
(a) *Result with the ICVL pre-trained model.*



(b) *Result with the NYU pre-trained model.*



(c) *Result with the MSRA pre-trained model.*



(d) *Result with the HANDS17 pre-trained model.*

Figure 3.10: *Sample results obtained by applying Pose-REN (Chen et al., 2019) on control data for all pre-trained models.*

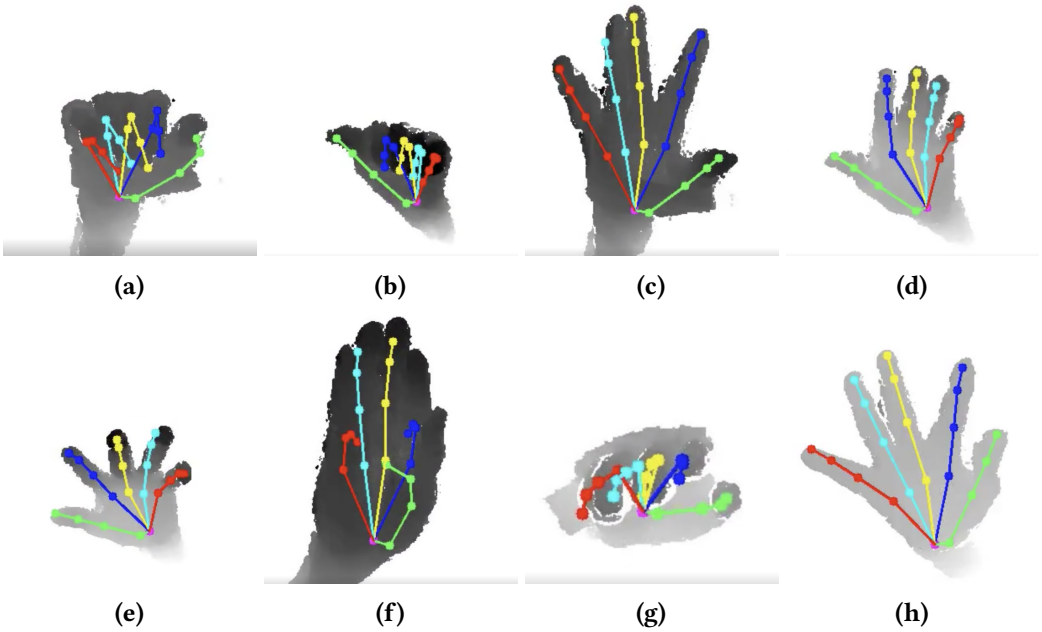


Figure 3.11: *Sample results obtained by applying Pose-REN model (Chen et al., 2019) trained on HANDS17 model with patients data, obtained in October 2018.*

Acquisition 4 - October 5th. 2018	
Patients with rheumatoid arthritis	3
Sequences	23
Sequences with orthosis	6
Control sequences with healthy hands	100
Size (GB)	147.5
Observations	The possibility of seeing the result of the hand pose estimation in real-time enhanced the acquisition process. The model used was robust to many situations tested. Sometimes the patients had difficulties to follow the protocol and the hand was at the same depth as the arm at start. The method had some difficulty with fingers closed in abduction movements. Patients were in different stages of the disease.
Experiments	All sequences were recorded with the SR300 sensor. As *librealsense2* was released, we used the new record/playback tool to capture data. Some experiments were made with a hand pose estimation algorithm.
Conclusions	Hand tracking method worked well but in some cases had difficulties to segment the hand from the rest of the image.

Table 3.5: *Summary of the data acquisition experiment - October 5th. 2018*

3.3 Final protocol and GUI interaction

With the amount of data acquired, the focus of the work in the subsequent stages was on the data analysis. With the angle computation, we developed a GUI tool to capture data and estimate the hand angles in real-time. For range of motion measurements, we studied reference guides in occupational therapy and estimated the average maximum and minimum values from the movements of a patient. This estimation tool, however, still lacked ground-truth comparison. Therefore, we felt the necessity of acquiring more data to complement the experiments.

Two new acquisition sessions were performed with patients in the Hospital das Clínicas from FMRP, in the Ribeirão Preto campus of the University of São Paulo. The sessions took place in September 6th. and 13rd. 2019. In those acquisition sessions, we obtained additional data from five patients with rheumatoid arthritis using a sensor *Intel RealSense SR300*, using the same setup of the acquisitions described in Table 3.5 and Figure 3.9. The patients evaluated were in different stages of the disease, but some movement sequences were challenging for the algorithm. For each patient, hand and movement type (flexion/extension and abduction/adduction) we acquired data from patients with and without the orthosis, as well as their manual range of motion measurements from a goniometer, in order to compare results. Each movement sequence recorded from the SR300 sensor is saved in a bag file and can be reproduced as input for a virtual RealSense sensor. For those sessions, we formalized an acquisition protocol, described in Appendix A.4. Table 3.6 summarizes the acquisitions. Figure 3.12 shows sample results of this acquisition. The main issues found in this acquisition were in cases where the arm is visible in the camera.

Acquisitions 5 and 6 - September 6th. and 13rd. 2019	
Patients with rheumatoid arthritis	5
Sequences	56
Experiments	All sequences were recorded following the protocol and using the GUI tool, and all patients had their range of motion measurements computed by a manual goniometer for comparison.
Observations	The depth cluttering was manually made in the GUI, facilitating the protocol for the patient. Some patients were in a much more advanced state of AR, and the method had some difficulty to deal with the movement patterns. For the first time, an orthosis affected the result of pose estimation.
Conclusions	The protocol was successful and manual goniometer data was provided for comparison.

Table 3.6: *Summary of the data acquisition experiments - September 6th and 13rd. 2019*

The data obtained in the sessions was complemented with the sequences obtained in 2018, and we also obtained data for flexion movements for 5 new subjects that compose the control group. Table 3.7 presents a summary of our dataset. As a conclusion, we managed

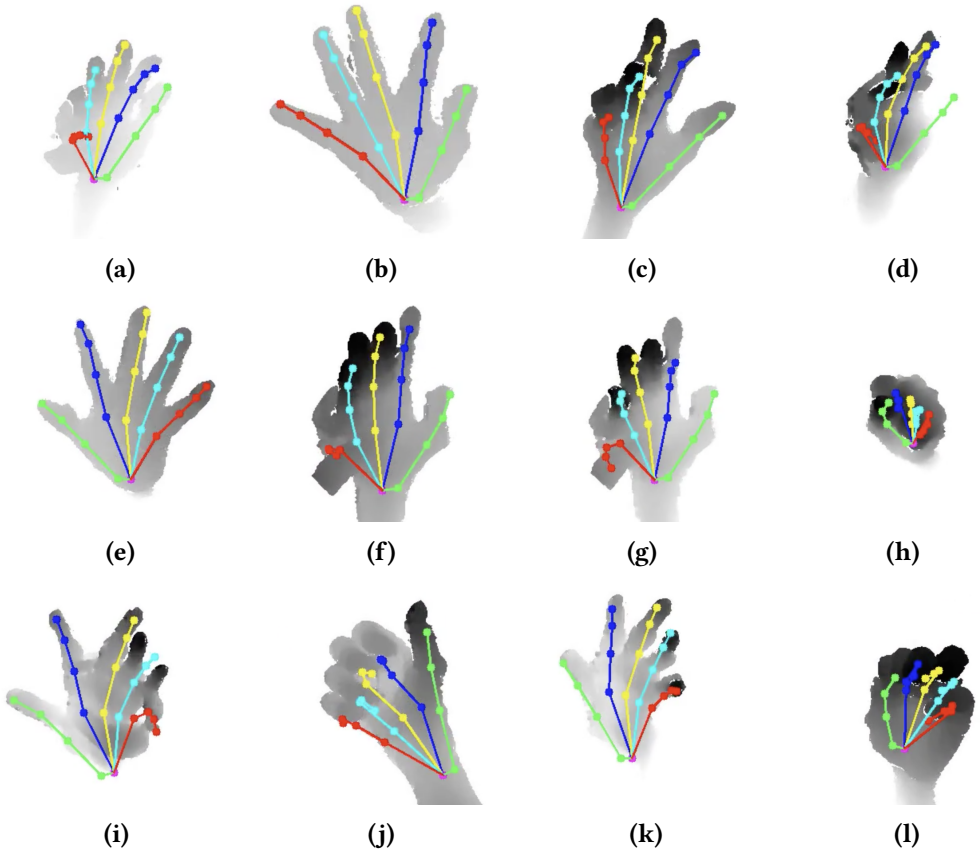


Figure 3.12: *Sample results obtained by applying Pose-REN model (Chen et al., 2019) trained on HANDS17 model with patients data, obtained in September 2019.*

to build a dataset with 891 movement sequences of flexion and abduction from patients and control subsets. Our dataset is the first of its kind and we made it publicly available[3]. The dataset contains samples captured from 12 healthy subjects and 8 RA patients, performing flexion and abduction movements with each of their hands. For each RA patient and each hand with ulnar deviation, we obtained two flexion and two abduction sequences. The data acquired from patients is limited due to the availability of patients and occupational therapists. The raw data captured from the sensor included RGB channels, which could potentially expose and enable the identification of patients. We reprocessed all the data to remove the RGB channels and to store the depth maps in an accessible way that is compatible with Pose-REN and does not affect the results reported.

[3] http://vision.ime.usp.br/~cejnog/handanalysis

Summary	
Patients with rheumatoid arthritis	**8**
Number of people in the control set	**12**
Patient Sequences	**79**
Control Sequences	**108**
Patient clips	**310**
Control clips	**581**
Total clips	**891**
Total number of frames	**85755**
Frames used on clips	**60192**
Percentage of frames used	**70.2%**
Size (GB)	**482**

Table 3.7: *Summary of our final dataset.*

3.4 Discussion

In this chapter, we presented the formation of the dataset used in this work, detailing each of its steps: the selection of the sensors, problems found in acquisition, project decisions taken, the selection of the hand pose estimation method adequate for the setup among many new methods that arose from the deep learning revolution, data anonymization and publication of the dataset. All steps and decisions taken during the process are important, in a way that mistakes could leverage or stall the success of the project or even make its execution unfeasible. We opted to report design decisions that turned out to be mistakes, such as the first setup and the use of sensors in parallel.

The amount of data obtained covered a small number of patients in the context of machine learning (8 patients, 12 control subjects) due to many factors: the limited availability of the appointments with patients, the continuous development of the setup during the doctorate, the 314km distance between São Paulo and Ribeirão Preto. Considering these variables, the amount of data obtained was solid and we were able to obtain multiple data for each movement and patient. FMRP-USP provided full support in the data acquisition process.

The dataset was made available online and any hand pose estimation method based on RGBD input can be used in further steps of the pipeline. To our knowledge, this is the first dataset for hand pose estimation to contain data from Rheumatoid Arthritis patients, which makes it challenging for current state-of-art pose estimation methods and we hope that it can contribute to the use of computer vision techniques in hand occupational therapy.

The main limitation of the proposed dataset is that we did not acquire ground-truth hand joint annotations per frame, due to the simplified setup and inherent limitations on the sensors that were used. This impacts the subsequent pipeline steps and validation in many ways: in the hand pose estimation task, we decided to use a model trained in a different standard dataset (HANDS17), and since both sets are different we cannot be certain whether such model can generalize to unseen shapes; for clip extraction, the

annotation was based on peaks and valleys and is validated manually. In summary, the absence of hand joint annotations per frame potentially amplifies uncertainties from different sources in the pipeline. These issues are further explored in Chapter 5.

Chapter 4

Proposed method: hand tracking, analysis and classification

This chapter details the framework proposed for hand range of motion estimation. Following the pipeline illustrated in Figure 4.1, the proposed methods for 3D hand pose estimation (Section 4.1) and hand movement analysis (Section 4.2) are presented. The proposed pipeline is able to extract and analyze hand joint angle measurements from RGBD images acquired in real-time.

Figure 4.1: *Proposed pipeline, highlighting the steps discussed in this chapter. The data acquisition step was discussed in Chapter 3*

4.1 Hand Pose Estimation

Given the unusual features of our dataset, the hand trackers that generate the best results on standard benchmarks do not necessarily perform well on rheumatoid arthritis patients. After a range of preliminary experiments, we concluded that, at the time of the design of our experiments, out of the most recent real-time 3D hand pose estimation methods, the one that gives the best results in our data is Chen et al.'s Pose-REN method (Chen *et al.*, 2019), trained with the HANDS17 dataset (Yuan, Garcia-Hernando, *et al.*,

2018b)[1]. Pose-REN method is based on the estimation of feature maps using Convolutional Neural Networks (CNNs). These feature maps are combined using an ensemble network, in order to generate a consistent hand pose. The method is illustrated in Figure 4.2.

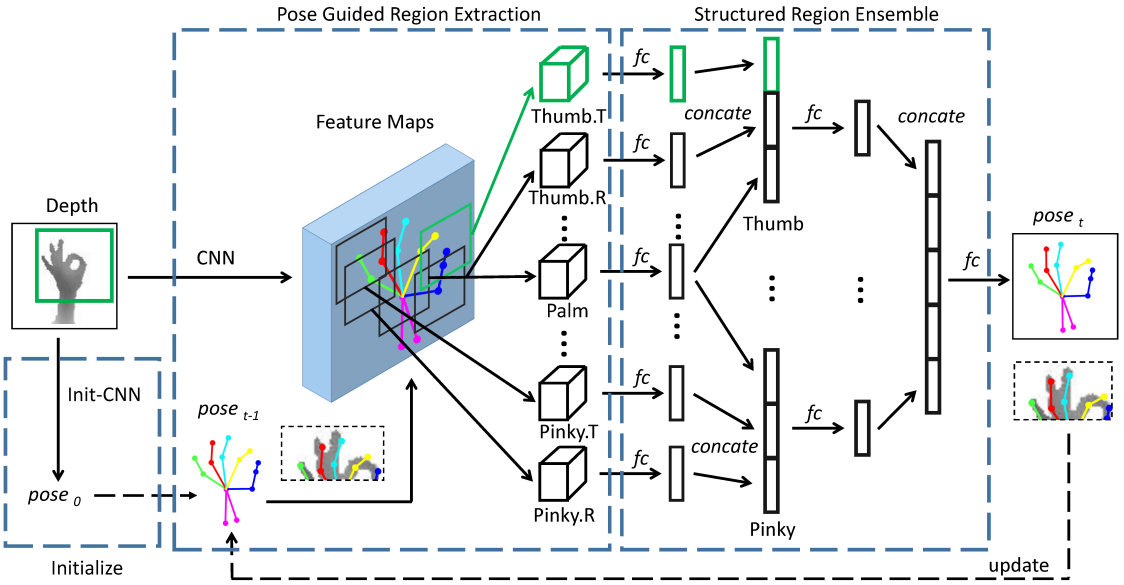


Figure 4.2: *Pipeline used on Pose-REN hand pose estimation method. Reproduced from Chen et al. (2019) (Copyright license nr. 4918240801176)*

The method takes as input a depth image D and returns as output the 3D locations $\mathcal{P} = (p_{xi}, p_{yi}, p_{zi}), i \in \{0, \cdots, N_j\}$ of the hand joints, where N_j is the number of joints of the hand model. The architecture is recurrent: the current estimate of the hand pose $\mathcal{S}_t$ is used as input to help refining it at $\mathcal{S}_{t+1}$. In the first iteration, a coarse hand pose $\mathcal{S}_0$ is estimated using a simple CNN. The network then enhances this pose in two steps: pose-guided region extraction and structured region ensemble. For a pose $\mathcal{S}_t$, in the region extraction, each point of the skeleton is projected from world to pixel coordinates, and a bounding box around each joint is cropped, generating the feature regions $\mathcal{F}_i^t$. In the ensemble network, those feature regions are processed by fully-connected (fc) layers, generating for each joint j feature vectors $h_j^{l_1}$, where l_1 indicates the first fc layer. These feature vectors are integrated hierarchically according to the topology of the hand, i.e. joints that belong to the same finger are concatenated in the same vector. This vector is then fed into another fc layer, whose output is a feature vector $h_i^{l_2}$ for each finger i. The vectors of all fingers are concatenated and again fed to a fc layer, whose output is $\mathcal{S}_{t+1}$, a $3 \times N_j$ matrix of point positions in 3D. This pose is then fed into the initial layer, starting a new iteration of the network, and gradually features around the location of the fingers contribute more to the feature vectors than distant features, optimizing the output pose. This method is currently among the best performing methods in all state-of-art datasets. Its implementation is available online [2].

[1] After the development of our experiments and the writing of this paper, the JGR-P2O (Fang *et al.*, 2020) was published along with its source code. An evaluation of that method for AR patient image sequences is suggested as future work.

[2] https://github.com/xinghaochen/Pose-REN

With respect to the taxonomy proposed by Yuan, Garcia-Hernando, *et al.* (2018b) (see discussion in page 16), the Pose-REN method is a 3D method that uses probability density maps, with hierarchical regression and the training is performed in a single step.

The skeleton used by HANDS17 dataset has 21 points of reference: the center of the wrist (W) and for each finger x the proximal interphalangeal (PIP_x), the distal interphalangeal joints (DIP_x) and the tip ($fingertip_x$). The exception is the thumb, which is represented by the carpometacarpal joint (CMC) and a single interphalangeal joint (IP). Fingers are represented by the respective roman number (I-V: I for the thumb, V for the little finger). In our pipeline, we will refer to a depth image as $D(x, y, t)$ and to the skeletons obtained by the hand pose estimation algorithm as $\vec{S}(t)$. This skeleton is illustrated in Figure 1.5.

As a preprocessing, we perform depth filtering. As mentioned in Chapter 3, in most cases this is sufficient to segment the hand in the scene.

4.2 Hand movement analysis

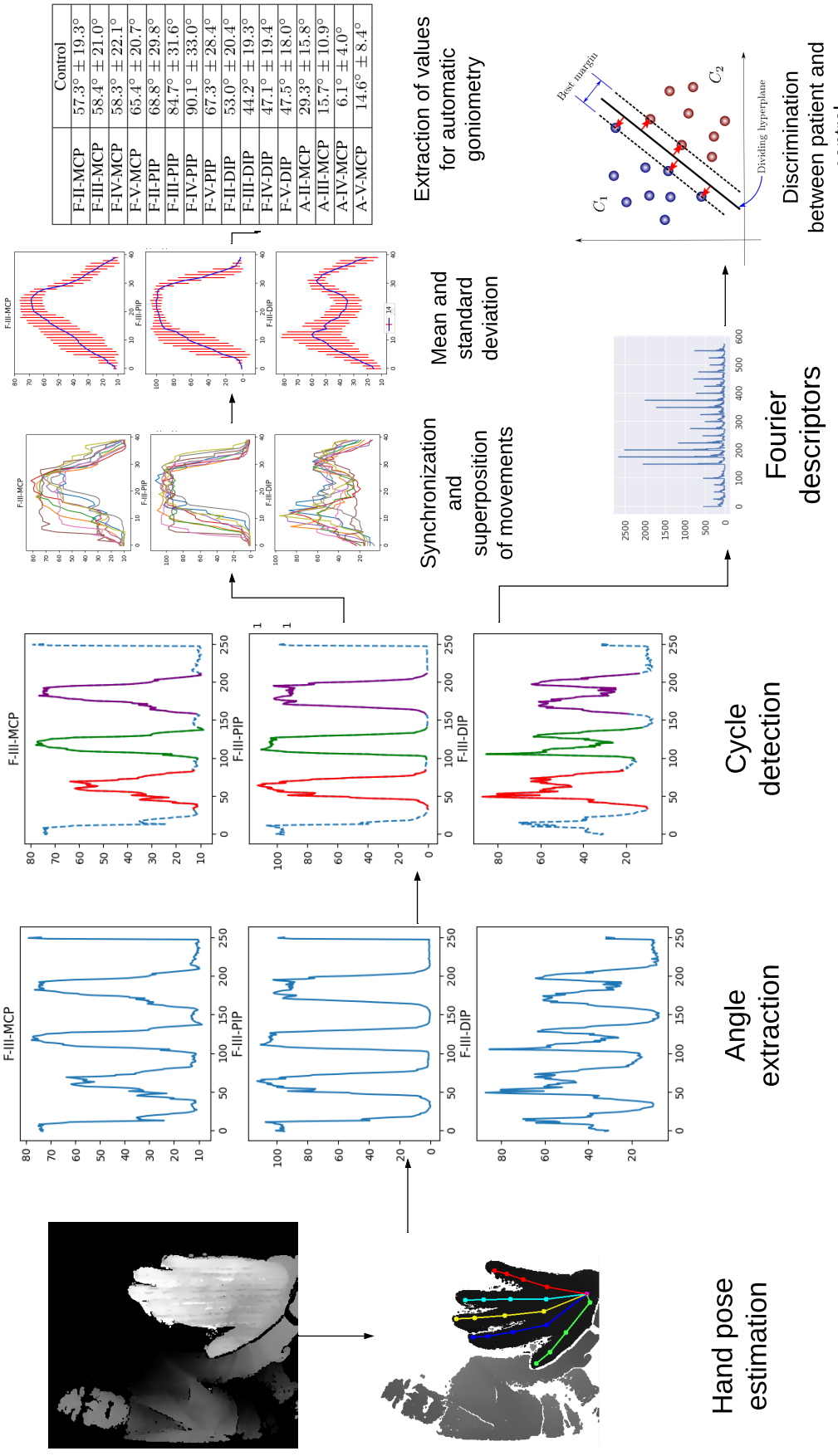
After obtaining the hand skeleton joints with the pose estimation algorithm, our goal is to estimate the range of motion measurements, in order to evaluate the patient and assess its movement capabilities. The diagnosis of the current state of each patient is provided in the form of a table (illustrated by Table 4.1), and the complete hand analysis pipeline is illustrated in Figure 4.3.

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P1 - L	MTC (°)	0	80	-8	96	0	94	-6	92
	IFP (°)	-14	72	-18	88	-36	96	-48	96
	IFD (°)	0	40	0	50	0	28	-12	44
	Abduction (cm)	11.3		8		3.6		3.4	
P1 - R	MTC (°)	0	82	0	102	-8	70	-12	92
	IFP (°)	-24	72	-36	86	-32	76	-48	94
	IFD (°)	0	30	0	44	8	28	0	42
	Abduction (cm)	10.5		4		4.3		3.5	

Table 4.1: *Measurements extracted from one of the patients during the data acquisition session.*

Since the range of motion measurements take into account the flexion and abduction angles, the key step for hand movement analysis is to compute such angles from the skeletons. The process of angle extraction is further described in Section 4.2.1.

With one skeleton per frame, each recorded sequence yields a signal that is composed by time series, one for each estimated measurement. This time series is noisy and contain many movements of flexion or abduction per sequence. We will refer to each cycle of



	Control
F-II-MCP	$57.3° \pm 19.3°$
F-III-MCP	$58.4° \pm 21.0°$
F-IV-MCP	$58.3° \pm 22.1°$
F-V-MCP	$65.4° \pm 20.7°$
F-II-PIP	$68.8° \pm 29.8°$
F-III-PIP	$84.7° \pm 31.6°$
F-IV-PIP	$90.1° \pm 33.0°$
F-V-PIP	$67.3° \pm 28.4°$
F-II-DIP	$53.0° \pm 20.4°$
F-III-DIP	$44.2° \pm 19.3°$
F-IV-DIP	$47.1° \pm 19.4°$
F-V-DIP	$47.5° \pm 18.0°$
A-II-MCP	$29.3° \pm 15.8°$
A-III-MCP	$15.7° \pm 10.9°$
A-IV-MCP	$6.1° \pm 4.0°$
A-V-MCP	$14.6° \pm 8.4°$

Figure 4.3: *Hand movement analysis pipeline. In this Chapter, we cover angle extraction (Section 4.2.1), Cycle detection (Section 4.2.2) and detail the process for discrimination between patient and control (Section 4.4). In Chapter 5 we cover experiments for automatic goniometry (Section 5.1) and the classification patient vs control (Section 4.4).*

flexion or abduction inside a sequence as a *clip*. With that, it is necessary that the beginning and the ending frames of each movement cycle is identified. The algorithm for automatic detection is described in Section 4.2.2. In each sequence, since the duration of each clip can be different, we perform an alignment of all signals extracted for a patient and hand, identifying mean and standard deviation values. With the resulting signal, we can identify peaks and valleys in order to determine the movement capabilities from each joint. This process is detailed in Section 4.3.1. With this, we can determine minimum and maximum values for the flexion and abduction angles of a patient.

4.2.1 Angle Extraction

Using the skeletons $\vec{S}(t)$ obtained by the hand pose estimation method, the analysis aims to obtain measurements of flexion/extension and adduction/abduction. Such measurements are computed for each frame of all sequences obtained in the acquisition. Our ultimate goal is to estimate these angles with accuracy similar to that obtained using manual measurements with goniometers, but in a more efficient and less intrusive way.

The estimation of the flexion angles is obtained by extracting the vectors between the adjacent joints in the structure. For the finger x, the flexion angles from the joints MCP, PIP and DIP are defined respectively as:

$$\widehat{\text{F-x-MCP}} = \arccos\left(\overrightarrow{\text{MCP}_x - W} \cdot \overrightarrow{\text{PIP}_x - \text{MCP}_x}\right) \tag{4.1}$$

$$\widehat{\text{F-x-PIP}} = \arccos\left(\overrightarrow{\text{PIP}_x - \text{MCP}_x} \cdot \overrightarrow{\text{DIP}_x - \text{PIP}_x}\right) \tag{4.2}$$

$$\widehat{\text{F-x-DIP}} = \arccos\left(\overrightarrow{\text{DIP}_x - \text{PIP}_x} \cdot \overrightarrow{\text{fingertip}_x - \text{DIP}_x}\right) \tag{4.3}$$

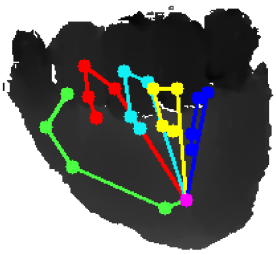
For the thumb, the flexion angles of CMC and IP joints are obtained analogously. Figure 4.4 illustrates the flexion angle computation, highlighting the arcs correspondent to the calculus on an example with a closed hand.

As for abduction, there are some difficulties to compute it because the angle between two phalanx bones actually depends on two systems of joints, rather than a single joint that connects both. This makes it hard to dissociate abduction from flexion angles, particularly on hands with deformities. For this reason, it is common that occupational therapists actually measure abduction by the distance between two consecutive fingertips. We also compute the opening between the fingers, which is not a usual measurement for occupational therapy, but is straightforward and can indicate other types of patterns in a way that is invariant to the size of the hands. The opening angle is computed as the angle between the mean point between the MCP joints of both fingers and each PIP joint.

$$\text{A-x-tip} = \|\text{fingertip}_{x-1} - \text{fingertip}_x\|_2 \tag{4.4}$$

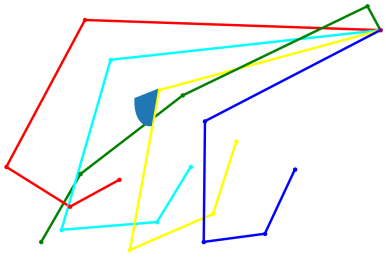
$$OP - x = \arccos\left(\overrightarrow{\text{PIP}_x - mid(\text{MCP}_x, \text{MCP}_{x+1})} \cdot \overrightarrow{\text{PIP}_{x+1} - mid(\text{MCP}_x, \text{MCP}_{x+1})}\right) \tag{4.5}$$

Figures 4.5 and 4.6 show the variation of angle measurements $\widehat{\text{F-III-MCP}}$, $\widehat{\text{F-III-PIP}}$, $\widehat{\text{F-III-TIP}}$ and ABD_3 per frame, computed for all frames in a sequence obtained with a control individual and a patient, respectively. These figures also show the correspondences between given poses and maximum and minimum values on the angle graphics, showing

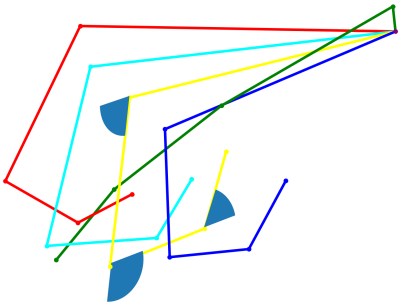


(a) *Skeleton acquired from depth image.*

(b) *Wireframe obtained from the skeleton, illustrating all hand joints.*



(c) *Wireframe highlighting the angle F – IV – MCP.*



(d) *Wireframe highlighting angles F – IV – MCP, F – IV – PIP and F – IV – DIP.*

Figure 4.4: *Example of application of the angle formulae for practical example of a closed hand, detailing the wireframe skeleton and highlighting the joint vectors and correspondent angles F – IV – MCP, F – IV – PIP and F – IV – DIP.*

that the method of hand pose estimation reaches consistent results for flexion movements. For the patient with ulnar deviation, the angle sequences show a higher variability, which is caused by the higher variability of the hand shapes of the patient.

Results obtained from the patient and control hands show that the Pose-REN method is able to generalize for unseen shapes, and despite the inaccuracy for unusual hand poses, the overall performance for angle detection shows that the method can be used in our pipeline.

4.2.2 Cycle detection

Each clip is composed by multiple movements of flexion/abduction. Since the objective of the movement analysis is the identification of the minimum and maximum angles for each patient (see Table 4.1), the proposed approach aims to identify each movement inside the clips, identifying the average minimum and maximum values for each angle. With this, we leverage the presence of outliers in previous steps. Therefore, we defined that after computing the angles, the next step of the pipeline is the identification of the begin and

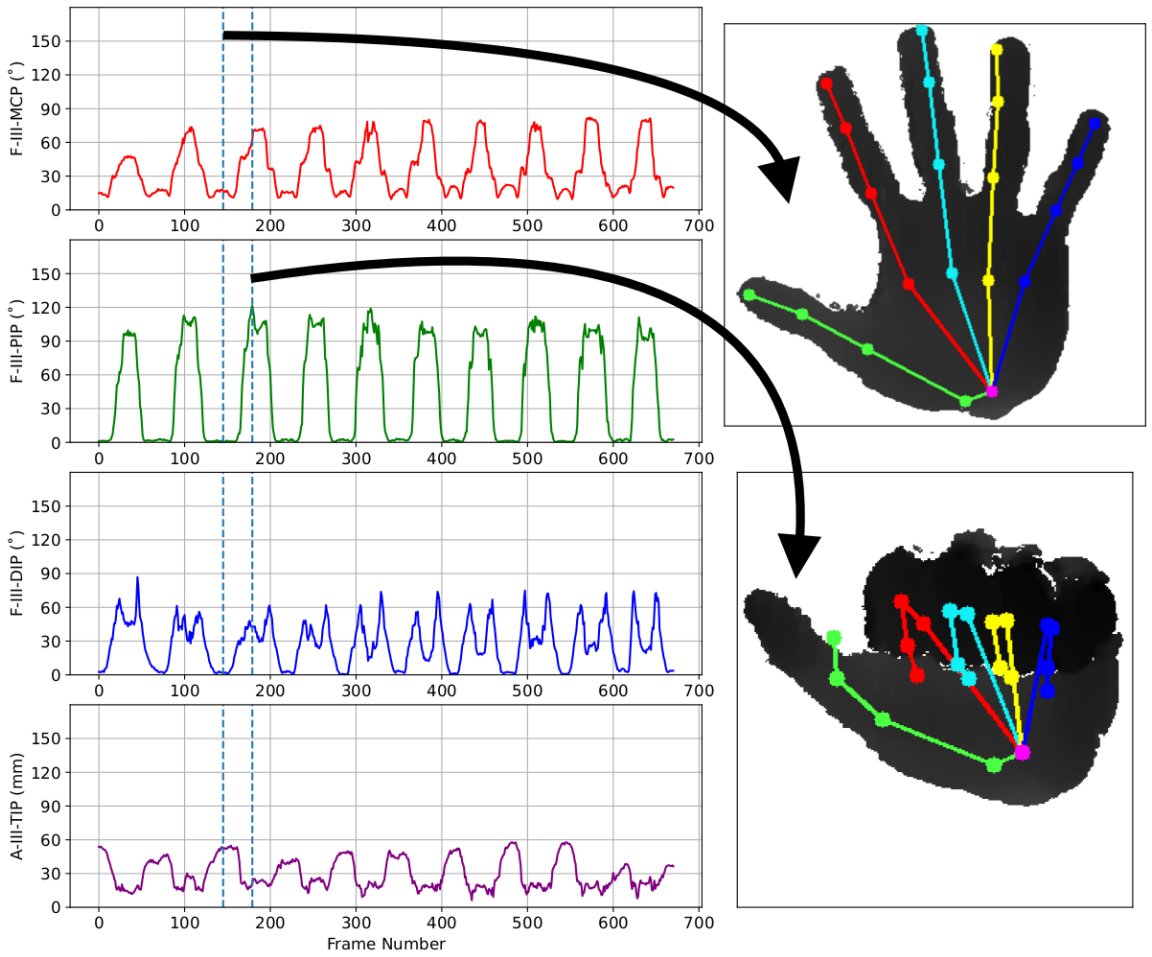


Figure 4.5: *Angle estimates, highlighting correspondences to poses obtained by the pose estimation algorithm on a healthy individual from the control set - Smaller angles represent open hands while larger angles correspond to closed hand poses.*

end of each clip.

In Section 4.2.1 we detailed the angle extraction, showing that in the case of flexion movements the minimum angles correspond to open hand poses and the maximum angles correspond to closed hands (Figure 4.5). Thus, we defined that the movements begin with the open hand and end with the next frame with the open hand, after the closing movement. For abduction/adduction, the clip should contain one cycle of the movement of opening and closing the fingers, starting and ending with the hand with the fingers closed.

For the upcoming analyses, the cycles were manually extracted from the sequences, using a visual tool to mark the frames from beginning and ending. However, such proceeding is not suitable for a real-time pipeline, since it is time-consuming and relies on visual interpretation.

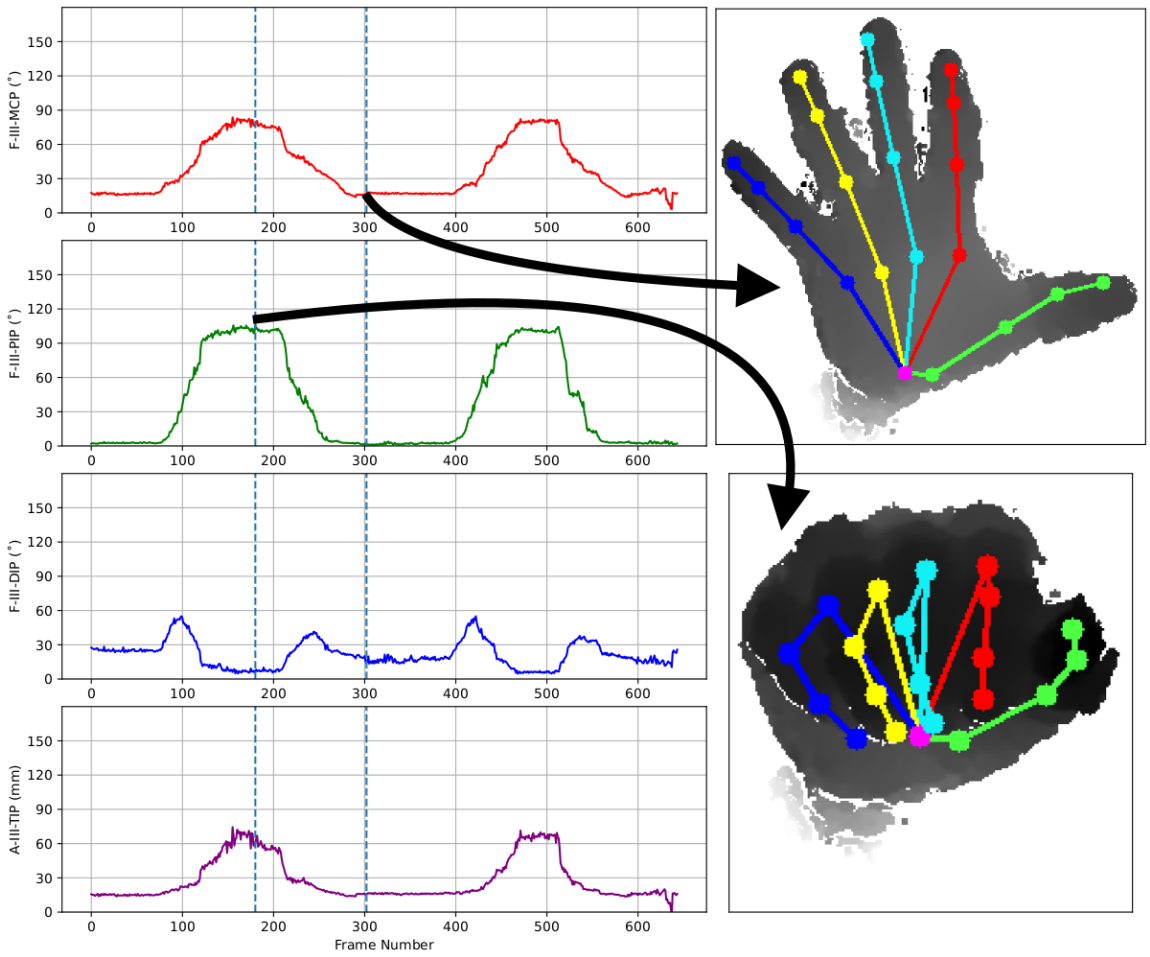


Figure 4.6: *Angle estimates, highlighting correspondences to poses obtained by the pose estimation algorithm on a RA patient - Smaller angles represent open hands while larger angles correspond to closed hand poses.*

4.3 Extraction of values for automatic goniometry

4.3.1 Synchronization and superposition of movements

After the segmentation of the sequence in clips containing one movement, we seek to characterize the range of motion of each joint, considering the multiple clips regarding to the same patient and hand.

Since the length of the clips can be different, the synchronization is made by resampling the angle signals with a standard range. For this, we perform an interpolation in each angle signal, such that the length of each clip is set as 50 frames.

With that, we are able to compute the average value and the standard deviation considering all processed clips for both patients and control set. This result is shown in Figure 4.7. Note that the graphs for different angles have different y-scales. This result shows that both sets follow the same movement pattern and have subtle differences, focused mainly in the beginning of the movement.

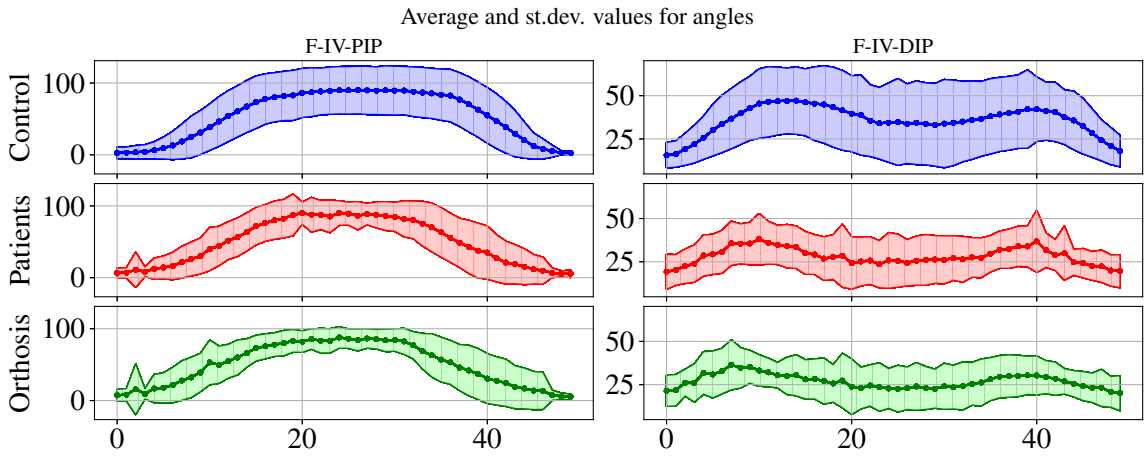


Figure 4.7: *Comparison of average angle values and standard deviations obtained in control set (blue), patients with (green) and without orthosis (red) for flexion movement, for finger 4.*

4.3.2 Results for automatic goniometry

With the synchronization of movements, we can also perform individual measurements for each patient, providing a table similar to Table 4.1 for each patient and hand. The abduction/adduction sequences are analyzed separately, and fill the bottom line with the extent of finger openings of a patient. This comparison is shown in details in the experiment proposed in Section 5.3.

4.4 Discrimination between patient and control

Another possible application for the pipeline is the differentiation of sequences between patients and control. For this, we use the cycles obtained in previous steps and propose the use of Fourier descriptors in order to represent the multidimensional signal. This classification experiment is important to validate whether the current angle extraction pipeline is able to characterize the effect of Rheumatoid Arthritis in the flexion movement pattern.

4.4.1 Fourier descriptors

For all sequences of movement acquired with the patients and with the control individuals we computed the flexion and abduction angles frame by frame, and manually extract the landmark frames in the beginning and in the end of each movement. We will refer to the angle representation of a movement sequence as a *clip*, representing the i-th angle as $a_i(t)$.

For each detected cycle of movement, we normalize the sequences of hand points trajectories by subsampling them, so that all clips have the same duration (i.e., the same number of measurements). The sample representation used for the classification experiment is based on the extraction of Fourier coeficients $\mathcal{F}(i)$ for each angle i. The 25 first coefficients for each angle are concatenated and stored as a sample representation. The

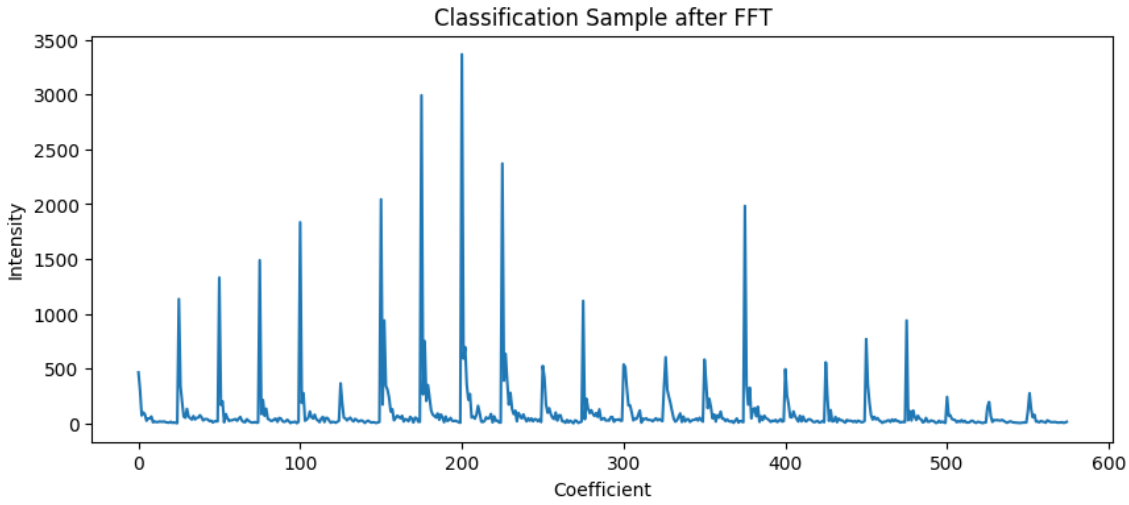
coefficients of the Fourier transform $F_{a_i}(u)$ for an angle a_i are computed by:

$$\mathcal{F}_{a_i}(u) = \frac{1}{N}\sum_{t=0}^{N-1} a_i(t)e^{\frac{-j2\pi ut}{N}}; 1 \leq u \leq 25 \tag{4.6}$$

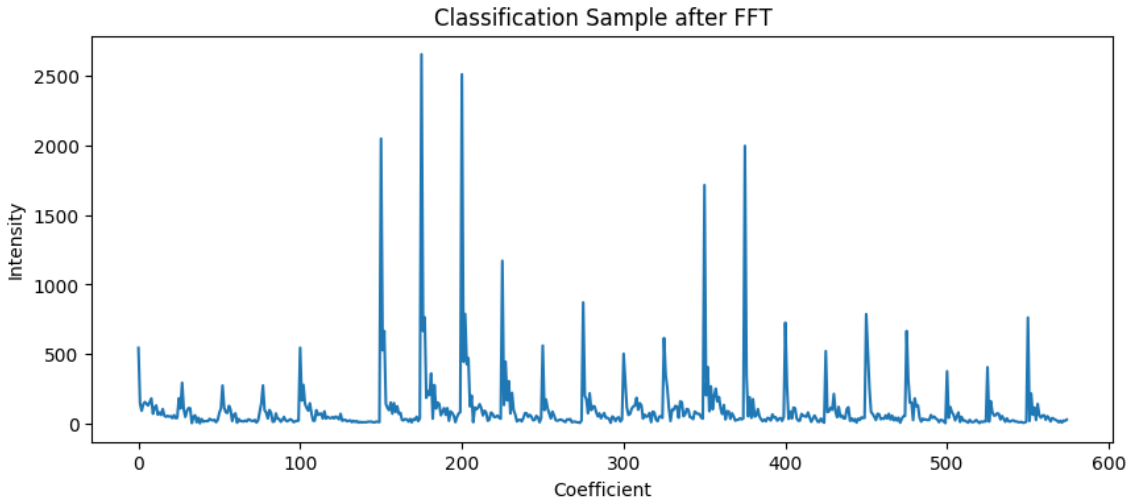
Let N_a be the number of angles computed for each clip. The final representation $\mathcal{F}(u)$ of a clip is the concatenation of all Fourier descriptors of all angles.

$$\mathcal{F} = concat(\mathcal{F}_{a_i}(u)), 1 \leq i \leq N_a, 1 \leq u \leq 25\ . \tag{4.7}$$

Figure 4.8 shows examples of training samples, obtained after the FFT processing. Note that each sample has $25 \times N_a = 575$ dimensions.



(a) *Control example.*



(b) *Patient example.*

Figure 4.8: *Examples of Fourier descriptors, obtained through the concatenation of Fourier descriptors for each angle of a clip.*

4.4.2 Classification

We performed a series of experiments to classify sequences into patients or control. Since the number of samples is small, we defined three types of classification experiments: 80-20% split, leave-one-person-out, and leave-one-person-out with sample synthesis (LOO + SS).

The goal of the initial experiments was to validate the feature extraction method based on Fourier descriptors and to choose an adequate classification algorithm. To validate our method, we defined a baseline descriptor which is built by simply concatenating of the minimum and maximum value of each angle of each joint of the hand.

To choose the classification algorithm, different supervised classifiers have been tried in both Split 80-20 and LOO during the experiments, namely: AdaBoost, Decision Tree, Gaussian Process, Linear SVM, Naive Bayes, Nearest Neighbors, Neural Net, QDA, Random Forest and RBF SVM, using the implementation available in the Python package Scikit-Learn(Pedregosa *et al.*, 2011). The methods selected were:

- **K-Nearest Neighbors** (Goldberger *et al.*, 2004): Assigns to each sample the most frequent value of the K nearest neighbors of the training set. The method tested used K=3.
- **Linear SVM (Support Vector Machine with Linear kernel)** (Platt, 1999): Classifier based on Support Vector Machine, which tries to estimate the hyperplane that maximizes the margin between the classes (distance to the nearest point of the hyperplane).
- **RBF SVM (Support Vector Machine with Radial Basis Function Kernel)** (Platt, 1999): Classifier based on SVM which uses a Kernel feature, in order to add nonlinearity to the hyperplane. The RBF (Radial Basis Function) kernel uses an exponential function. The parameters used are γ and C: γ defines how much influence a single example has, and a high C aims to classify all training examples correctly.
- **Gaussian Process** (Rasmussen and Williams, 2005): Based on Gaussian Process, this classifier estimates the hyperparameters of a prior kernel using the training data, and integrates out the kernel after tuning. The kernel used in the test is RBF(1.0).
- **Decision Tree** (Breiman *et al.*, 1984): Model that learn simple rules from the data features. The main parameter is the tree depth: deeper trees allow more complex rules and more precise models.
- **Random Forest** (Breiman, 2001): Ensemble method based on combining the predictions of many decision tree classifiers, which uses averaging to improve the predictive accuracy. The parameters are the number of estimators (trees in the forest), the max depth of each tree, and the size of the random subsets of features considered when splitting a node.
- **Neural Net** (Hinton, 1990): Classifier method based on a Multi-Layer Perceptron Network. Optimization is done by minimizing the log-loss function using stochastic gradient descent.

- **AdaBoost** (Hastie *et al.*, 2009): AdaBoost classifier optimizes a sequence of weak classifiers, fitting a sequence of repeatedly modified versions of the data. Each modified version of the data multiplies the weight of each training sample, allowing the classifier to focus on the difficult examples on each iteration. The base estimator used is Decision Tree, with max depth = 1.
- **Naive Bayes** (Chan *et al.*, 1982): Implements the Gaussian Naive-Bayes algorithm for classification. This algorithm is based on Bayes theorem with the Naive assumption of independence between each sample. With the assumption, we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i|y)$. The likelihood of the functions is assumed to be gaussian.
- **QDA (Quadratic Discriminant Analysis)** (Ledoit and Wolf, 2004): Classic classifier that fits a quadratic decision surface in the data, generated by fitting class conditional densities to the data and using Bayes' rule.

4.5 Discussion

Section 4.1 presented the method Pose-REN, chosen due to its ease of implementation and generalization capacity in the wild for hand pose estimation, in particular with the model trained on HANDS17 dataset. The HANDS17 dataset, as discussed in Section 2.5, was the first million-scale dataset for hand pose estimation, and the pre-trained model encodes a much larger complexity, which translates into generalization capacity. By using this model, the method is able to perform more robustly in real cases.

Section 4.2 presented all the subsequent steps of the pipeline, including the angle extraction from the skeleton obtained in previous steps, automatic and manual procedures for cycle detection on the signals, extraction of values for automatic goniometry through the synchronization and superposition of movements of the same patient, and the discrimination between patient and control, which used Fourier descriptors to encode angle signals and presented all the classification methods adequate for this task.

In the angle extraction procedure, we are aware that the wrist point is not the correct point for measuring the wrist joint location in the calculation of flexion angles in each finger and should be corrected for each finger, skewing the quality of angle estimatives. Post-processing steps can be proposed to correct those positions according to the position of each bone using the raw depth image as a guide. However, considering the complexity of the hand shapes, this post-processing step poses a challenging problem.

We believe that the pipeline proposed allows further generalization: any hand pose estimation method based on depth input can be used. For angle extraction, we based the measurements on HANDS17 model, whose hand joints are the same as in MSRA model. The hand models used in NYU and ICVL datasets are based on different hand joint positions, and the angle measurements should be computed differently. This is exemplified in Figure 3.10.

Results on the comparison between automatic and manual procedures for cycle detection, comparison between the proposed automatic goniometry and manual goniometry for

patients, and classification between control and patient flexion sequences will be presented in Chapter 5.

Chapter 5

Experimental Results

This chapter presents experiments to assess and validate the proposed method. Section 5.1 presents a signal analysis experiment, illustrating all the steps of the hand analysis pipeline, translating the results as measurement tables, evaluating whether the hand range of motion assessment is meaningful and describing the inherent patterns to the movement. Section 5.2 presents the classification experiments done to classify sequences into patient and control classes, with the goal of showing that the adopted shape descriptors and classifiers are able to encode the difference between the shapes of skeletons. Results show that the use of SVM classifiers and Fourier descriptors reach an accuracy of approximately 90% in classifying sequences into patient and control. Section 5.3 shows a comparative experiment of the Range of Motion measurements obtained automatically from the patients with annotated goniometer measurements.

5.1 Characterization of movement signals

This experiment aimed to evaluate the proposed method for the angle evaluation pipeline, using the methodology described in Section 4.2. We assume that the prerequisite steps were already executed: we used the Pose-REN method with the pre-trained HANDS17 model in all sequences of the dataset to compute hand skeletons in every frame of movement acquired. We adopted the formulae proposed in Section 4.2.1 to compute flexion and abduction angles per frame. The process was made for all movement sequences of the dataset, composed of patients and control subjects.

In this analysis, we show data for joints $F-IV-MCP$, $F-IV-PIP$, $F-IV-PIP$ and $A-IV-tip$. The data visualization for all joint angles and complete figures are available in the Annex C.

The first goal of the experiment was to visually validate the angle measurements. For this, we evaluated the maximum and minimum values of each angle and their correspondant images for examples of the dataset. For the flexion movements, the maximum and minimum values of MCP and PIP angles correspond directly to closed and open hands. Figures 5.1 and 5.2 illustrate the angle evaluation for each frame of a sequence.

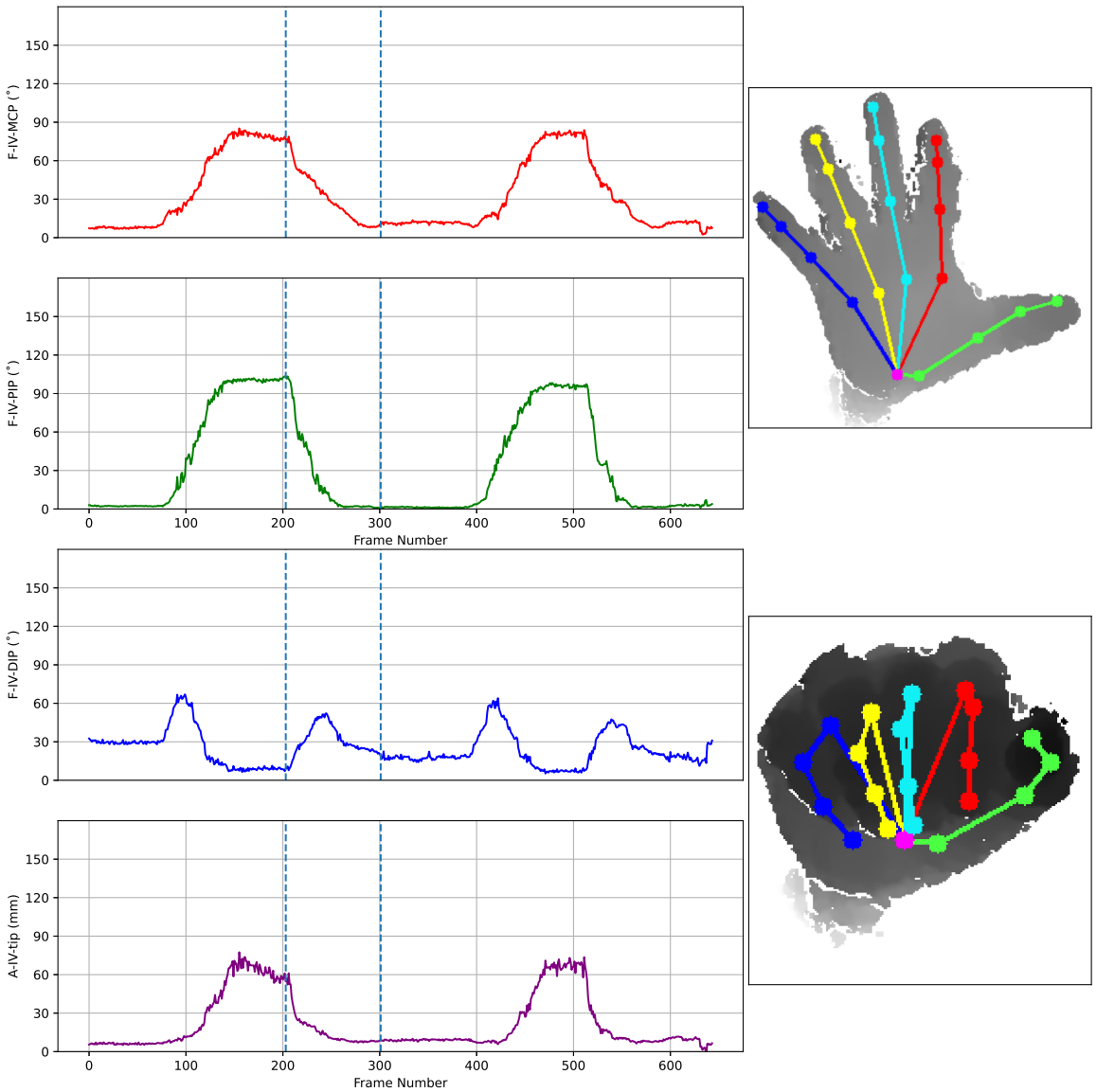


Figure 5.1: *Angle evaluation of a patient. Left and middle columns show graphs with angle joint measurements obtained frame by frame of a sequence acquired following the defined protocol. Right column present frames of maximum and minimum values for the angle $F - IV - MCP$ in the sequence, corresponding to the instants highlighted by vertical dashed lines in the graphs: top image is the lowest angle value and bottom corresponds to the highest.*

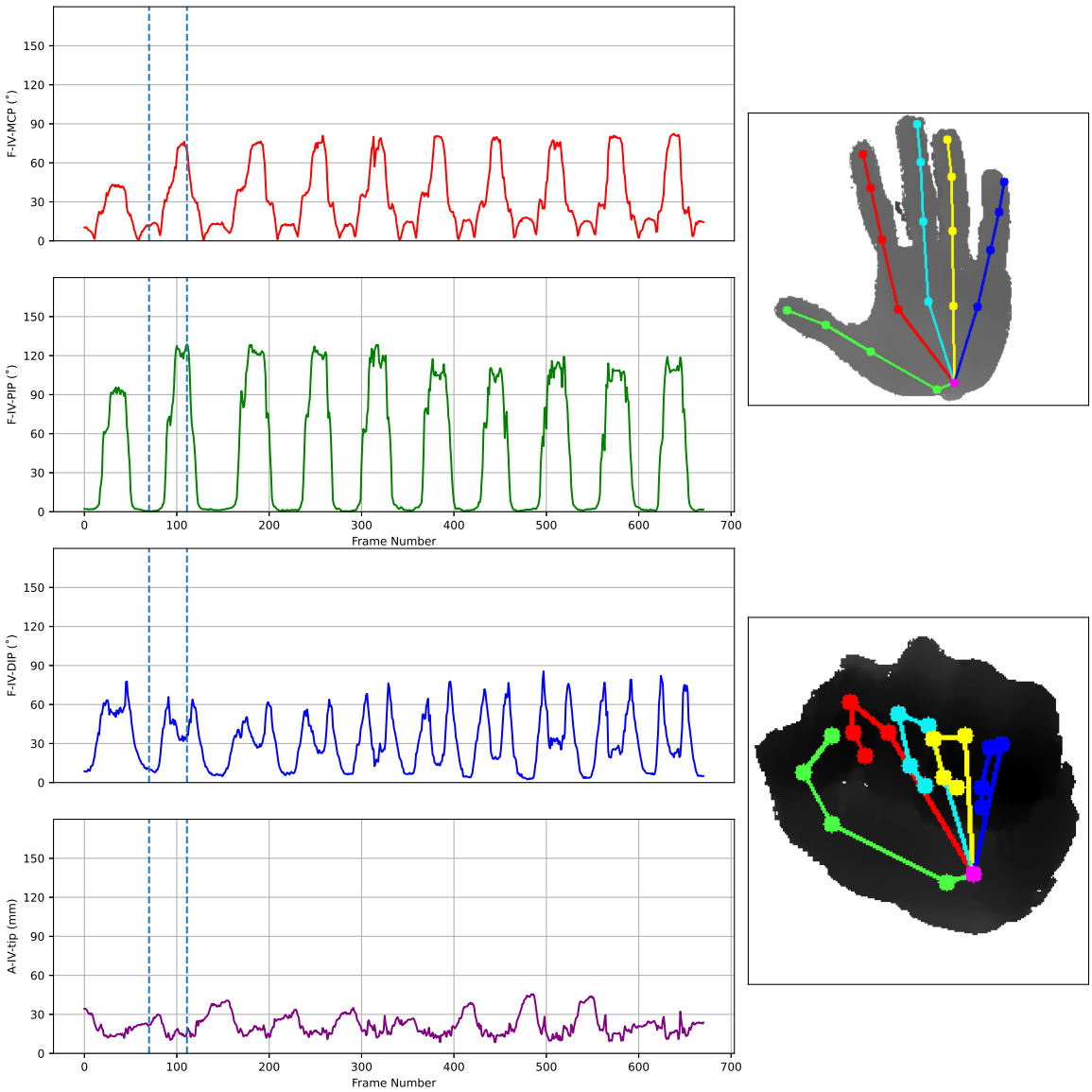


Figure 5.2: *Angle evaluation of an individual in control group. Left and middle columns show graphs with angle joint measurements obtained frame by frame of a sequence acquired following the defined protocol. Right column present frames of maximum and minimum values for the angle $F - IV - MCP$ in the sequence, corresponding to the instants highlighted by vertical dashed lines in the graphs: top image is the lowest angle value and bottom corresponds to the highest.*

Each pulse on the graphs in Figure 5.2 correspond to a flexion movement, and some of the local optima are highlighted. The maximum and minimum values for the angle $F - IV - MCP$ are highlighted and correspond to the hand skeleton frames plotted on the right. The initial analysis of these graphs shows that the angle pattern for open hands correspond to points closer to the global minima for all MCP, PIP and DIP angle measurements. As for the closed hand pattern, MCP and PIP angles present a similar pattern of prominent peak with a local maxima, although in Figure 5.2 the joints corresponding to PIP angles of some cycles present a pattern of irregularity near the local maxima. For the DIP angles, a similar pattern of irregularity near the local maxima is observed in both patient and control sequences. The main differences between patient and control are in the pattern of PIP angle graphics and in shape and magnitude of DIP angles, that reach maximum values of approximately 90° for control and peaks at approximately 60° for patients.

Guided by the visual correspondences between angle signals and frames, we manually annotated the frames corresponding to the beginning and the end of each movement, naming each movement interval as a *clip*. Figure 5.3 highlights the two intervals of flexion movement manually annotated from the movement sequence illustrated in Figure 5.1. The number of clips is variable for each sequence, and in this stage we filtered sequences in which the hand pose estimation result was inaccurate.

The next validation step was to obtain a summary of measurements for each individual. For this, we resampled all clips, representing them by an interpolated version of 50 frames. With the resampling, we are able to deal with movements of different speed, comparing and grouping sequences with different length. Figure 5.4 shows the superposition of clips in each of the previous sequences illustrated in Figures 5.1 and 5.2. Figure 5.5 shows the clips extracted from all flexion sequences of the same subject.

Analyzing Figure 5.4, it is possible to note the general patterns of flexion movement for the same subject. In nearly all sequences we perceive the peak in MCP and PIP angles and the inflexion in DIP angles. In Figure 5.5, the pattern maintains itself, but in the patient sequences we can perceive two groups of signals in MCP and DIP angles, which leads to a higher standard deviation for the average signal.

The final step is to consolidate the results for each subject: given that all clips were annotated and resampled to the same length, we obtain average and standard deviation values for each angle in each frame of the resulting signals, thus characterizing the flexion movement for the subject. Figure 5.6 shows the average and standard deviation for the control and patient individuals illustrated in previous figures, with the mean signals for each individual shown in the background colored by the respective classes.

Figure 5.6 shows that the general tendency maintains the characteristics observed at the previous examples for the majority of patient and control subjects. Those classes behave similarly, with the main differences occurring in the DIP joint (fifth and sixth rows of the Figure) - in some control subjects the pattern observed is a slope, differing from the previously observed inflexion behavior, and the angles reach higher values in these cases. In order to provide a complete objective feedback with angle values for each patient, we need the abduction/adduction measurement. For this, we perform a similar procedure on abduction sequences, resampling and grouping the sequences of the same subject, in order

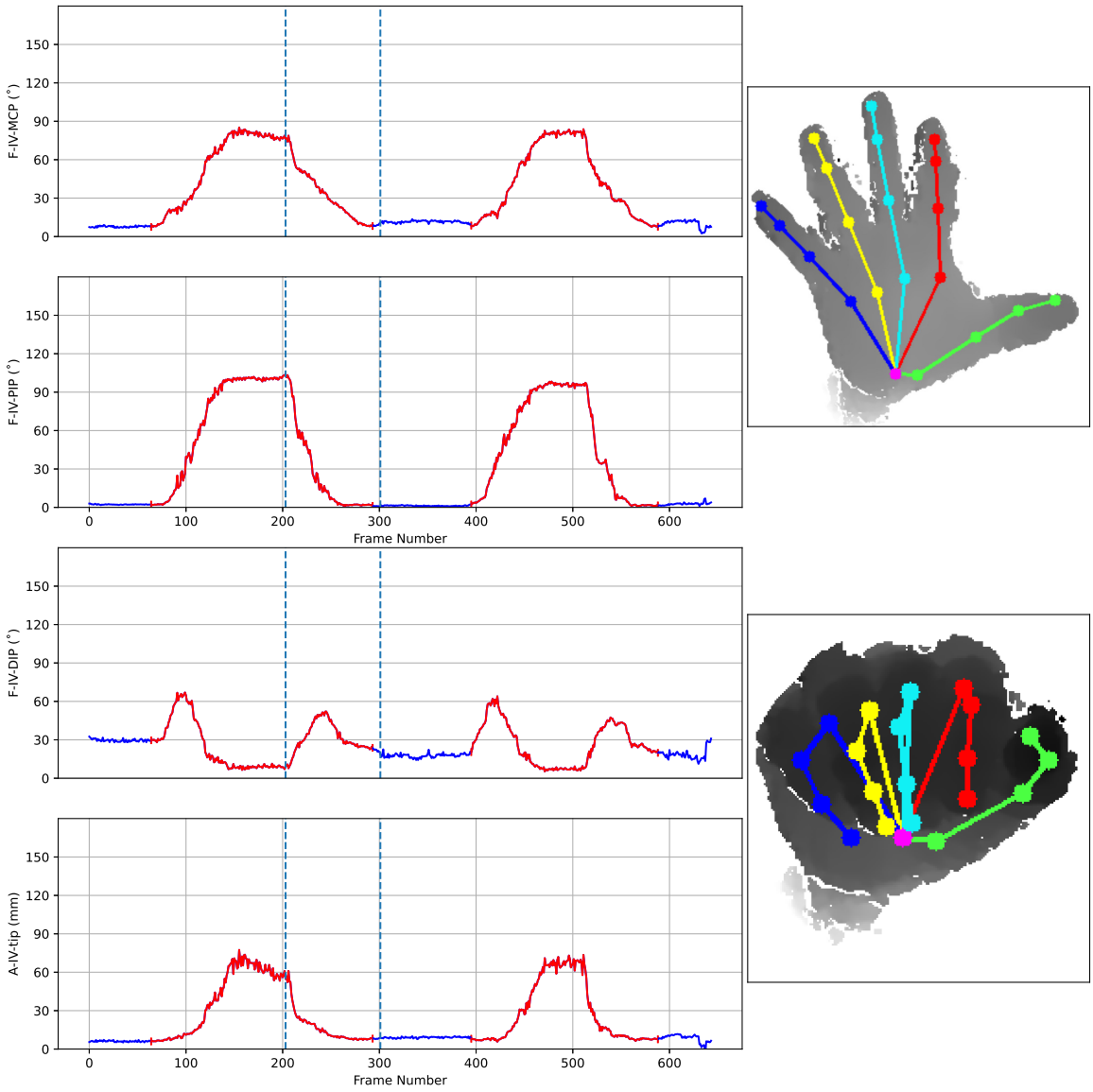


Figure 5.3: *Manual annotation of movement intervals in the angle sequence described in Figure 5.1. Extracted clips are marked in red.*

to extract mean values for each moment of the movement. Using flexion and abduction assessments, we are able to compose Tables 5.1 and 5.2 , which are built in the same way as Table 4.1.

Comparing these tables, it is noticeable that the angles of the control subject in left and right hands have similar magnitudes, with subtle variations on the IFP joint. For this subject, abduction range was not computed because there were no clips for abduction movement. For the patient (Table 5.2), the difference between the MTC angles are much more noticeable between hands, in which the left hand (with AR) has limited range and the maximum angle is lower. This is a pattern observable in the results of some patients, indicating that RA affected one of the hands more than the other. This observation is highlighted in Table 5.2. The individual measurements for each patient are available in Appendix B, and are summarized in Figures 5.7, 5.8, 5.9, 5.10.

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C05 - L	MTC (°)	18.97	**68.36**	15.77	**67.47**	10.01	**68.18**	12.53	**72.16**
	IFP (°)	2.51	90.97	1.63	109.98	1.23	117.82	2.54	88.56
	IFD (°)	3.40	60.52	3.41	48.88	6.98	54.49	3.79	53.28
	abd (cm)	0.00		0.00		0.00		0.00	
C05 - R	MTC (°)	18.45	**69.91**	14.89	**76.30**	7.94	**79.83**	7.88	**82.08**
	IFP (°)	2.42	94.76	1.89	101.86	1.32	109.96	1.62	87.97
	IFD (°)	2.71	64.67	2.85	57.60	7.75	60.19	4.35	57.12
	abd (cm)	0.00		0.00		0.00		0.00	

Table 5.1: *Measurements extracted from one of the control subjects. Highlighted values indicate maximum MCP flexion angles for both hands, which for control subjects in general are comparable.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P07 - L	MTC (°)	15.74	**44.05**	10.56	**47.77**	4.91	**47.51**	6.76	**58.84**
	IFP (°)	0.74	85.21	1.23	98.16	1.37	91.04	1.08	71.82
	IFD (°)	19.45	47.15	17.70	43.76	16.36	48.84	13.13	54.07
	abd (cm)	2.52		1.66		1.69		2.50	
P07 - R	MTC (°)	23.63	**76.20**	14.88	**80.43**	9.05	**81.23**	7.89	**80.91**
	IFP (°)	0.94	88.63	2.76	103.09	1.49	99.76	1.74	83.98
	IFD (°)	20.45	48.65	6.66	45.11	8.06	54.33	18.23	56.34
	abd (cm)	2.87		2.48		1.34		0.98	

Table 5.2: *Measurements extracted from one of the patients during the data acquisition session. Highlighted values indicate maximum MCP flexion angles for both hands. In this patient specifically such values are much different between left and right hand, which reflects different rheumatoid arthritis stages for each hand.*

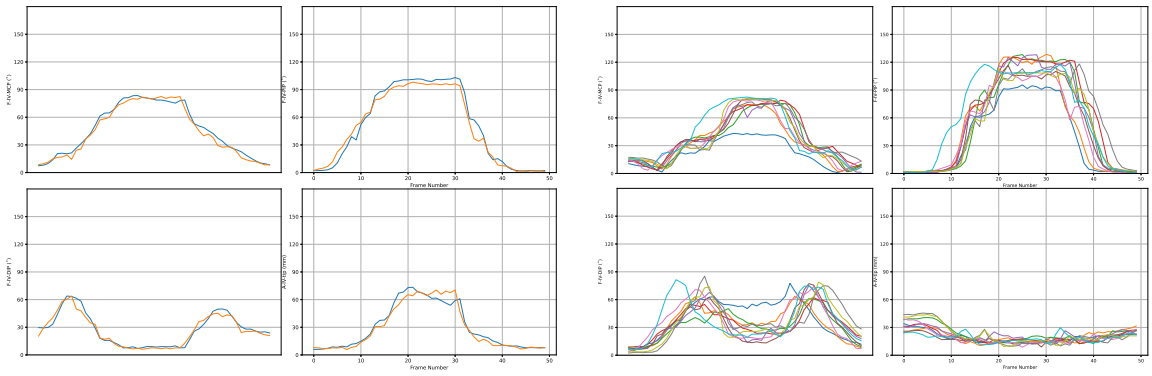


Figure 5.4: *Extracted clips from sequences shown in Figure 5.3: patient (left) and control (right). Trajectories have been re-sampled.*

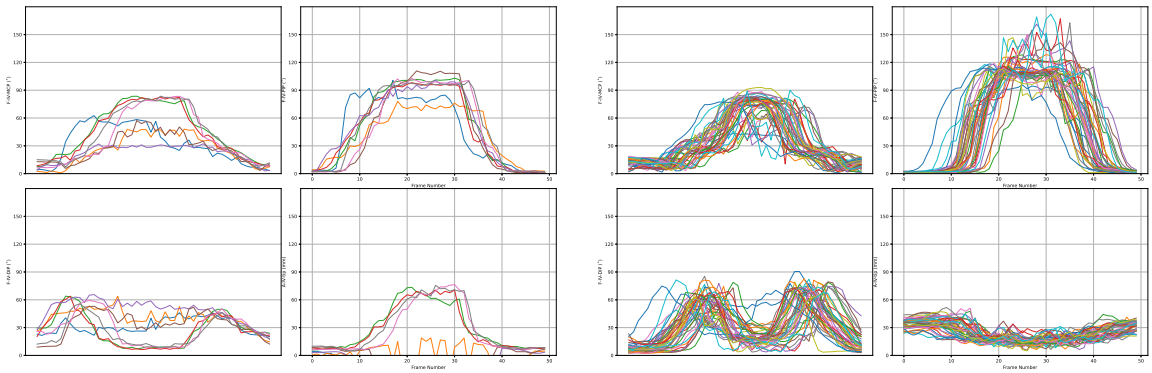


Figure 5.5: *All trajectories extracted from clips of the same person: patient (left) and control (right). Trajectories have been re-sampled.*

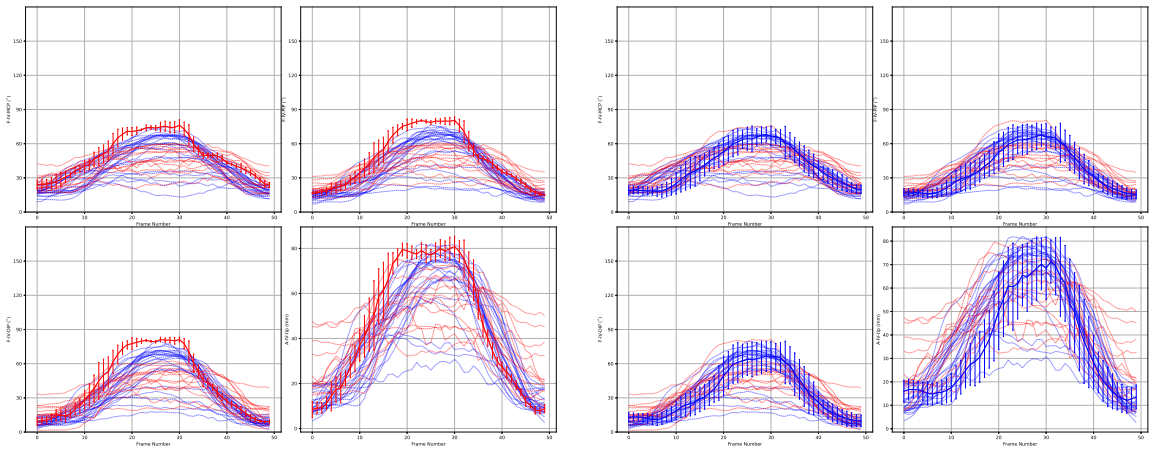


Figure 5.6: *Summarization in terms of mean and standard deviation of all trajectories extracted from clips from the same person: patient (left) and control (right). "Average clips" of other subjects are shown in the background. Patient samples are colored in red, and control samples are colored in blue.*

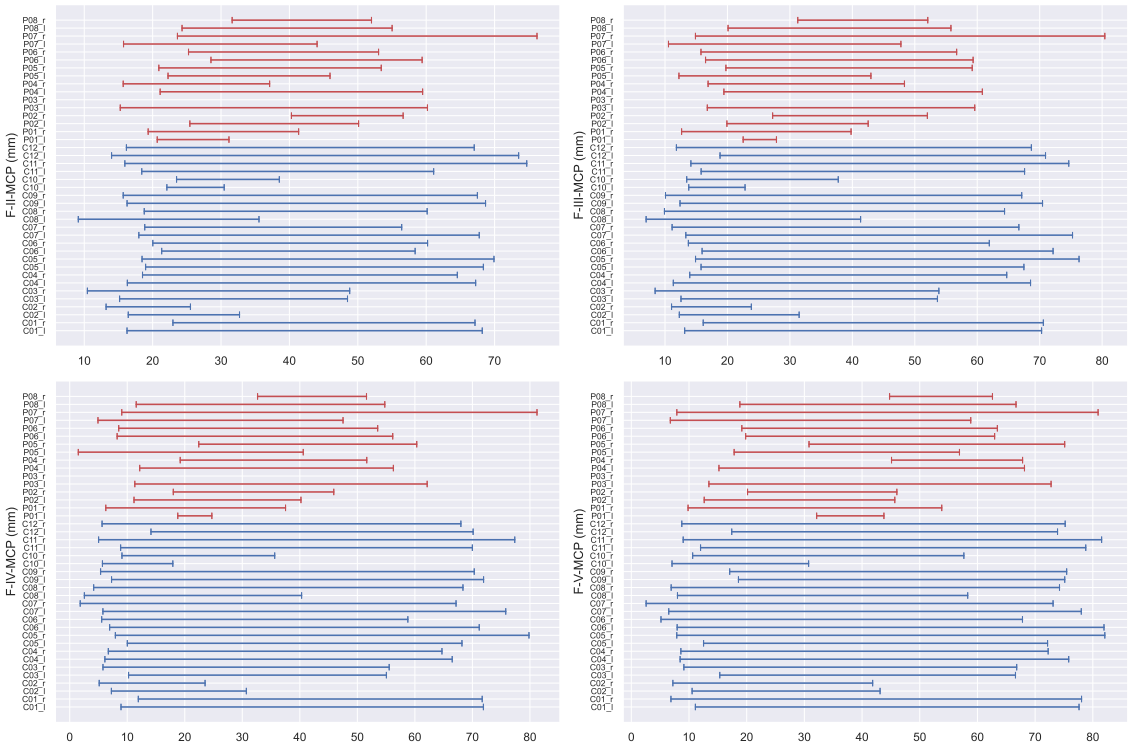


Figure 5.7: *Summarization of the average minimum and maximum values for MCP joints in all subjects of the dataset. Patients are identified in red and control subjects in blue.*

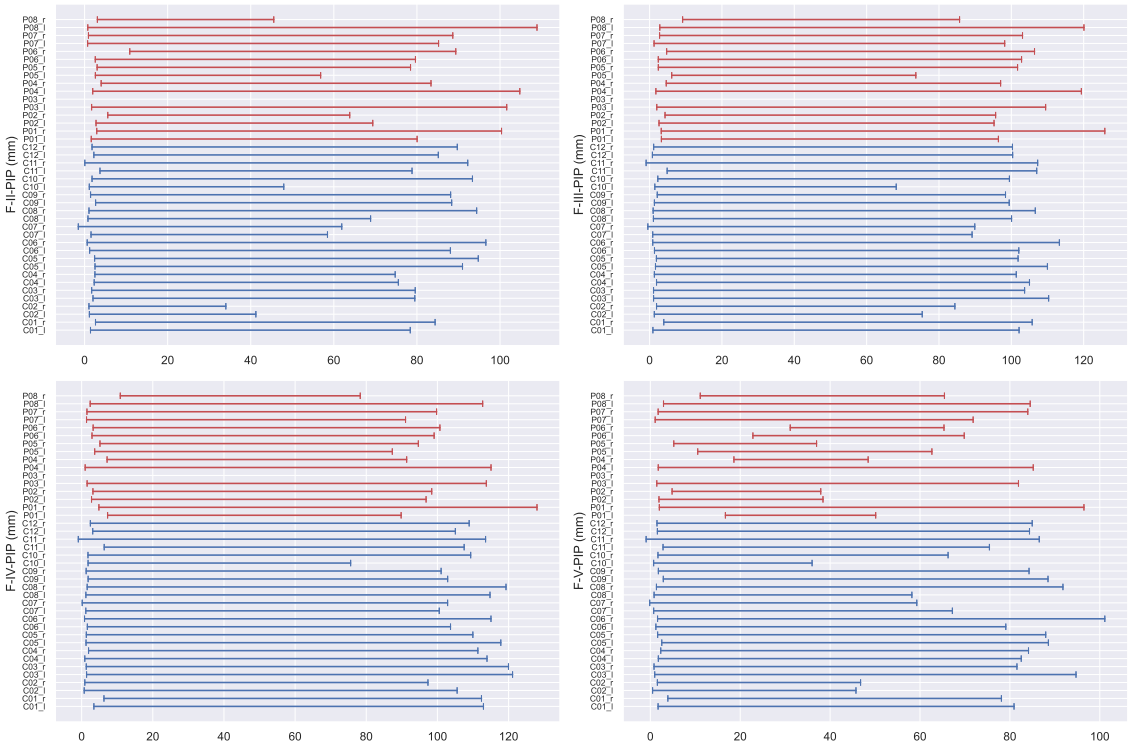


Figure 5.8: *Summarization of the average minimum and maximum values for PIP joints in all subjects of the dataset. Patients are identified in red and control subjects in blue.*

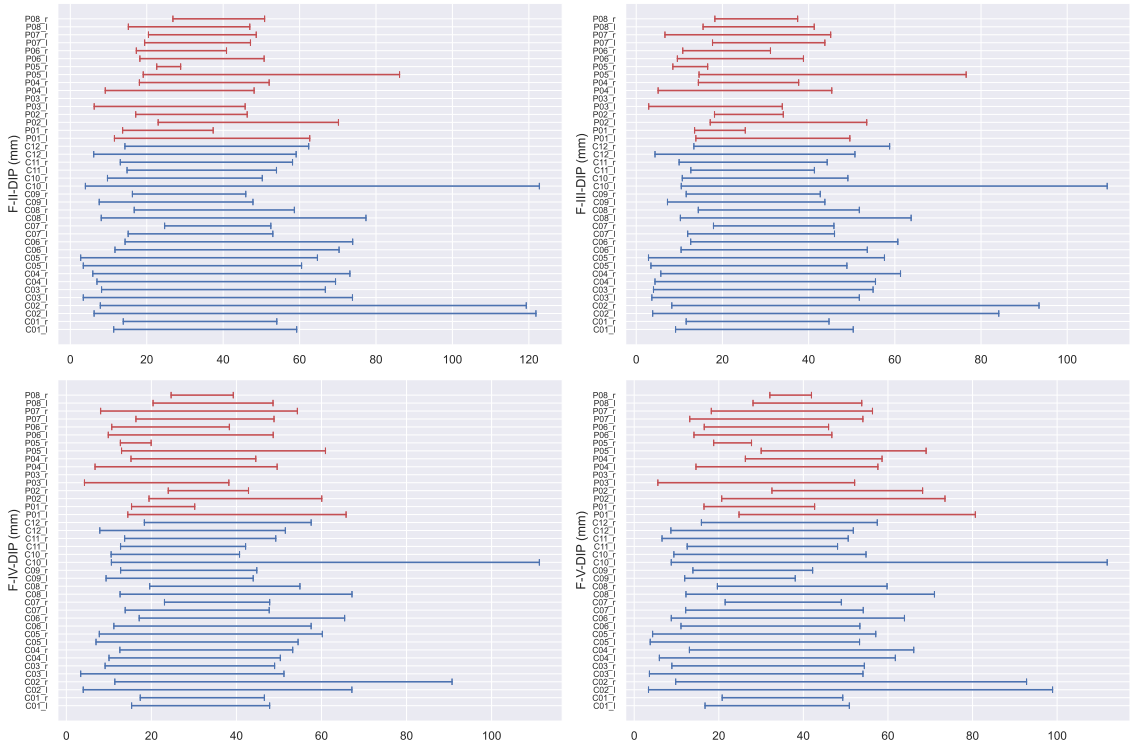


Figure 5.9: *Summarization of the average minimum and maximum values for DIP joints in all subjects of the dataset. Patients are identified in red and control subjects in blue.*

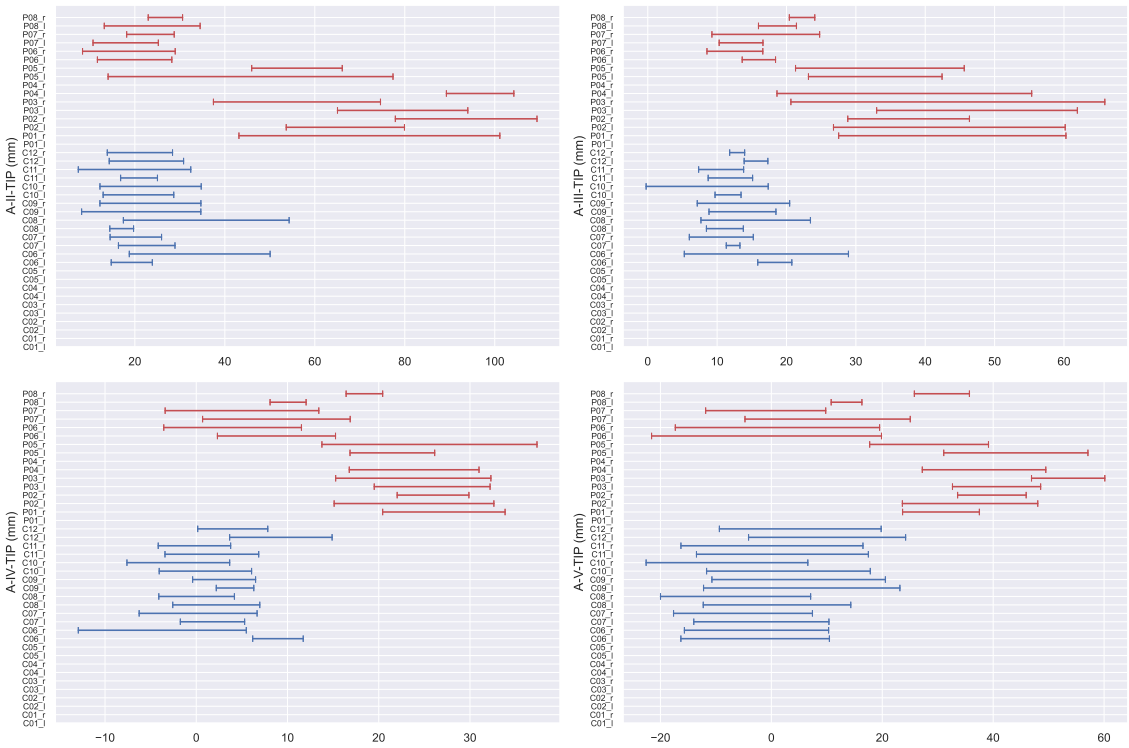


Figure 5.10: *Summarization of the average minimum and maximum values for abduction in all subjects of the dataset. For these measurements only the sequences with abduction movement were considered. Patients are identified in red and control subjects in blue.*

From Figures 5.7, 5.8, 5.9 and 5.10, we observe that the tendency of variation between hands occurs in the majority of RA patients, especially in MCP and DIP joints. This is reflected by the variance between consecutive red rows in Figures 5.7 and 5.9, which contrasts to the much lower variance between consecutive blue rows. We also observed a big variation in abduction measurements, which can be explained by the nature of the abduction measurement used in this work. Abduction is measured as the difference between consecutive fingertips, in millimetres. Such measurement tends to present a high variability for different sizes of hands, which is reflected in Figure 5.10 and the differences between intervals observed in the lines of the graph, where in previous figures such difference is smaller. We can also observe that patients (in red) in general reach a higher maximum value and present higher differences between both hands (subsequent lines of the graphs) for abduction.

Another possible application with the measurements obtained is to characterize patients and control subjects in general. This is done in similar fashion to the previous procedure, grouping all control clips and all patient clips and extracting mean and standard deviation values. Figure 5.11 present the "average clip" with error bars for each set, illustrating that the general behaviour is similar among patient and control subjects, with subtle variations on the magnitude of the standard deviation for MCP and DIP angles. For the patient signals, it is noticeable that the minimum value for almost every flexion angle is higher than the average minimum value for control signals. The biggest difference, however, is in abduction, which presents a much higher variability for patients, especially at the joint II-TIP.

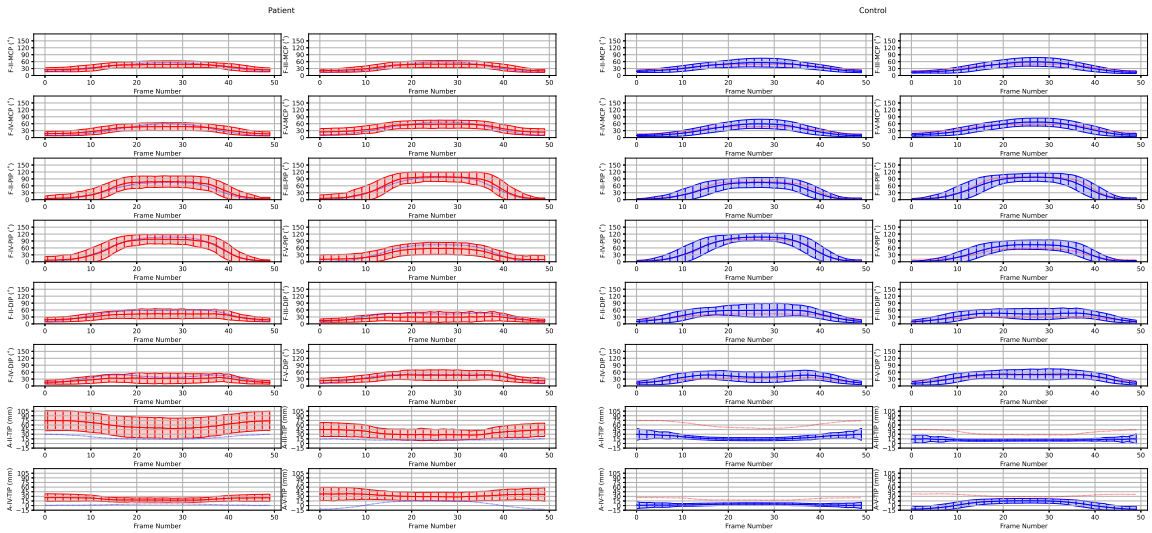


Figure 5.11: *Summarization in terms of mean and standard deviation of patient set (left) and control set (right). Patient set contains all clips extracted from patients and show a slightly higher variability. Control set contains all clips extracted from the control group.*

5.2 Classification

In order to provide an initial step in the use of data from whole movement sequences for analysis, we performed a series of experiments to classify sequences into patients or control. These experiments aim to validate the angle extraction pipeline proposed in this work for the classification task. For this experiment, we focused solely on flexion movement sequences. Ideally the motivation for classification was to provide a "grading" system that evaluates the state of the disease in each patient. However, given the small amount of clips and the absence of specialized hand pose estimation methods for the context, this classification is done to prove whether simple descriptors and classifiers can separate the results provided by the hand pose estimation setup in two simple classes, establishing the feasibility of such analysis. With enough data, more details can be added to the classification pipeline and the original grading system idea can be implemented.

In order to show that movement descriptors can enhance the accuracy of classifying between control and patients by taking into account the dynamic aspect of the movement, we proposed the computation of Fourier descriptors for each clip. The current evaluation assessment method extracts maximum and minimum angle metrics, and by correctly classifying sequences instead of key frames, the occupational therapy community can work on more complex input data and discover different types of features that characterize the disease. A future possibility is the use of the classification as a decision support tool in the context of telemedicine and remote diagnosis.

For this experiment, we tested different classification algorithms, in order to select the methods that yielded better results, and compared the feature extraction method based on Fourier descriptors (described in Section 4.4.1) with a baseline descriptor built by simply concatenating the minimum and maximum values from each computed angle. This process is described by the following equation:

$$\mathcal{B} = concat(\min(a_i), \max(a_i)), \tag{5.1}$$

for $i = 1, \cdots, N_a$.

Since the number of samples is small, we defined three types of classification experiments: 80 – 20% split, leave-one-person-out, and leave-one-person-out with sample synthesis (LOO + SS). We performed paired experiments with both baseline and Fourier descriptors, using Split and Leave-one-person-out strategies.

- Split (80-20): since the sample shuffling can affect the data distribution, we perform 10 instances of classification, each with a random split of 80% of the samples for training and the remaining for testing. We then report the mean and standard deviation of the accuracies obtained.

- Leave-one-person-out (LOO): we choose one person and take all clips from that person as the test set. Training is done with all other sequences. This test shows whether the pattern obtained from a patient or a control subject can generalize well for unseen subjects. We grouped the results in control and patient groups, showing the mean and standard deviation of the accuracy of both groups.

- Leave-one-out with sample synthesis (LOO + SS): random noise is applied to sequences of the dataset, generating new sample sequences and balancing the training and test set. After this process, we apply the leave-one-out strategy, by selecting a subject for testing and training with sequences of all other subjects. Different levels of noise and train set sizes have been tested, in order to validate whether small inaccuracies from the hand pose estimation method affect the angle estimation and the overall patient/control classification, and the results are compared to LOO average results. This experiment is decribed in details in Subsection 5.2.1.

Additionally, different supervised classifiers have been tried in both Split 80-20 and LOO during the experiments, namely: AdaBoost, Decision Tree, Gaussian Process, Linear SVM (Support Vector Machine), Naive Bayes, Nearest Neighbors, Neural Net, QDA (Quadratic Discriminant Analysis), Random Forest and RBF SVM (Support Vector Machine with Radial Basis Function Kernel). A brief description of each classifier is available in Section 4.4.2. Among the classifiers, the Linear SVM presented the best performance. The results of both experiments are shown in Tables 5.3 and 5.4.

The best combination of classifier and descriptor in both experiments was the Linear SVM with the Fourier descriptor, reaching an accuracy of 94.1% in the Split 80-20

Experiment (%)	Control (%)	Patient (%)	General (%)
Fourier Linear SVM	**96.31** ± 3.07	**91.97** ± 6.55	**94.14** ± 5.56
Baseline QDA	96.66 ± 3.81	89.11 ± 6.82	92.88 ± 6.69
Fourier Nearest Neighbors	97.08 ± 3.42	86.31 ± 9.39	91.69 ± 8.88
Baseline AdaBoost	94.99 ± 4.29	83.08 ± 12.56	89.04 ± 11.11
Baseline Neural Net	88.87 ± 8.93	88.65 ± 8.03	88.76 ± 8.49
Baseline Linear SVM	92.72 ± 3.69	84.60 ± 7.51	88.66 ± 7.17
Fourier AdaBoost	95.86 ± 1.51	78.97 ± 11.65	87.41 ± 11.85
Fourier Neural Net	94.37 ± 5.88	78.50 ± 13.42	86.44 ± 13.05
Baseline Nearest Neighbors	95.30 ± 4.68	73.79 ± 13.97	84.54 ± 14.97
Baseline Random Forest	98.96 ± 1.62	65.33 ± 15.04	82.15 ± 19.93

Table 5.3: *Best performance classifiers on the Split experiment (in percentage of accuracy).*

Experiment (%)	Control (%)	Patient (%)	General (%)
Fourier Linear SVM	**94.33** ± 10.53	**81.57** ± 31.34	**89.63** ± 21.67
Fourier Neural Net	92.89 ± 12.01	73.11 ± 36.33	85.60 ± 25.85
Baseline AdaBoost	89.54 ± 15.54	74.07 ± 30.09	83.84 ± 23.28
Baseline Linear SVM	89.08 ± 20.29	74.57 ± 33.44	83.74 ± 26.85
Baseline Neural Net	87.35 ± 22.97	73.91 ± 35.87	82.40 ± 29.14
Fourier AdaBoost	91.92 ± 8.93	65.71 ± 34.74	82.26 ± 25.59
Fourier Decision Tree	90.76 ± 10.56	62.01 ± 28.66	80.17 ± 23.78
Baseline QDA	90.58 ± 17.43	59.71 ± 38.39	79.21 ± 30.93
Baseline Random Forest	95.82 ± 9.59	49.71 ± 40.60	78.83 ± 34.06
Fourier Nearest Neighbors	92.50 ± 16.03	53.64 ± 40.62	78.18 ± 33.49

Table 5.4: *Best performing classifiers on the leave-one-person-out experiment.*

experiment, and 89.6% in the leave-one-person-out experiment. It is worth mentioning that, except for some specific cases, most classifiers did not perform much worse than the SVM results here reported. Our interpretation is that the proposed hand tracking and angle measurements successfully capture the differences between control and patient movements in a robust way. Therefore, the classification task itself does not critically depend neither on the features nor on the classifier, which is a good advantage of the proposed framework.

For control subjects, the accuracy reached in the majority of methods is high, surpassing 90% with low standard deviation in most cases. The main differences can be seen in the patient set, whose accuracy varies between 65% and 91% in the split experiment, and between 49% and 81% in the leave-one-person-out experiment. Although the Fourier Linear SVM was the method that performed better in both experiments, the methods Baseline QDA and Fourier Nearest Neighbors, which presented competitive results in the Split experiment, reported a lower accuracy in the LOO experiment: Baseline QDA varied from 92.88% to 79.21%, and Fourier Nearest Neighbors varied from 91.69% to 78.18%, with a loss of approximately 13% for both methods between experiments that indicates difficulties when dealing with unseen subjects. The Fourier Linear SVM method presented a more robust behavior, with a variation from 94.14% to 89.63%, losing approximately 5% between experiments and with an average accuracy 4.04% higher than the second best method, Fourier Neural Net.

The Fourier descriptor was consistently better than the baseline descriptor, with an average difference of 5%. The baseline result reached the average accuracy of 84% in the leave-one-person-out, which indicates that the minimum and maximum angles are important measurements and can be used to identify patient and control. However, the information added by Fourier descriptors is able to consistently improve the performance, working as a fine-tuned descriptor.

The high accuracy of the Linear SVM in both experiments is a good indicative, especially in the leave-one-person-out, which shows that the descriptor can be generalized for unseen subjects. For patients, the accuracy was slightly lower, which is expected as the data is more diverse, since each patient's hand is in a different stage of ulnar deviation. This higher variance in hand shapes and movement patterns creates data clusters that are more challenging for the classifier.

5.2.1 Data generation with sample synthesis

One important issue in the classification experiment is that the dataset is composed by 581 control clips and 310 patient clips, as described in Table 3.7. This poses the dataset as a slightly imbalanced dataset, which is usually biased towards the majority class (Burnaev *et al.*, 2015). Common strategies to deal with this issue are undersampling of the majority class, oversampling of the minority class and data augmentation / sample synthesis techniques.

For the third experiment we performed sample synthesis (SS) (Dougherty *et al.*, 2002) to address the imbalance between the amount of samples from patients and control. In this process, we generate synthetic data from the samples, enabling us not only to deal with

data imbalance but also to evaluate the results of of our analysis method in the presence of hand pose estimation noise. For that, we applied Gaussian noise for each joint position in the skeleton. One could question why we have not applied standard data augmentation strategies on the depth maps, instead of injecting noise on the pose estimation data. The data augmentation strategies used in other computer vision applications usually follow two strategies: (a) RGB value perturbations (such as changing brightness, contrast, injecting Gaussian noise, etc.) and (b) homography transformations, cropping and padding. These strategies cannot naïvelly be applied to depth maps for the following reasons:

(a) The behaviour of noise in depth maps is different from that of pixel RGB values. The noise from depth sensors that are based on active infra-red patterns tend to alternate between Gaussian-like patterns and patches with unknown depth values. In fact, Chenggang Yan, Li, *et al.* (2020) have exploited the intrinsic low-rang and self-similarity property of depth images to propose a denoising method. A proper data augmentation method should start from a 3D scene model and apply a transformation that would do the inverse of what the method of Yan et al. does.

(b) Homography-based distortions would be unrealistic for our data acquisition setting and would generate on unexpected depth values. An alternative would be to geneate a 3D point cloud from each depth map, perturb the 3D position of each point and re-generate depth maps by ray tracing and interpolations.

There are works in the literature that discuss the best ways of augmenting depth maps, for problems that are different than hand pose estimation, such as depth completion (Hammond, 2019). However, the complexity of such methods would certainly slow down the training process. Furthermore, our depth maps were acquired in realistic non-ideal conditions (particularly when the patients were wearing orthoses). This means that the depth maps already had a noise that is very typical of that kind of sensor and those conditions. We therefore believe that there was no need to inject further noise on depth maps to synthesize new samples. Instead, we focused our sample synthesis method on modelling potential imperfections of the hand pose estimation method (rather than on the depth maps), which is why it was more sensible to inject noise on the resulting 3D point positions. This is in line with other papers about pose estimation methods: many of them make use of skeletons and model priors for sample synthesis and for the training process (Zhang *et al.*, 2020; Wu *et al.*, 2020; Molchanov *et al.*, 2015).

Therefore, for a sequence

$$\vec{S}(t) = \{x_i(t), y_i(t), z_i(t)\}$$

for $i = 1, \cdots, N_j$ and $t = 1, \cdots, T$, we generate the augmented sequence

$$\vec{S'}(t) = \{x_i(t) + \mathcal{N}(0, \sigma), y_i(t) + \mathcal{N}(0, \sigma), z_i(t) + \mathcal{N}(0, \sigma)\} \ ,$$

where $\mathcal{N}(\mu, \sigma)$ represents a Gaussian function with μ mean and σ standard deviation, measured in millimeters. This procedure is applied in each frame to generate new clip samples.

In this sense, we augmented the training and the testing sets and performed the Leave-

one-person-out experiment. We analysed the variations of control and patient sets and evaluate how the Gaussian noise affects the classification accuracy. For each patient hand and noise magnitude, we generated 100 augmented samples from the original sequences. On the composition of the training set, we used different sample sizes, of 100 and 400 samples of the remaining patients, such that this set is composed by the same number of control and patient samples randomly chosen among the augmented dataset. The test set is composed of all 100 samples of the unseen patient. We repeated the leave-one-out experiment using all subjects for testing and different values of σ (1, 2 and 4). We used the Linear SVM classifier with Fourier descriptors, which yielded the best results in previous experiments, and Baseline descriptors for comparison. This experiment was repeated 8 times for each parameter configuration, and average results are compared in Table 5.5.

σ	ts=100		ts=400	
	Control	Patients	Control	Patients
Baseline, σ = 1	79.53% ± 24.86%	68.95% ± 34.86%	87.52% ± 18.67%	73.17% ± 31.37%
Baseline, σ = 2	81.07% ± 21.73%	73.78% ± 30.38%	84.88% ± 17.79%	72.56% ± 32.60%
Baseline, σ = 4	78.71% ± 19.64%	71.56% ± 27.10%	82.92% ± 20.25%	73.82% ± 28.68%
Fourier, σ = 1	87.16% ± 20.28%	71.76% ± 37.63%	89.67% ± 17.42%	73.86% ± 35.05%
Fourier, σ = 2	87.94% ± 18.80%	74.18% ± 35.93%	89.80% ± 16.73%	74.27% ± 35.86%
Fourier, σ = 4	84.42% ± 17.69%	75.79% ± 33.31%	84.82% ± 16.86%	73.37% ± 31.96%

Table 5.5: *Average and standard deviation SVM precision values for different train sizes and noise amounts.*

Comparing the descriptors, the Fourier descriptor again reached higher levels of accuracy than the baseline descriptor in all cases. The training set size also influenced the results, such that in nearly all cases a larger training yielded a higher level of accuracy. The exceptions were the Fourier descriptor with σ = 4 and the Baseline descriptor with σ = 2, that reached an average accuracy higher with the training set of size 100. Nonetheless, the results were consistent and show that the pipeline is able to handle different levels of noise, and that the angle analysis can deal with small inconsistencies from the tracker for the classification task.

Table 5.6 presents a comparison with previous experiment accuracy rates. This result indicates that the leave-one-out experiment poses a more accurate representation of the classification experiment for unseen hands. The presence of augmented data lowers the accuracy for the patient set (around 73% for all values of σ), with significant values of standard deviation. This small performance loss was expected due to the variance of hand poses in the patient set, as discussed earlier.

Figure 5.12 show the average accuracy per subject of the dataset. It is observable from the Figure that lower accuracies are concentrated in specific cases, notably two control individuals ($C02\_fr$ and $C10\_fr$), and three patients ($P05\_fl$, $P06\_fr$, $P07\_fl$ and $P07\_fr$). For the remaining cases, the pipeline was able to predict the class with accuracy over 80% for σ = 1. With higher values of σ, it is noticeable that the accuracy decreases for some cases (*e.g.* $C01\_fr$ and $P01\_fr$), and increases in some of the described cases of lower accuracy (*e.g.* $P06\_fr$). This is explained by the tendency of data normalization and detail

Experiment	Control (%)	Patient (%)	General (%)
LOO	94.66% ± 8.45%	83.00% ± 31.69%	90.37% ± 21.14%
LOO + SS, σ = 1	88.41% ± 18.90%	72.80% ± 36.20%	82.66% ± 27.66%
LOO + SS, σ = 2	88.63% ± 18.36%	73.51% ± 35.96%	83.06% ± 27.25%
LOO + SS, σ = 4	87.29% ± 18.09%	73.86% ± 34.85%	82.35% ± 26.38%

Table 5.6: *Accuracy comparison (in %) between the Linear SVM with sample synthesis using different values of σ (in mm) with the result obtained in the Leave-one-person-out experiment.*

loss that comes with higher values of noise.

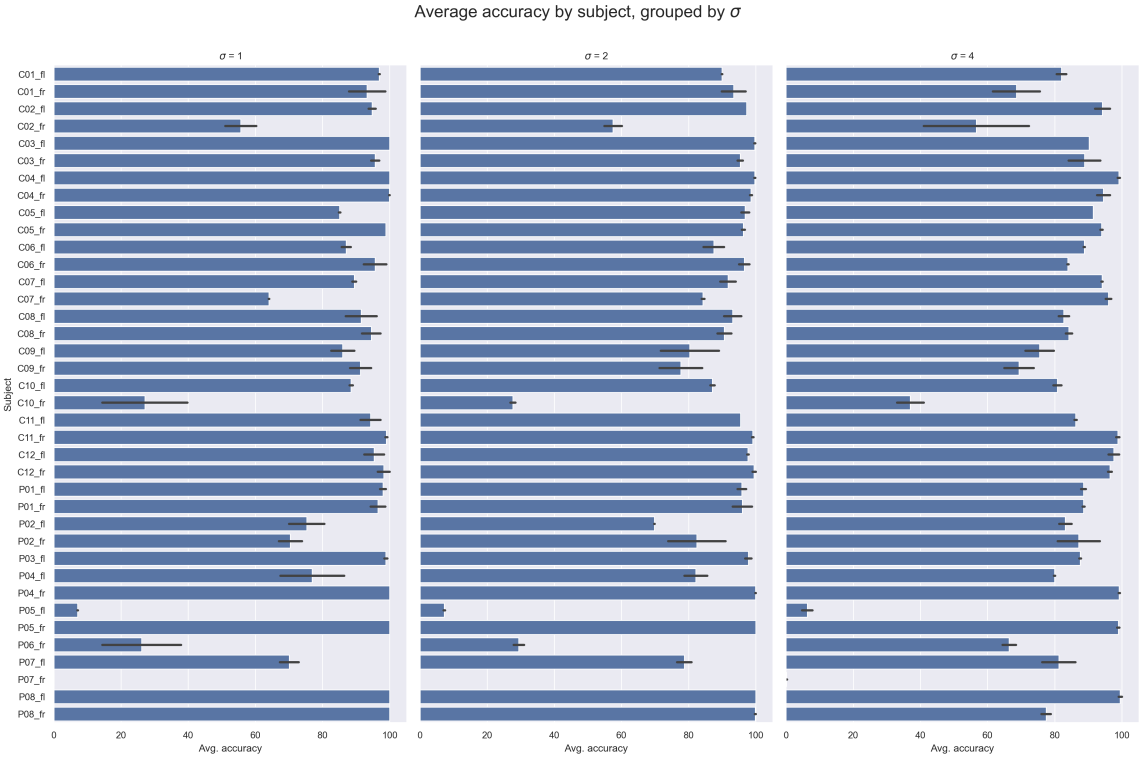


Figure 5.12: *Average accuracy by subject, grouped by σ.*

The main findings of these experiments are:

- Without noise, we are able to reach a good accuracy score for classification between control and patients, even with scenarios of unseen shapes.
- With the presence of noise, the accuracy score is lower especially in patients. The training set size has little influence on the accuracy, and the use of Fourier descriptors does enhance the results for left hand of the patients.
- Fourier descriptors are a classical approach for describing series, and are used in this context to prove the usefulness of describing movements taking into account its dynamic aspects - any state-of-art series descriptor can be used and should provide yet more accurate results.

5.3 Comparison with the goniometer

We concluded in previous experiments that the hand analysis pipeline is able to characterize flexion and abduction movements, the angles extracted translate into open and closed hand patterns, and that a straightforward classifier has high accuracy in distinguishing between patient and control hand shapes. In this experiment we compared patient measurements obtained by the sensor with reliable goniometer measurements, aiming to evaluate whether the hand analysis pipeline can provide objective feedback to the occupational therapists in practical scenarios.

For this, with support of the Occupational Therapy department from FMRP-USP, we obtained the range of motion goniometer measurements for flexion and abduction of five patients. Those data were acquired for patients in different stages of the disease. The measurements provided were the maximum and minimum value for each flexion angle, and the maximum distance between tips for abduction.

For RA, the standard procedure is the range of motion value evaluation. For phalanx joints, i.e. MCP, PIP and DIP, such values are computed simply by subtracting the maximum flexion angle with the maximum extension angle of each joint. Since flexion and extension are measured with relation to the same plane, we represent this value as the minimum and maximum value for the flexion angle of each joint. In some patients, this minimum value is negative, indicating that the resting position is negative with relation to the plane defined by the palm of the hand. This configuration is named hyperextension, and the proposed method is unable to identify negative values. In our proposal, we extracted clips with single hand opening movements, in which we can identify the range of each angle measurement using maximum and minimum values.

In order to compare the measurements, we decided to use an observation-based methodology. For each patient and hand, the sensor measurement should be computed from the set of all measurements manually annotated into clips (using the methodology described in Section 5.1). Figure 5.13 plots all measurements per angle from one of the evaluated patients, with violin plots (Hintze and Nelson, 1998) with estimated distributions of each angle. Small white points are measurements taken from each clip frame, and the background curves are calculated using a kernel density estimation method from the underlying points.

From all the measurements, we compute quantiles for each patient and angle, and compose the maximum and minimum values using four different approaches:

1. Global maximum and minimum values for each measurement;
2. 0.05 and 0.95 quantiles for each measurement;
3. 0.10 and 0.90 quantiles for each measurement;
4. Average of minimum and average of maximum values from all clips of the same subject (used to compose Tables 5.1 and 5.2).

Figure 5.14 show the interval comparison between the annotated goniometer ROM (in red) and the other strategies for the patient characterized in Figure 5.13. The observed behaviour varies with the type of joint: for MCP and DIP joints, the goniometer interval tends

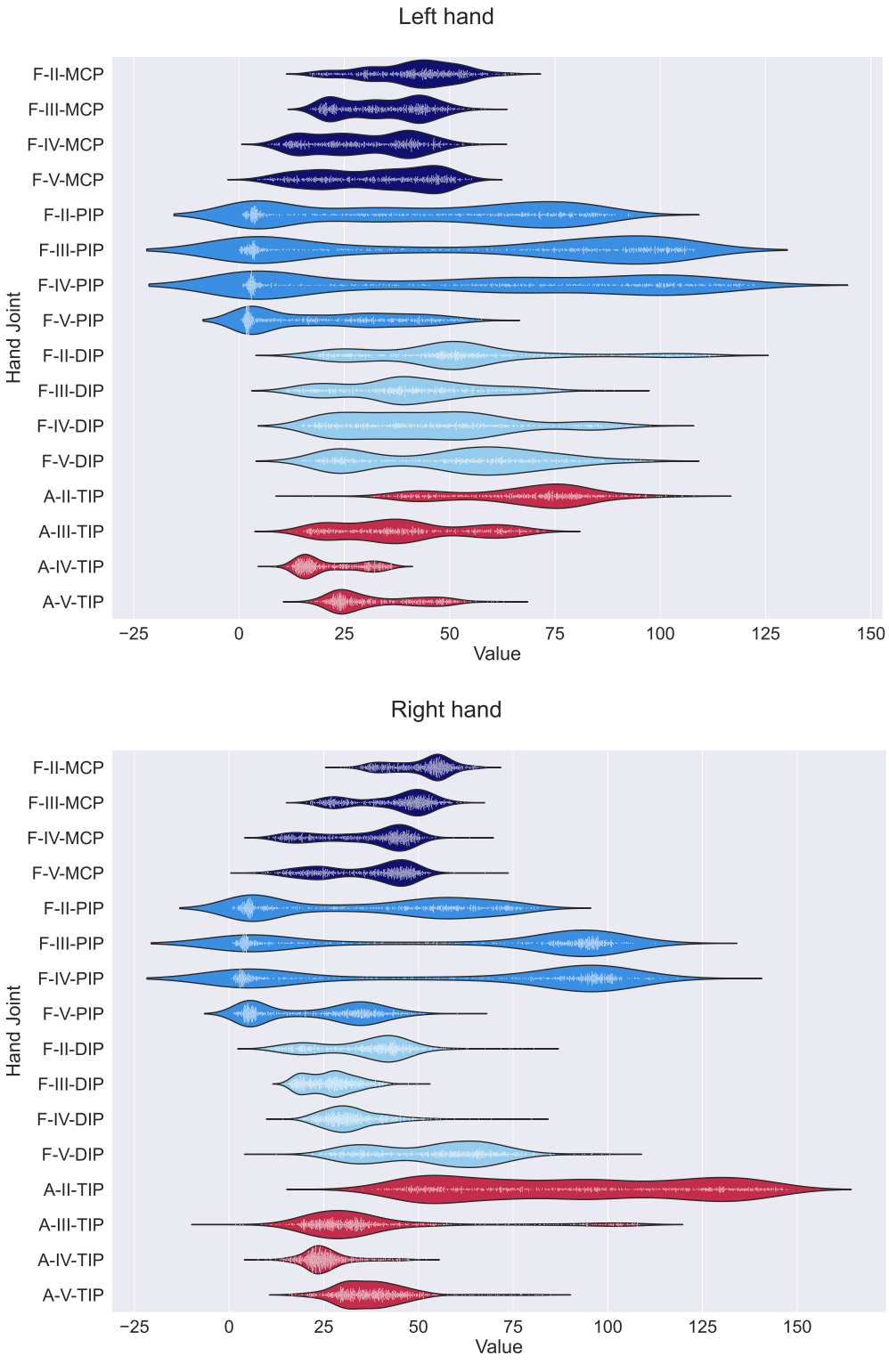


Figure 5.13: *Angle observations from a patient.*

to be larger than the sensor intervals; for PIP joints, the sensor minimum measurement is around 0 and the goniometer indicates a measurement around 20° for all joints, indicating that the hand does not open completely. For abduction, the measurement obtained by

the goniometer was the maximum distance between two fingers, and we considered that the movement interval for abduction is between 0 and the measurement provided by the goniometer.

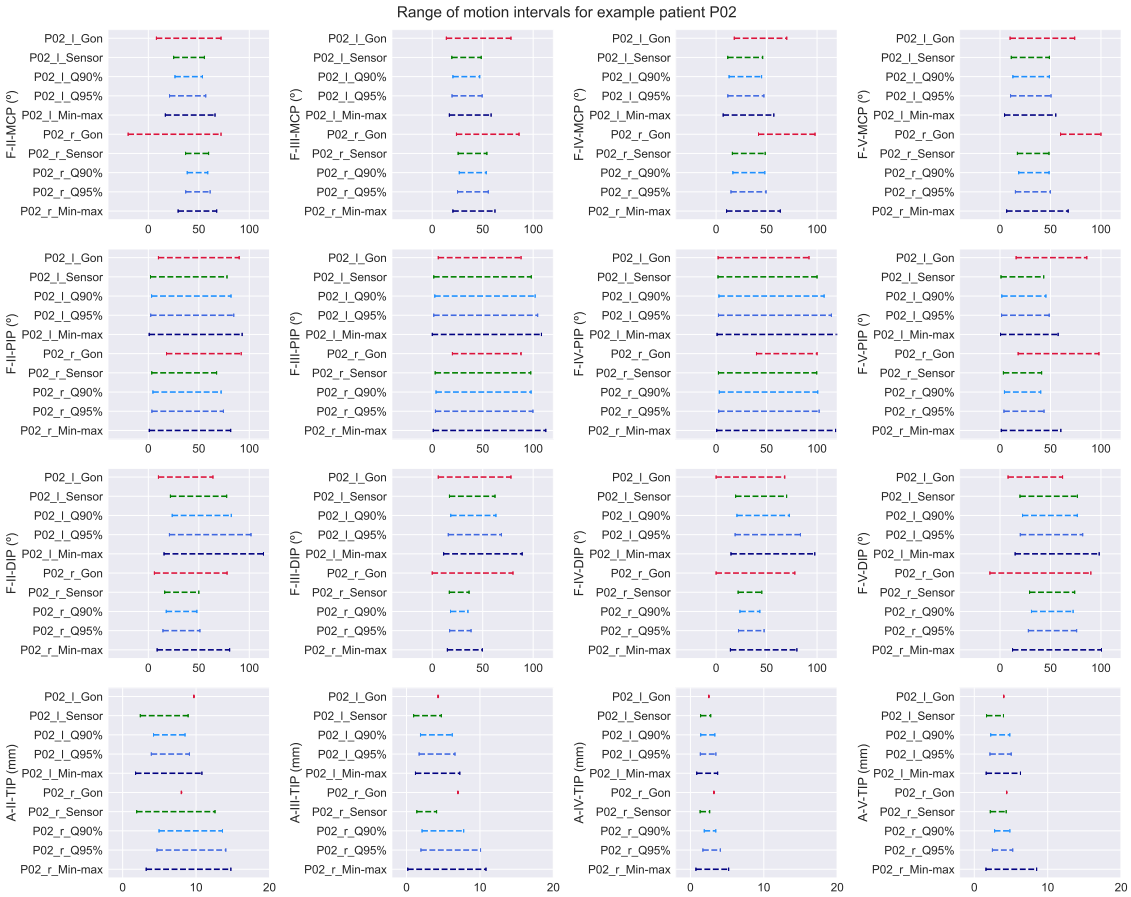


Figure 5.14: *Angle range intervals for observed patient using the four strategies decided.*

In order to choose the most consistent strategy for range of motion extraction we compute average values for range of motion per joint. The behaviour for the observed patient (Figures 5.13 and 5.14) arguably extends for the rest of the dataset: for MCP joints, the sensor ROM is substantially smaller than the goniometer, and the best approximation is the strategy 1 (using global maximum and minimum values). On the other hand, for PIP joints, the range of motion provided by the goniometer is consistently smaller than the sensor measurement; in this case, using the strategy 4 (average maximum and average minimum) leverages the error. For DIP joints, the best strategy was the 0.95 quantile (strategy 2), with the exception of the joint $F - II - DIP$, for which the strategy 4 provided a better aproximation. For abduction, the strategy 2 is a better approximation in all scenarios with the exception of the joint $A - II - TIP$, for which the strategy 4 yielded a smaller magnitude error. Figure 5.15 shows the average ROM value, while Figure 5.16 shows the average absolute magnitude error per joint.

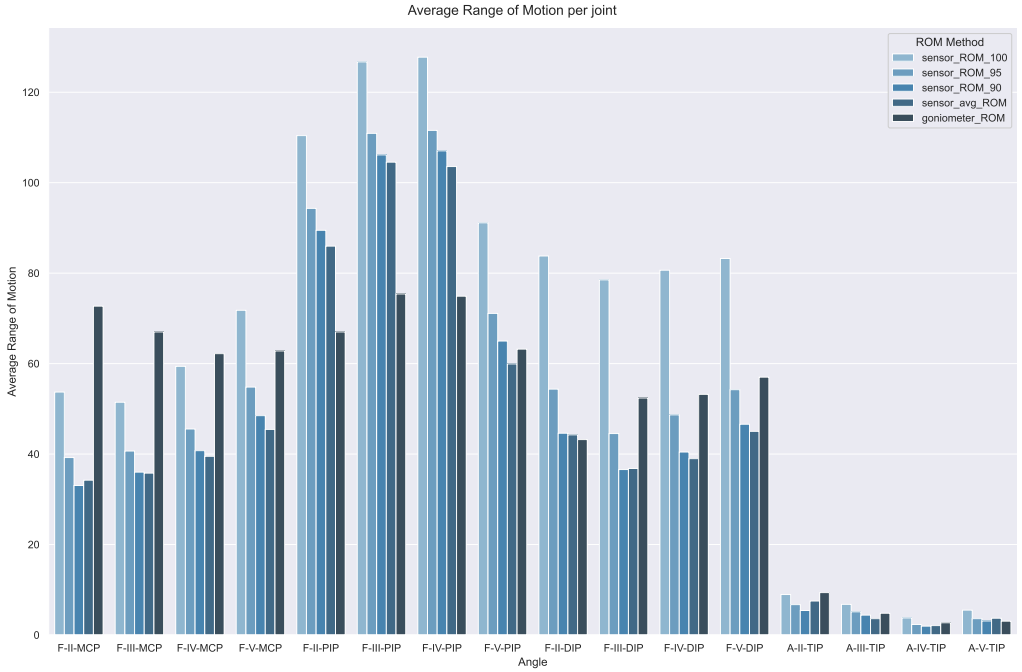


Figure 5.15: *Average range of motion per strategy, comparing with the goniometer.*

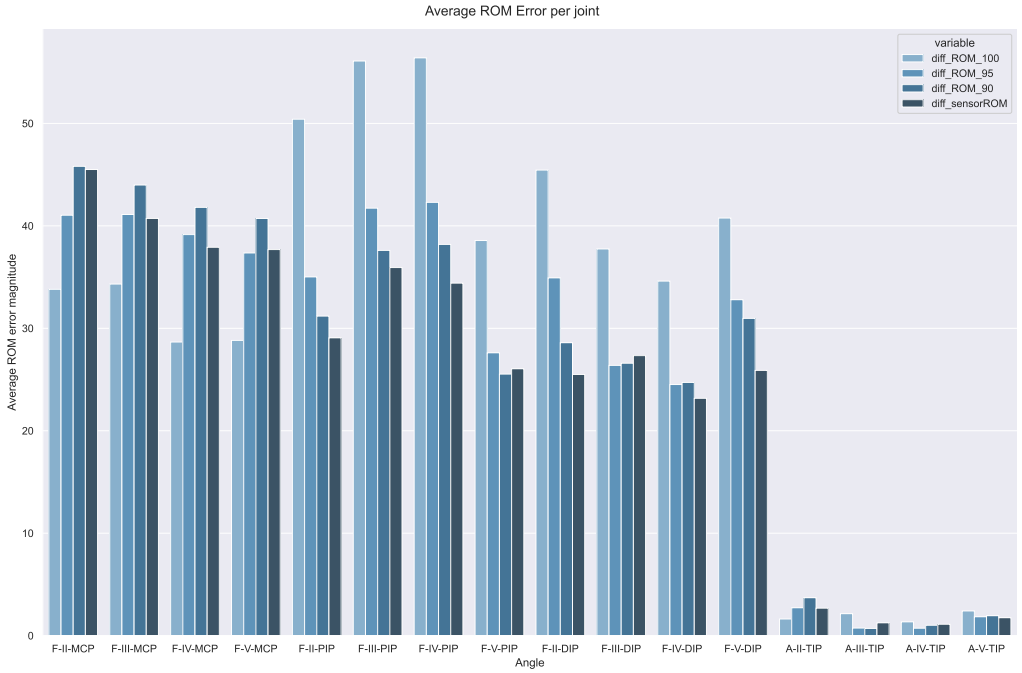


Figure 5.16: *Average absolute difference of range of motion per strategy, comparing with the goniometer.*

Analysing the error magnitude, we consider the approximation provided by the sensor not sufficient to provide reliable results for occupational therapy. In order to detail the error sources that compose such result, we computed the Pearson correlation coefficient between the observed variables. The dataset used in this analysis was composed by one

record per hand (5 patients, 2 hands) and per angle (16 angles), resulting in a total of 160 samples. For each sample, we computed maximum and minimum values for each strategy, maximum and minimum values extracted by the goniometer, as well as the ROM measurement using all strategies and the ground-truth goniometer ROM measurement. The Pearson correlation coefficient $\rho_{X,Y}$ is a measure of linear correlation between two sets of data X and Y, assuming values in the interval $[-1, 1]$. Values closer to 1 indicate high positive correlations, negative values indicate negative correlation and low magnitude values indicate that the variables are uncorrelated. This coefficient is computed by the covariance between two variables divided by the product of their standard deviations, as detailed in Equation 5.2.

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_x \sigma_y} \tag{5.2}$$

Figure 5.17 shows a heatmap with this measure, indicating that the correlation between the range of motion obtained by the goniometer and the sensor ROM peaks at 0.52, for the strategy 4. This indicates a low correlation between such measurements. Furthermore, the correlation between the observated ROMs for the four strategies are above 0.9, indicating that the all strategies have similar behaviours, which means that outlier filtering has limited effect in this task. In addition, the correlation between goniometer minimum values and sensor minimum values is much lower than the correlation between maximum values, thus the main imprecisions come from lower angle measurements. One possible explanation is the incapacity of the hand pose estimation method on representing extension angles (negative measurements).

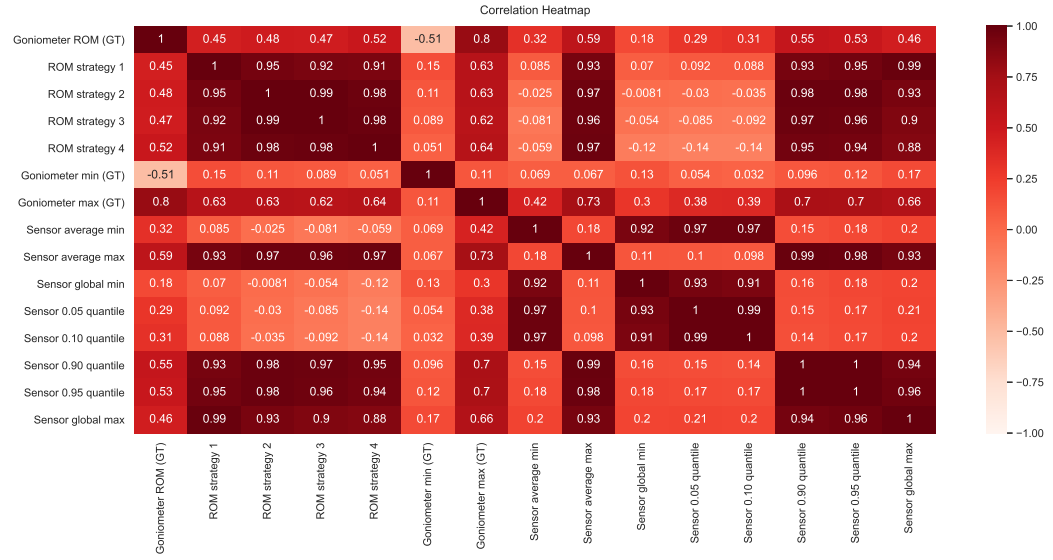


Figure 5.17: *Correlation heatmap between observations - values close to 1 indicate a high linear correlation, close to -1 indicate a high negative linear correlation, and low magnitude values indicate that the variables are uncorrelated.*

Aware of this issue with the hand pose estimation method, we recalculated all the measures transforming all negative minimum values to zero. Figure 5.18 shows the corre-

lation indexes in the new scenario. We observed an increase on correlation values between goniometer ROM and all strategies (peaking at 0.55 for the strategy 4), and between goniometer minimum values and sensor measurements. This indicates that the extension movement impacts the results of the ROM comparison, but the increase of the Pearson correlation coefficient is insufficient to justify the use of the method in practical scenarios.

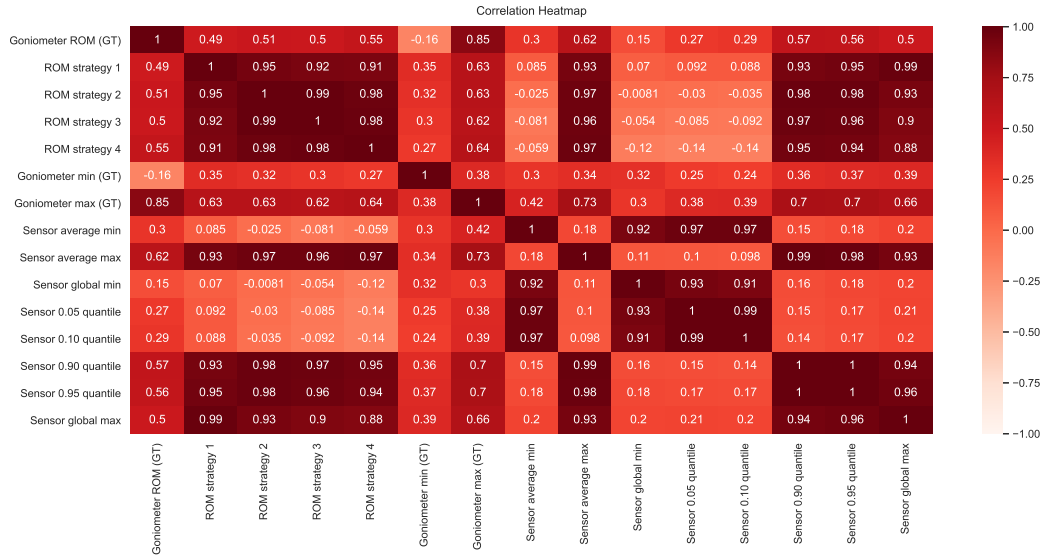


Figure 5.18: *Correlation heatmap between observations (disconsidering negative GT angle measurements) - values close to 1 indicate a high linear correlation, close to -1 indicate a high negative linear correlation, and low magnitude values indicate that the variables are uncorrelated.*

5.4 Remarks

In Section 5.1, we showed that the hand analysis pipeline yields coherent results for hand angles, for which the highest and lowest measurements are associated with open and closed hand patterns. We further described the general behaviour of control subjects and patients in flexion movements, as well as how the limitations caused by rheumatoid arthritis affects the angles. The pipeline uses a pre-trained model for hand pose estimation and was trained with regular hand shapes, therefore showing sufficient generalization capacity in terms of estimating hand poses in unseen scenarios. For occupational therapy, we believe that the characterization of flexion angles in terms of temporal signals provides new possibilities for comparison and characterization of the disease state of each patient.

The experiment executed in Section 5.2 shows that the use of simple descriptors and classifiers is enough to differentiate movement patterns from control and patient subjects. This initial result reflects that computer vision approaches can be used to identify features that characterize rheumatoid arthritis on patients. The high accuracy yielded from the leave-one-person-out experiment also indicates that the movement patterns are indeed separable as two distinct classes, and further exploration of the characteristics of such patterns can provide new findings about rheumatoid arthritis.

In Section 5.3, we compare ground-truth minimum and maximum values obtained from a goniometer with sensor measurements, in order to validate the practical use of the framework for range of motion estimation. Watching the obtention of measurements with the goniometer, we noted that the therapist guides the patient to put the adequate strength in the movements of opening and closing the hand. The measurements are then calculated by guiding the patient to keep the hand in the current state for some seconds - the guidelines provided by the therapist have fundamental importance for the precise assessment. For the sensor acquisition, those guidelines are performed before the recording, and this setup difference yields bias and influences the process.

The results obtained in this experiment show that the range of motion intervals generated by the sensor and the goniometer have a low correlation, despite the efforts on evaluating different strategies of providing min/max values. This shows that more studies and enhancements on hand pose estimation are needed in order to use the framework in practical range of motion acquisition scenarios.

We also noted that the lack of annotated data from patients sensibly limits the ROM measurement accuracy, since we are applying our system on a dataset where hand shapes and patterns are very different from those used in the pre-trained model. With bigger and specific purpose datasets, the increase of the generalization capacity of hand pose estimation methods can help this pipeline to achieve more reliable results.

Chapter 6

Conclusion

6.1 Conclusion

Main contributions

This thesis sought to evaluate the viability of an automatic pipeline for patients with rheumatoid arthritis, using state-of-the-art hand pose estimation methods. A new approach has been introduced and evaluated experimentally using real data. We defined an acquisition setup with a patient set and a control set, obtaining flexion and abduction movement sequences. In the case of patients, the ulnar deviation and the use of orthosis affect the acquisition and the resulting hand pose, which makes the problem more challenging than state-of-the-art hand pose estimation benchmarks. We defined an acquisition protocol for flexion and abduction movements and detailed the main project decisions for the formation of the dataset.

We estimated the hand pose using the Pose-REN algorithm and presented a method to convert 3D point coordinates to flexion and abduction angles. We proposed a strategy to register sequences of movements and represent cycles of movements as feature vectors based on frequency domain descriptors. This representation of the movement patterns is then used for classification, aiming to identify patterns and distinguish between patients and control data.

The proposed method is able to accurately estimate skeleton angles and range of motion measurements from control and patients, even with the 3D hand pose estimation algorithm being trained in a completely different dataset of healthy hand movements. Results for classification are promising, showing that a simple movement cycle is enough to distinguish patients from control.

Findings of the experiments

The main findings of the experiments are:

- The angles extracted by the hand analysis pipeline encode correctly flexion and abduction movements, characterizing visually each movement in terms of angle

variation.

- A simple classifier and motion descriptor is able to distinguish between control and patient classes, even with unseen subjects.
- When compared to real goniometer range of motion measurements, the error magnitude is still high, indicating that the there is a lot of room for improvement in the application for real patient assessments.

Impact for computer vision community

It is important to note that this thesis proposes a challenging application for hand pose estimation with a baseline solution. The pipeline built for estimation of hand angles can be used with different hand pose estimation methods and sensor configurations. The work of Ng *et al.* (2021) uses the flexion and abduction angles formulae to apply the self-attention hand pose estimation method in a setup with two sensors, computing the average angle value in both sensors in order to reach more robust results. We believe that with the advances in the results for hand pose estimation, further methods could be used with tests "in the wild", and the generalization capacity of current pose estimation methods makes possible the application in other knowledge areas, especially in assessments for medicine and occupational therapy.

Impact for occupational therapy

For occupational therapy, the thesis proposes a framework to analyse flexion and abduction angles as time signals. Compared to current movement analysis that uses maximum and minimum values, the analysis of a signal that encodes the complete movement pattern can help the research on rheumatoid arthritis and future characterization of movement patterns from patients in different stages of the disease.

In terms of objective feedback, the range of motion comparison experiment resulted in high error values, and the min/max intervals provided by the goniometer have low correlation to the intervals generated by the sensor. This limits the use of our framework to provide objective feedback for the patients. However, new methods and new datasets can enhance this result.

In terms of applicability, the acquisition protocol is simple and requires a single depth sensor RealSense SR300. The project originally sought to use 2D hand pose estimation methods in order to provide feedback with cell phone cameras, but despite the recent interest in such solutions, the state-of-art methods still have limited accuracy. With the progress of the area, we believe that such solutions can become feasible, making the setup much cheaper and accessible.

Impact of the dataset

The created dataset is important in the sense of providing depth images of patients with rheumatoid arthritis in contrast with control images. Such images have unusual shapes and pose a challenge for hand pose estimation methods. The dataset has the limitation of

not providing the ground-truth values for each frame, due to acquisition setup limitations. Nonetheless, experiments show that the dataset is able to provide valuable information in form of movement description for occupational therapists.

Lessons learnt from acquisition setup

In the context of machine learning for human/hand pose estimation, the ideal data acquisition procedure uses sensors to obtain ground-truth measurements for the skeletons of the subjects, allowing the training of hand pose estimation models that contains hands with shapes of patients with rheumatoid arthritis and thus leveraging the error propagation from the hand joints estimation. This process, however, involves a careful and time-consuming setup and positioning the sensors on the exact positions of the joints. The data acquisition protocol for this project required the assessment of hand shapes during medical assessments from patients in rehabilitation, and the conception of the setup demanded simplicity and comfort of the patients, with markerless assessments. In this context, we opted for a pre-trained model using the HANDS17 dataset, which provides robust estimatives and is able to deal with most movement sequences. Since the model was not trained with patients and hand shapes with disabilities, that decision limited the reliability of data and added an error source to the pipeline from the model used in hand pose estimation. The production of a dataset with ground-truth joint values for rheumatoid arthritis and other disabilities would improve the model used in HPE and provide more exact results, but was unfeasible in the current project.

Limitations and reproductibility of the pipeline

The pipeline is built such that new methods can be tested in the hand pose estimation step. The only limitation is the angle calculation, which might need to be reformulated for other hand models. The pipeline is compatible with every method that uses the HANDS17 or MSRA dataset, whose model is composed by 21 joints. We believe that if hand angle evaluation becomes a common task for hand pose estimation, the pipeline can be straightforwardly adapted to include other methods.

We are optimistic that the recent developments in 2D hand pose estimation, with million-scale datasets and training of proper models, can provide a better generalization capacity for 2D hand pose estimation methods. As discussed in Section 2.4, monocular image-based and pose estimation is a much more complex problem if compared to depth-based, since it relies on color-based joint estimation and naked hands do not have enough texture to allow reliable tracking of its parts. The amount of hand shape variation and self-occlusion are also major challenges. With the use of the InterHand2.6M dataset, we can expect that new models are trained and can reach promising results. For now, the use of 3D hand pose estimation methods is recommended.

6.2 Future Works

This work demonstrates the viability of using a computer-vision based system for movement analysis in the context of occupational therapy. However, the adoption of such

approach in a real-world clinical setting is more challenging and requires further research and development in many aspects.

We suggest multiple directions for future works following this thesis.

Hand pose estimation

For hand pose estimation, the proposed framework and acquisition protocol serves as a baseline for angle estimative of patients. We highlight the subsequent work of Ng *et al.* (2021), that uses a multiple sensor setup for hand angle estimation on patients with stroke. That work proposes the hand angle estimation as application of the method, uses multiple sensors for robustness, and points towards the future directions of producing a dataset with hand shapes of stroke patient. We believe that the construction of a purpose-specific ground-truth dataset with patients with rheumatoid arthritis will enhance the generalization capacity of hand pose estimation methods for this task.

Hand analysis

The main limitation of the current pipeline is that the movement clips are manually annotated. In order to provide a fully automatic pipeline, an algorithm for clip detection has to be applied. We currently see two strategies for this task:

1. Use the annotated intervals to train a machine learning model that automatically encodes the class information (i.e. flexion or abduction movements).

2. Use the main features described by the angle signals (inflexions and slopes) to detect similarities between a basis-curve and intervals of raw data.

For real-time flexion movement detection, we suggest a pattern-based algorithm. The first step is to remove apparent outliers and irregularities in MCP flexion angles. For this, we apply the smoothing by simplified least squares method in each angle separately. After the smoothing, we seek to find peaks and valleys in the signals. This step is made by finding zero-crossing values in the derivative of the signal. In order to avoid outliers, we also filtered the results, eliminating points that are at a distance bigger than a threshold to the global minimum and maximum. Clip interval candidates are composed by the interval between two local minima. Since we are dealing with multiple time signals, the resulting clips are the intersection of the intervals composed from each individual signal.

Another limitation of the current hand angle estimation method is that it only estimates positive angles, being unable to express extension and adduction values. The computation of angle formulae for other hand models, such as the MANO model, are also important.

In the context of occupational therapy, further work can be done in describing and quantifying the movements of flexion/extension and abduction/adduction: we described the shape of flexion and abduction measurements per frame, but a study on occupational therapy can associate the characteristics of rheumatoid arthritis with the observed behavior. Further analysis can also associate curve descriptors with the state of the disease in each

patient; this can be done by machine learning classifiers, taking into account the movement dynamic aspects for evaluating the hand state.

We can also perform analyses with the measurements obtained in sequences from patients with orthoses - such sequences can be used as comparison baseline for characterizing the state of the disease since the orthoses are designed to correct the ulnar deviation.

The proposed framework can be further used for evaluation of the hand state in the same patient. In the context of this work, we planned to perform acquisitions with a patient in different moments, but since the setup was in development we were not able to analyze the measurement variation through time. With enough data, it is possible to use such information to help the therapist in the treatment planning.

Another possibility of application is to ideally perform the classification of hand shapes and movement descriptors according to the disease state - creating a "grading" scheme that can be used to provide continuous evaluation of a single patient.

It is also feasible the development of an interface that can guide the patient on the flexion/abduction movement execution using augmented reality, through the playback of reference sequences indicating for the patient the movement that should be performed in each moment.

Other areas

We consider that the range of applications extend to areas other than occupational therapy - the characterization of different types of movements than flexion and abduction can be tested for tasks like human-computer interaction and sign language recognition, given the appropriate datasets for training and evaluation.

Appendix A

Data acquisition protocol

This appendix details the steps used in the final data acquisition protocol, used for the acquisition performed in patients at 2019 September 6th and 13rd (as described in Table 3.6).

A.1 Setup installation and configuration

1. Fasten the camera in a position such that the hand is acquired in a frontal position.
2. Mark the floor indicating the positioning of the patient.

A.2 Upon patient arrival

1. Extract goniometer measurements from the patient by the occupational therapist.
2. Instruct the patient showing the movement that should be done from the current position.
3. Position the patient on the mark.
4. If the patient is wearing a long sleeve shirt, roll up the sleeves such that it does not disturb the quality of the hand detection.
5. Record the sequences, in the following order:
   (a) Three flexion sequences with the right hand (without orthosis).
   (b) Three flexion sequences with the left hand (without orthosis).
   (c) Three abduction sequences with the right hand (without orthosis).
   (d) Three abduction sequences with the left hand (without orthosis).
   (e) Three flexion sequences with the right hand (with orthosis).
   (f) Three flexion sequences with the left hand (with orthosis).

(g) Three abduction sequences with the right hand (with orthosis).

(h) Three abduction sequences with the left hand (with orthosis).

A.3 Analysis

1. Annotate each movement sequence, indicating the frame numbers of the beginning and the end of the movement (indicating the intervals where the movement happens).
2. Generate tables with minimum, maximum and average values for each angle measured in each clip.
3. Compare the values obtained with the sensor and the values obtained by the goniometer.
4. Train classification model to classify between control and patient.

A.4 Recommendations

- The hand pose estimation method requires that the hand should be the nearest object to the camera in order to reach more precise results.
- Each recorded sequence should be composed by 10 movements (named clips). At the end of the 10 movement sequences, the patient should lower the arm, wait 10 seconds and start the next sequence. The acquisition software should be configured for this pipeline.

Appendix B

Results for each subject

In this appendix, we show the flexion angle and abduction assessments for each subject composing our dataset. The tables are composed in the same way as Tables 5.1 and 5.2, with minimum and maximum values separated per hand, joint and finger. For control subjects and patients with abduction measurements equal to zero, no abduction clips have been recorded.

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C01 - L	MTC (°)	16.25	68.20	13.15	70.30	8.92	71.91	11.11	77.61
	IFP (°)	1.43	78.40	0.92	102.14	3.44	112.94	1.72	80.92
	IFD (°)	11.35	59.24	9.13	50.35	15.35	47.80	16.74	50.87
	abd (cm)	0.00		0.00		0.00		0.00	
C01 - R	MTC (°)	22.97	67.14	16.12	70.59	11.93	71.70	6.87	78.05
	IFP (°)	2.63	84.37	3.92	105.76	6.27	112.39	3.90	78.11
	IFD (°)	13.85	54.04	11.59	44.72	17.38	46.57	20.80	49.31
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.1: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C02 - L	MTC (°)	16.44	32.70	12.29	31.48	7.25	30.71	10.54	43.12
	IFP (°)	1.16	41.23	1.31	75.36	0.72	105.54	0.50	45.75
	IFD (°)	6.22	121.85	3.81	84.12	3.97	67.15	3.38	98.93
	abd (cm)	0.00		0.00		0.00		0.00	
C02 - R	MTC (°)	13.19	25.53	11.05	23.83	5.13	23.53	7.20	41.83
	IFP (°)	1.05	34.01	1.92	84.39	0.89	97.34	1.56	46.77
	IFD (°)	7.85	119.32	8.25	93.45	11.40	90.70	9.80	92.78
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.2: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C03 - L	MTC (°)	15.17	48.49	12.56	53.63	10.23	55.05	15.34	66.59
	IFP (°)	2.03	79.49	1.10	110.31	1.39	121.12	0.97	94.71
	IFD (°)	3.40	73.84	3.62	51.76	3.37	51.17	3.60	54.09
	abd (cm)	0.00		0.00		0.00		0.00	
C03 - R	MTC (°)	10.46	48.81	8.42	53.85	5.78	55.53	9.11	66.80
	IFP (°)	1.71	79.59	1.09	103.67	1.30	119.98	0.82	81.57
	IFD (°)	8.18	66.74	4.00	54.95	9.11	48.97	8.90	54.41
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.3: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C04 - L	MTC (°)	16.30	67.23	11.33	68.53	6.11	66.49	8.46	75.82
	IFP (°)	2.32	75.53	1.92	105.02	0.89	113.94	1.78	82.52
	IFD (°)	6.97	69.40	4.35	55.51	10.03	50.32	5.95	61.75
	abd (cm)	0.00		0.00		0.00		0.00	
C04 - R	MTC (°)	18.52	64.55	13.96	64.72	6.71	64.71	8.59	72.26
	IFP (°)	2.49	74.77	1.32	101.36	1.94	111.37	2.32	84.14
	IFD (°)	5.88	73.17	5.70	61.32	12.58	53.23	13.05	66.10
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.4: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C05 - L	MTC (°)	18.97	68.36	15.77	67.47	10.01	68.18	12.53	72.16
	IFP (°)	2.51	90.97	1.63	109.98	1.23	117.82	2.54	88.56
	IFD (°)	3.40	60.52	3.41	48.88	6.98	54.49	3.79	53.28
	abd (cm)	0.00		0.00		0.00		0.00	
C05 - R	MTC (°)	18.45	69.91	14.89	76.30	7.94	79.83	7.88	82.08
	IFP (°)	2.42	94.76	1.89	101.86	1.32	109.96	1.62	87.97
	IFD (°)	2.71	64.67	2.85	57.60	7.75	60.19	4.35	57.12
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.5: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C06 - L	MTC (°)	21.32	58.38	15.94	72.14	6.94	71.19	7.95	81.94
	IFP (°)	1.22	88.06	1.35	102.07	1.60	103.68	1.23	79.09
	IFD (°)	11.69	70.33	10.41	53.61	11.17	57.56	11.06	53.36
	abd (cm)	2.39		2.08		1.17		1.05	
C06 - R	MTC (°)	20.03	60.20	13.74	61.93	5.56	58.78	5.14	67.81
	IFP (°)	0.63	96.59	0.87	113.28	0.81	115.05	1.61	101.12
	IFD (°)	14.33	73.89	12.63	60.69	17.13	65.44	8.77	63.90
	abd (cm)	5.01		2.89		0.55		1.03	

Table B.6: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C07 - L	MTC (°)	17.98	67.76	13.33	75.26	5.77	75.79	6.49	78.00
	IFP (°)	1.56	58.47	0.87	89.15	1.17	100.52	0.74	67.20
	IFD (°)	15.14	53.00	11.94	46.00	13.84	47.70	12.17	54.16
	abd (cm)	2.89		1.33		0.53		1.04	
C07 - R	MTC (°)	18.84	56.42	11.14	66.67	1.83	67.15	2.57	73.11
	IFP (°)	-1.51	61.91	-0.48	89.89	0.16	102.87	-0.14	59.28
	IFD (°)	24.68	52.48	17.92	45.87	23.08	47.81	21.50	48.98
	abd (cm)	2.59		1.52		0.67		0.74	

Table B.7: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C08 - L	MTC (°)	9.13	35.55	6.98	41.32	2.55	40.31	8.00	58.31
	IFP (°)	0.82	68.85	1.06	100.06	1.19	114.78	0.83	58.19
	IFD (°)	8.09	77.37	10.24	63.78	12.64	67.17	12.23	71.00
	abd (cm)	1.97		1.38		0.70		1.43	
C08 - R	MTC (°)	18.77	60.14	9.91	64.37	4.17	68.36	6.91	74.22
	IFP (°)	1.06	94.38	0.98	106.62	1.56	119.28	1.36	91.80
	IFD (°)	16.72	58.63	14.40	51.76	19.60	54.93	19.67	59.78
	abd (cm)	5.43		2.35		0.42		0.71	

Table B.8: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C09 - L	MTC (°)	16.26	68.68	12.40	70.45	7.28	71.94	18.57	75.13
	IFP (°)	2.66	88.36	1.33	99.41	1.82	102.88	2.88	88.48
	IFD (°)	7.56	47.80	7.25	43.77	9.34	43.94	11.94	38.08
	abd (cm)	3.47		1.85		0.63		2.32	
C09 - R	MTC (°)	15.70	67.49	10.08	67.12	5.37	70.32	17.05	75.47
	IFP (°)	1.44	88.11	2.12	98.39	1.27	101.04	1.79	84.26
	IFD (°)	16.25	45.91	11.59	42.70	12.76	44.78	13.88	42.19
	abd (cm)	3.47		2.05		0.65		2.06	

Table B.9: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C10 - L	MTC (°)	22.08	30.46	13.80	22.82	5.69	17.92	7.06	30.73
	IFP (°)	1.14	47.95	1.45	68.18	1.81	75.61	0.74	35.99
	IFD (°)	3.95	122.75	10.45	109.28	10.60	111.24	8.76	111.84
	abd (cm)	2.86		1.34		0.61		1.78	
C10 - R	MTC (°)	23.50	38.52	13.48	37.74	9.08	35.63	10.63	57.65
	IFP (°)	1.80	93.37	2.27	99.47	1.78	109.35	1.70	66.25
	IFD (°)	9.73	50.24	10.68	49.10	10.53	40.72	9.38	54.81
	abd (cm)	3.48		1.74		0.37		0.66	

Table B.10: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C11 - L	MTC (°)	18.41	61.09	15.78	67.58	8.85	69.99	12.00	78.78
	IFP (°)	3.71	78.82	4.85	107.02	6.34	107.51	2.82	75.44
	IFD (°)	14.85	53.95	12.67	41.34	12.75	42.15	12.53	48.08
	abd (cm)	2.50		1.51		0.68		1.75	
C11 - R	MTC (°)	15.94	74.70	14.14	74.66	5.02	77.36	8.99	81.52
	IFP (°)	0.06	92.23	-0.96	107.28	-0.94	113.54	-0.94	86.53
	IFD (°)	13.08	58.15	9.95	44.30	13.71	49.25	6.58	50.60
	abd (cm)	3.24		1.38		0.38		1.65	

Table B.11: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
C12 - L	MTC (°)	13.99	73.52	18.81	70.93	14.13	70.12	17.41	73.88
	IFP (°)	2.24	85.15	0.78	100.36	3.11	105.04	1.57	84.36
	IFD (°)	6.12	59.10	4.35	50.74	7.87	51.50	8.67	51.81
	abd (cm)	3.08		1.73		1.49		2.42	
C12 - R	MTC (°)	16.16	67.02	11.83	68.67	5.63	68.00	8.75	75.21
	IFP (°)	1.80	89.71	1.13	100.31	2.45	108.92	1.48	84.96
	IFD (°)	14.29	62.38	13.37	58.79	18.35	57.57	15.89	57.47
	abd (cm)	2.84		1.40		0.78		1.98	

Table B.12: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P01 - L	MTC (°)	20.67	31.16	22.52	27.85	18.78	24.71	32.13	43.79
	IFP (°)	1.61	80.04	3.26	96.43	7.30	89.80	16.70	50.14
	IFD (°)	11.55	62.65	13.83	49.56	14.45	65.79	24.78	80.67
	abd (cm)	0.00		0.00		0.00		0.00	
P01 - R	MTC (°)	19.35	41.35	12.66	39.80	6.27	37.52	9.84	53.81
	IFP (°)	2.93	100.37	3.22	125.84	4.85	127.99	2.00	96.48
	IFD (°)	13.69	37.38	13.54	25.30	15.33	30.19	16.50	42.65
	abd (cm)	10.11		6.03		3.38		3.75	

Table B.13: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P02 - L	MTC (°)	25.44	50.11	19.91	42.53	11.18	40.20	12.64	45.67
	IFP (°)	2.76	69.37	2.60	95.22	2.78	96.82	1.92	38.41
	IFD (°)	22.99	70.17	17.17	53.48	19.44	60.06	20.69	73.49
	abd (cm)	7.99		6.02		3.26		4.80	
P02 - R	MTC (°)	40.30	56.63	27.23	52.01	18.01	45.90	20.15	46.03
	IFP (°)	5.59	63.82	4.27	95.66	3.17	98.40	4.84	37.91
	IFD (°)	17.12	46.29	18.16	34.12	23.97	42.83	32.57	68.19
	abd (cm)	10.94		4.64		2.99		4.59	

Table B.14: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P03 - L	MTC (°)	15.26	60.18	16.77	59.60	11.32	62.14	13.45	72.76
	IFP (°)	1.70	101.61	1.99	109.48	1.54	113.75	1.43	81.91
	IFD (°)	6.25	45.75	2.89	33.89	4.26	38.21	5.59	52.12
	abd (cm)	9.40		6.19		3.22		4.86	
P03 - R	MTC (°)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IFP (°)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IFD (°)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	abd (cm)	7.46		6.59		3.23		6.01	

Table B.15: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P04 - L	MTC (°)	21.10	59.48	19.43	60.82	12.17	56.26	15.19	68.15
	IFP (°)	1.96	104.76	1.74	119.36	0.99	115.07	1.76	85.19
	IFD (°)	9.14	48.10	5.08	45.37	6.75	49.57	14.60	57.62
	abd (cm)	10.42		5.53		3.10		4.95	
P04 - R	MTC (°)	15.70	37.10	16.90	48.34	19.21	51.63	45.11	67.82
	IFP (°)	3.99	83.39	4.60	97.03	7.15	91.36	18.60	48.46
	IFD (°)	18.08	52.03	14.42	37.68	15.20	44.55	26.29	58.60
	abd (cm)	0.00		0.00		0.00		0.00	

Table B.16: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P05 - L	MTC (°)	22.25	45.94	12.23	42.97	1.49	40.60	17.83	56.88
	IFP (°)	2.60	56.82	6.12	73.60	3.67	87.32	10.54	62.65
	IFD (°)	19.08	86.16	14.57	76.55	12.98	60.94	30.02	69.02
	abd (cm)	7.74		4.24		2.61		5.71	
P05 - R	MTC (°)	20.91	53.43	19.75	59.20	22.43	60.32	30.81	75.12
	IFP (°)	3.02	78.46	2.41	101.72	5.15	94.64	5.22	36.98
	IFD (°)	22.62	28.87	8.49	16.57	12.71	19.94	18.79	27.73
	abd (cm)	6.61		4.56		3.73		3.92	

Table B.17: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P06 - L	MTC (°)	28.53	59.41	16.49	59.34	8.24	56.15	19.82	62.98
	IFP (°)	2.58	79.65	2.41	102.83	2.91	99.07	22.82	69.84
	IFD (°)	18.20	50.72	9.55	38.80	9.86	48.64	14.13	46.73
	abd (cm)	2.82		1.84		1.53		1.99	
P06 - R	MTC (°)	25.25	53.03	15.76	56.70	8.52	53.53	19.16	63.45
	IFP (°)	10.88	89.35	4.72	106.41	3.23	100.70	31.12	65.34
	IFD (°)	17.28	40.87	10.79	31.14	10.68	38.34	16.56	45.96
	abd (cm)	2.90		1.66		1.15		1.95	

Table B.18: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P07 - L	MTC (°)	15.74	44.05	10.56	47.77	4.91	47.51	6.76	58.84
	IFP (°)	0.74	85.21	1.23	98.16	1.37	91.04	1.08	71.82
	IFD (°)	19.45	47.15	17.70	43.76	16.36	48.84	13.13	54.07
	abd (cm)	2.52		1.66		1.69		2.50	
P07 - R	MTC (°)	23.63	76.20	14.88	80.43	9.05	81.23	7.89	80.91
	IFP (°)	0.94	88.63	2.76	103.09	1.49	99.76	1.74	83.98
	IFD (°)	20.45	48.65	6.66	45.11	8.06	54.33	18.23	56.34
	abd (cm)	2.87		2.48		1.34		0.98	

Table B.19: *Measurements extracted from one of the patients during the data acquisition session.*

	Finger	2		3		4		5	
		min	max	min	max	min	max	min	max
P08 - L	MTC (°)	24.30	55.02	20.10	55.79	11.56	54.77	18.82	66.69
	IFP (°)	0.78	108.90	2.82	120.07	2.40	112.72	2.93	84.52
	IFD (°)	15.19	46.96	15.51	41.30	20.39	48.60	28.10	53.80
	abd (cm)	3.45		2.14		1.20		1.63	
P08 - R	MTC (°)	31.63	52.00	31.26	52.08	32.66	51.59	44.77	62.60
	IFP (°)	3.08	45.54	9.13	85.69	10.86	78.31	11.11	65.44
	IFD (°)	26.85	50.84	18.24	37.45	24.64	39.26	32.07	41.91
	abd (cm)	3.06		2.41		2.04		3.57	

Table B.20: *Measurements extracted from one of the patients during the data acquisition session.*

Appendix C

Visual evaluation for abduction sequences

In this appendix, we show full results for characterization of the abduction movement, with the complete graphics of movement and angle description, extending the analysis performed for the finger IV to all other fingers.

Figures C.1 and C.2 illustrate flexion sequences, highlighting the frames with higher and lower values for the measurement $F - IV - PIP$.

For clip extraction, we manually extracted the landmarks for abduction, with the cycle correspondent to one opening and closing of the fingers. Figures C.3 and C.4 show the annotated clips.

Figures C.5 and C.6 conclude the analysis by showing summarization in terms of mean and standard deviation for one patient and control individual.

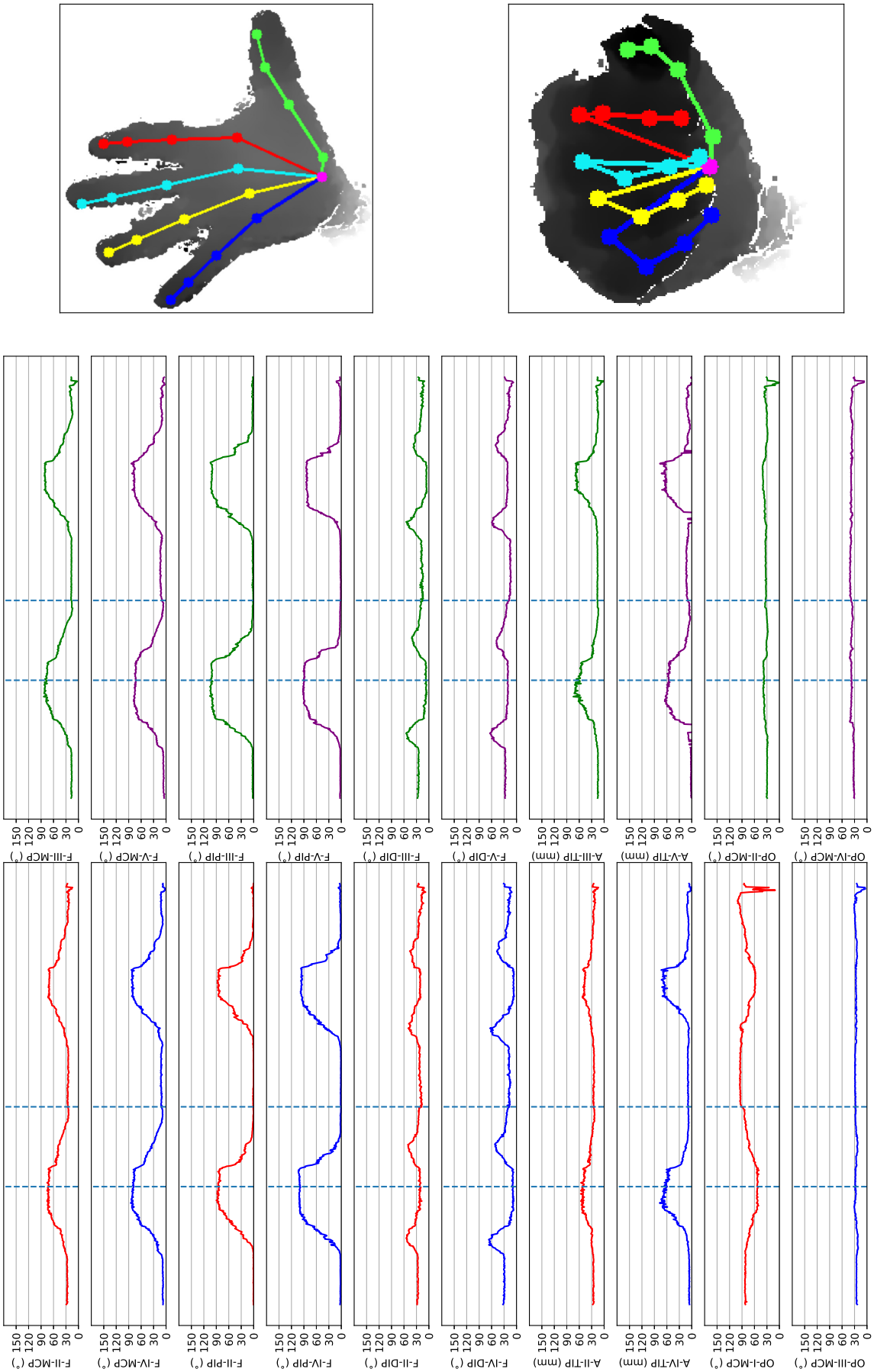


Figure C.1: *Angle evaluation of a patient. Left and middle columns show graphs with angle joint measurements obtained frame by frame of a sequence acquired following the defined protocol. Right column present frames of maximum and minimum values for the angle F – IV – MCP in the sequence, corresponding to the instants highlighted by vertical dashed lines in the graphs: top image is the lowest angle value and bottom corresponds to the highest.*

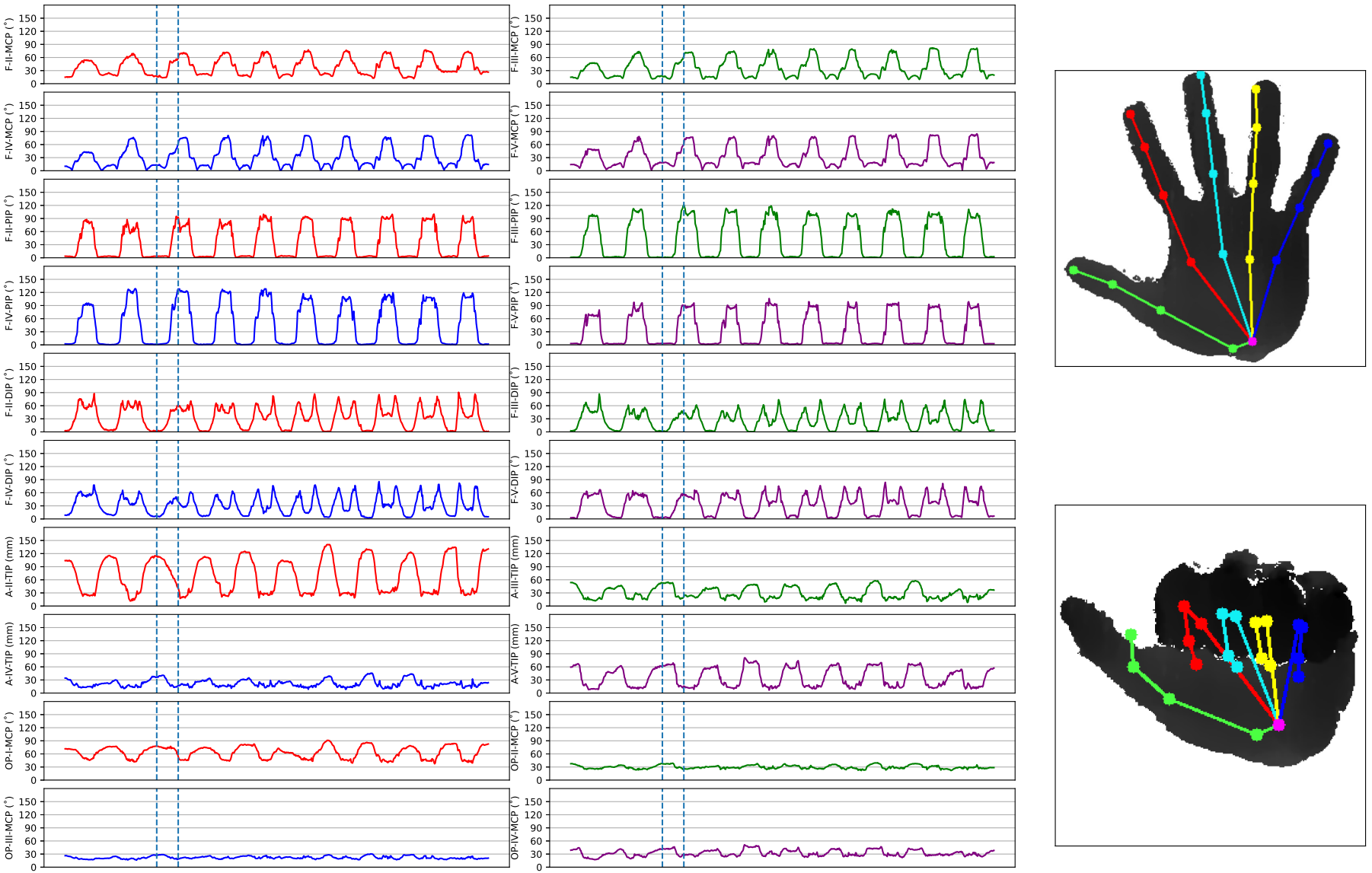


Figure C.2: *Angle evaluation of an individual in control group. Left and middle columns show graphs with angle joint measurements obtained frame by frame of a sequence acquired following the defined protocol. Right column present frames of maximum and minimum values for the angle $F - IV - MCP$ in the sequence, corresponding to the instants highlighted by vertical dashed lines in the graphs: top image is the lowest angle value and bottom corresponds to the highest.*

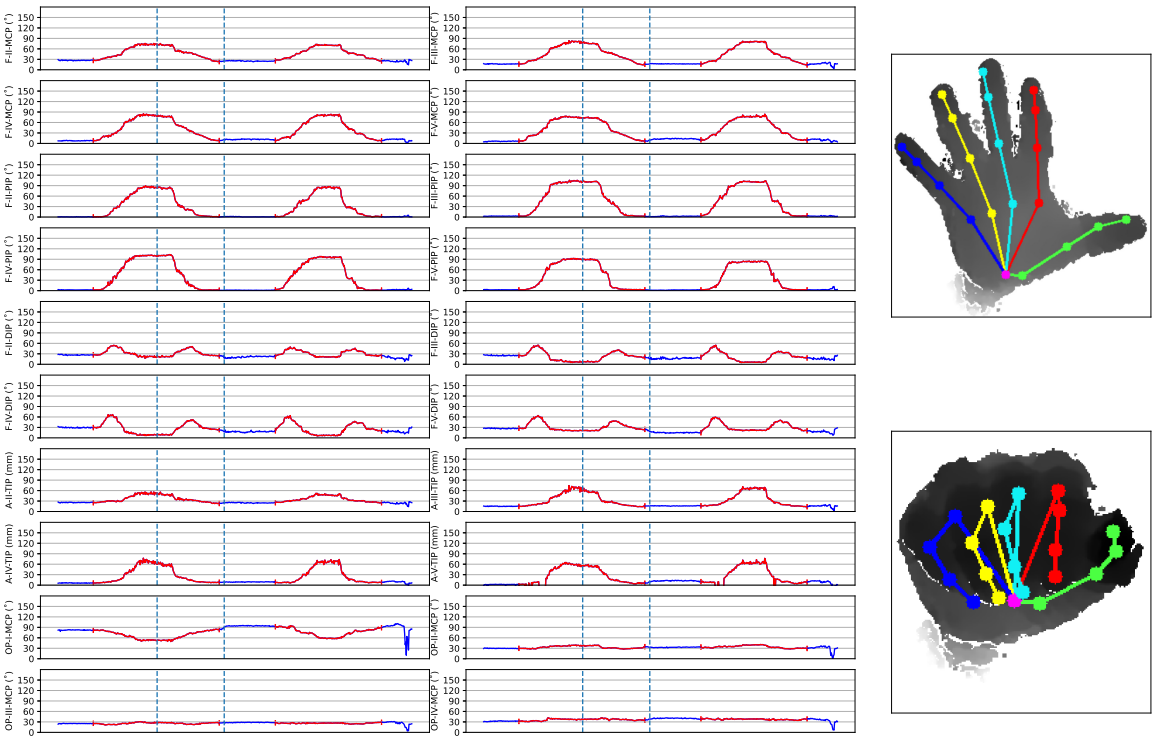


Figure C.3: *Manual annotation of movement intervals in the angle sequence described in Figure C.1. Extracted clips are marked in red.*

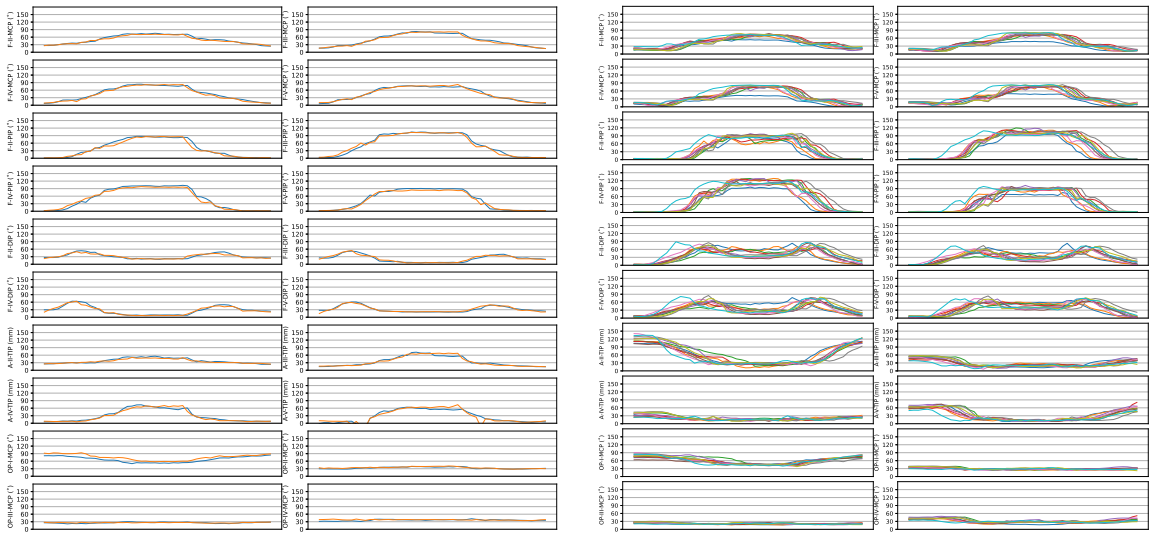


Figure C.4: *Extracted clips from sequences shown in Figure 5.3: patient (left) and control (right). Trajectories have been re-sampled.*

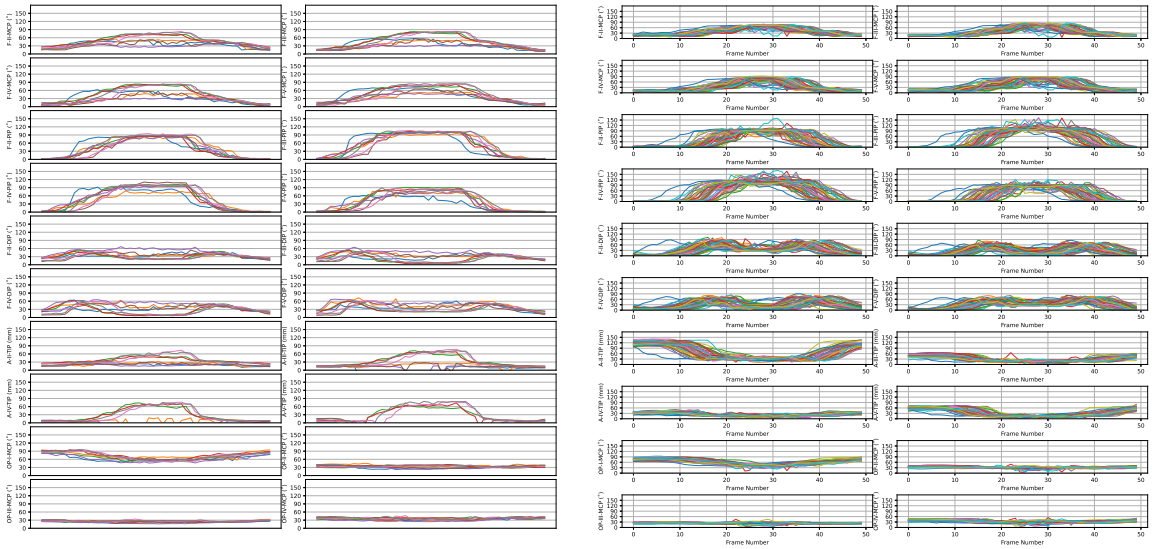


Figure C.5: *All trajectories extracted from clips of the same person: patient (left) and control (right). Trajectories have been re-sampled.*

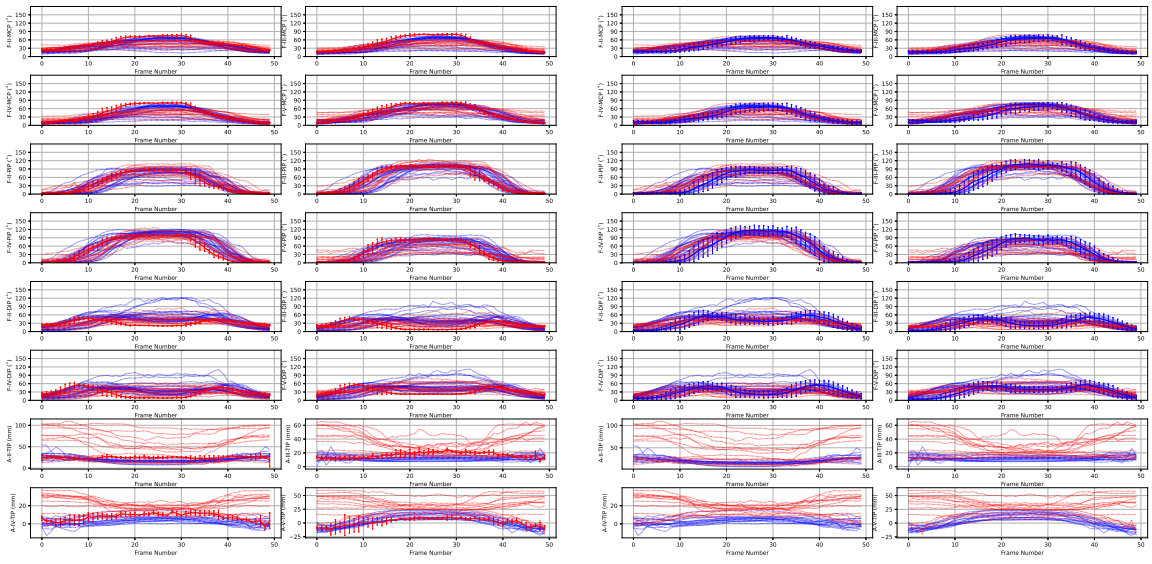


Figure C.6: *Summarization in terms of mean and standard deviation of all trajectories extracted from clips from the same person: patient (left) and control (right). "Average clips" of other subjects are shown in the background. Patient samples are colored in red, and control samples are colored in blue.*

Appendix D

Visual evaluation for abduction sequences

In this appendix, we show the results and characterization of the abduction movement. This is made in a similar fashion to the methodology applied in Section 5.1.

Figures D.1 and D.2 illustrate abduction sequences, highlighting the frames with higher and lower values for the measurement $A - IV - tip$. With this, we show the direct correspondence between closed hands and lower measurements and open hands with higher measurements.

For clip extraction, we manually extracted the landmarks for abduction, with the cycle correspondent to one opening and closing of the fingers. Figures D.3 and D.4 show the annotated clips.

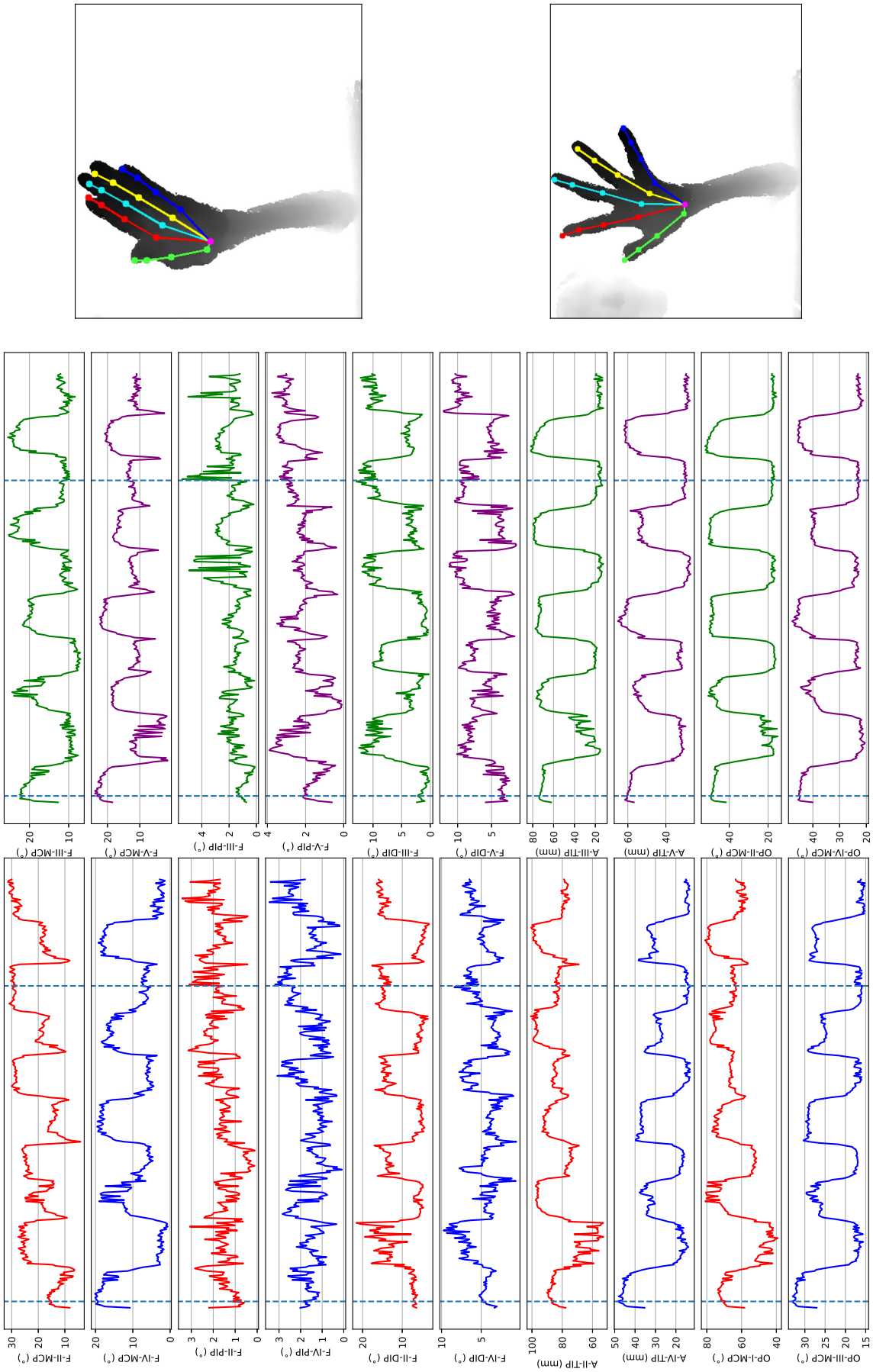


Figure D.1: *Angle evaluation of a patient. The vertical lines point the maximum and minimum values for the angle A – IV – tip, and the hand images on the right are the frames corresponding to the highlighted values; top image is the lowest angle value and bottom corresponds to the highest.*

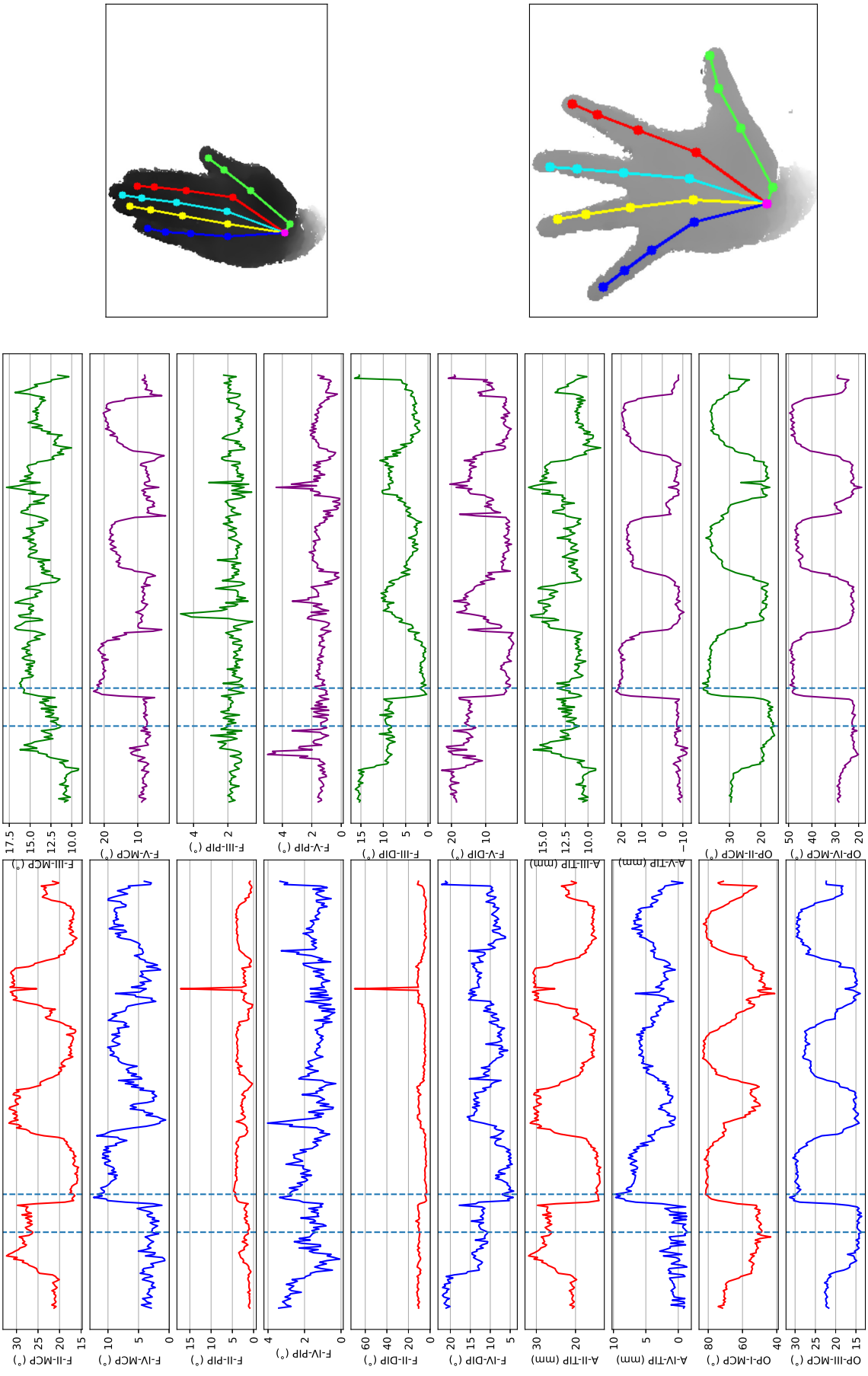


Figure D.2: *Angle evaluation of an individual in control group. The vertical lines point the maximum and minimum values for the angle A – IV – tip, and the hand images on the right are the frames corresponding to the highlighted values; top image is the lowest angle value and bottom corresponds to the highest.*

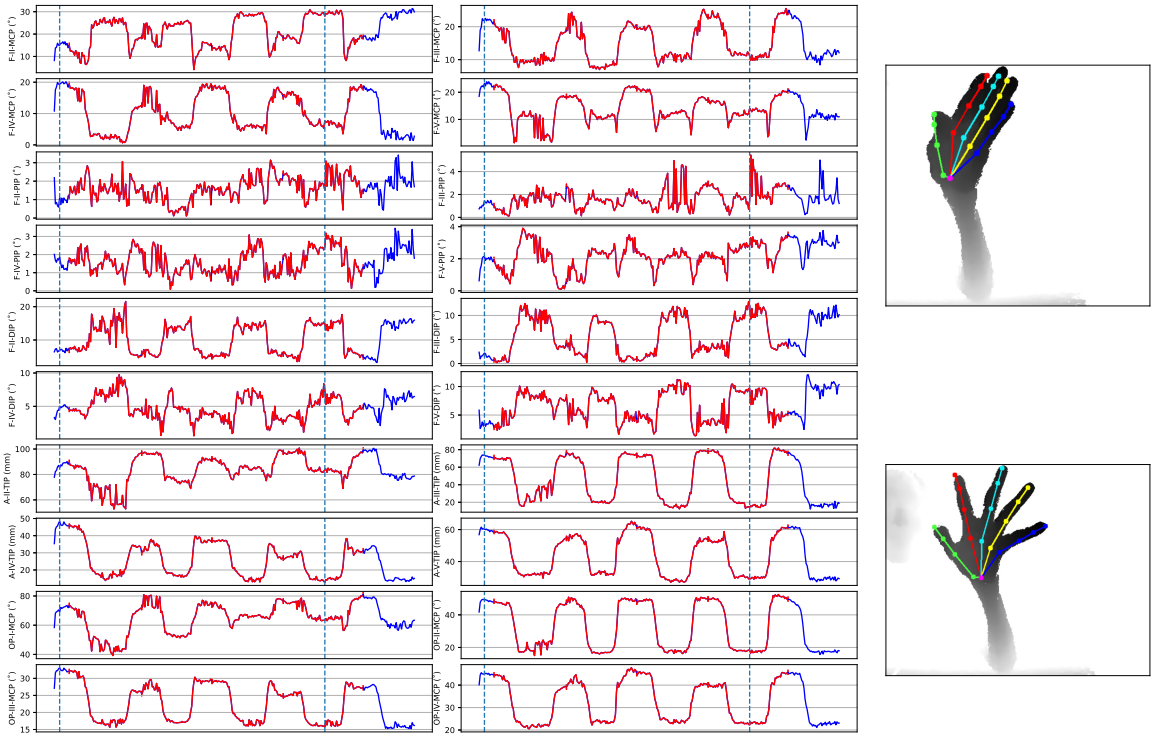


Figure D.3: *Manual annotation of movement intervals in the angle sequence described in Figure D.1. Extracted clips are marked in red.*

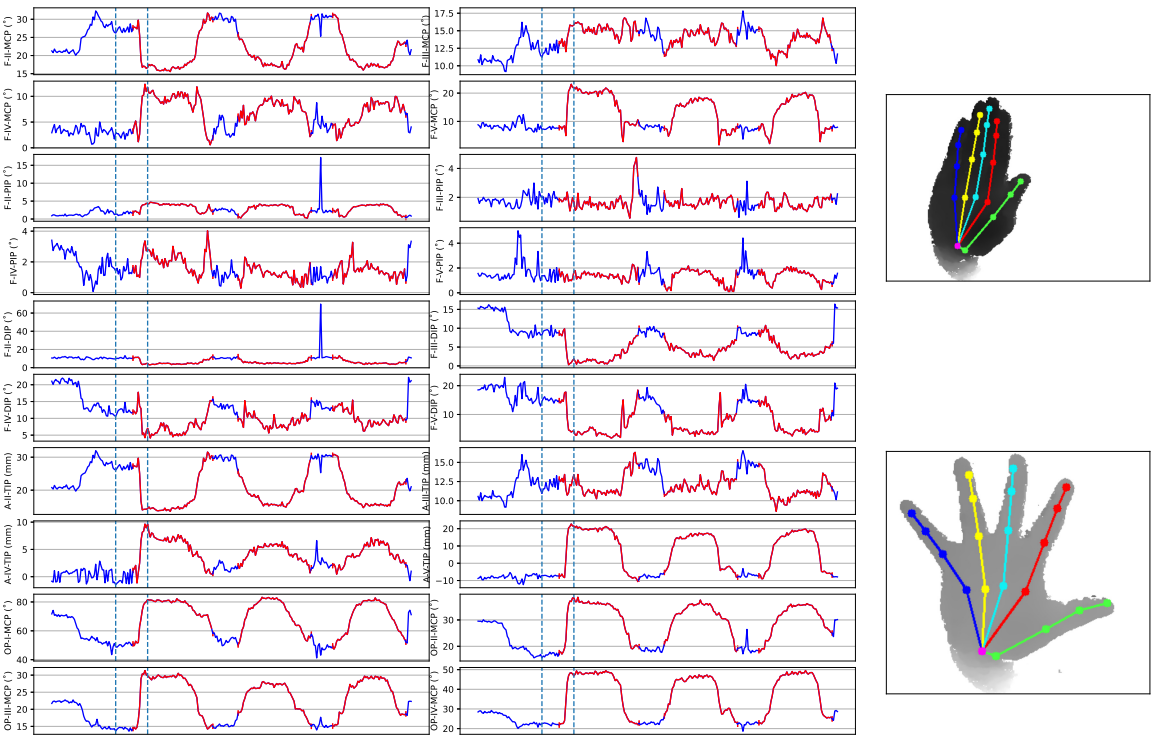


Figure D.4: *Manual annotation of movement intervals in the angle sequence described in Figure D.2. Extracted clips are marked in red.*

References

[Boukhayma *et al.* 2019] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. "3D hand shape and pose from images in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10843–10852 (cit. on p. 18).

[Bruton *et al.* 1999] Anne Bruton, Bridget Ellis, and Jonathan Goddard. "Comparison of visual estimation and goniometry for assessment of metacarpophalangeal joint angle". In: *Physiotherapy* 85.4 (1999), pp. 201–208 (cit. on p. 4).

[Burnaev *et al.* 2015] E. Burnaev, P. Erofeev, and A. Papanov. "Influence of resampling on accuracy of imbalanced classification". In: *Eighth International Conference on Machine Vision (ICMV 2015)*. Ed. by Antanas Verikas, Petia Radeva, and Dmitry Nikolaev. Vol. 9875. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Dec. 2015, p. 987521. doi: 10.1117/12.2228523. arXiv: 1707.03905 [stat.ML] (cit. on p. 64).

[Breiman *et al.* 1984] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984 (cit. on p. 47).

[Breiman 2001] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 47).

[Campos 2006] T E de Campos. "3D Visual Tracking of Articulated Objects and Hands". PhD thesis. Department of Engineering Science, University of Oxford, Trinity Term, 2006 (cit. on pp. 11, 12).

[Cao *et al.* 2021] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 43.1 (2021). doi: 10.1109/TPAMI.2019.2929257 (cit. on p. 19).

[Cejnog, de Campos, Elui, and R. M. Cesar Jr. 2019] L. W. X. Cejnog, T. E. de Campos, V. M. C. Elui, and R. M. Cesar Jr. "Hand range of motion evaluation for Rheumatoid Arthritis patients". In: *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. 2019, pp. 1–5 (cit. on p. 8).

[Cejnog, de Campos, Elui, and Roberto Marcondes Cesar Jr. 2021] L. W. X. Cejnog, T. E. de Campos, V. M. C. Elui, and Roberto Marcondes Cesar Jr. "A framework for automatic hand range of motion evaluation of rheumatoid arthritis patients". In: *Informatics in Medicine Unlocked* 23 (2021), p. 100544. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2021.100544. URL: https://www.sciencedirect.com/science/article/pii/S2352914821000344 (cit. on p. 9).

[Chan *et al.* 1982] Tony F Chan, Gene H Golub, and Randall J LeVeque. "Updating formulae and a pairwise algorithm for computing sample variances". In: *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Springer. 1982, pp. 30–41 (cit. on p. 48).

[Chen *et al.* 2019] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang. "Pose guided structured region ensemble network for cascaded hand pose estimation". In: *Neurocomputing* (2019) (cit. on pp. 4, 8, 15, 20, 27, 30, 31, 33, 37, 38).

[Campos and Murray 2006] T E de Campos and D W Murray. "Regression-based Hand Pose Estimation from Multiple Cameras". In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. New York, June 2006. URL: http://www.robots.ox.ac.uk/ActiveVision/Publications/decampos_murray_cvpr2006/decampos_murray_cvpr2006.html (cit. on p. 11).

[Carvalho *et al.* 2012] Rosana Martins Ferreira de Carvalho, Nilton Mazzer, Claudio Henrique Barbieri, *et al.* "Análise da confiabilidade e reprodutibilidade da goniometria em relação à fotogrametria na mão". In: *Acta Ortopédica Brasileira* 20.3 (2012), pp. 139–149 (cit. on p. 4).

[Donate *et al.* 2021] Paula B. Donate *et al.* "Cigarette smoke induces miR-132 in Th17 cells that enhance osteoclastogenesis in inflammatory arthritis". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 118.1 (2021), e2017120118. DOI: 10.1073/pnas.2017120118 (cit. on p. 1).

[Dougherty *et al.* 2002] Edward R Dougherty *et al.* "Inference from clustering with application to gene-expression microarrays". In: *Journal of Computational Biology* 9.1 (2002), pp. 105–126 (cit. on p. 64).

[Ellis *et al.* 1997] Bridget Ellis, Anne Bruton, and Jonathan R Goddard. "Joint angle measurement: a comparative study of the reliability of goniometry and wire tracing for the hand". In: *Clinical rehabilitation* 11.4 (1997), pp. 314–320 (cit. on p. 4).

[Erol *et al.* 2007] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. "Vision-based hand pose estimation: A review". In: *Computer Vision and Image Understanding* 108.1 (2007), pp. 52–73 (cit. on p. 11).

[Fang *et al.* 2020] Linpu Fang, Xingyan Liu, Li Liu, Hang Xu, and Wenxiong Kang. "JGR-P2O: Joint Graph Reasoning based Pixel-to-Offset Prediction Network for

3D Hand Pose Estimation from a Single Depth Image". In: *European Conference on Computer Vision (ECCV)*. Proprint published at arXiv:2007.04646. Source code available from https://github.com/fanglinpu/JGR-P2O. 2020 (cit. on pp. 16, 17, 38).

[Garcia-Hernando *et al.* 2018] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 409–419 (cit. on p. 16).

[Ge *et al.* 2017] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. "3D convolutional neural networks for efficient and robust hand pose estimation from single depth images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1991–2000 (cit. on p. 15).

[Goia *et al.* 2017] Daniela Nakandakari Goia, Carlos Alberto Fortulan, Benedito Moraes Purquerio, and Valéria Meirelles Carril Elui. "A new concept of orthosis for correcting fingers ulnar deviation". en. In: *Research on Biomedical Engineering* 33 (Mar. 2017), pp. 50–57. issn: 2446-4740. doi: 10.1590/2446-4740.02516. url: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2446-47402017000100050&nrm=iso (cit. on p. 2).

[Goldberger *et al.* 2004] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. "Neighbourhood components analysis". In: *Advances in neural information processing systems* 17 (2004), pp. 513–520 (cit. on p. 47).

[Guo *et al.* 2017] Hengkai Guo *et al.* "Region ensemble network: Improving convolutional network for hand pose estimation". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 4512–4516 (cit. on pp. 15, 23, 26–28).

[Gutiérrez-Martínez *et al.* 2014] J. Gutiérrez-Martínez, A. Ortiz-Espinosa, P. R. Hernandez-Rodriguez, and M. A. Núñez-Gaona. "System to measure the range of motion of the joints of the human hand ." In: *Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion* 66 Suppl 1 (2014), S122–30 (cit. on p. 4).

[Hampali *et al.* 2019] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. "Ho-3D: A multi-user, multi-object dataset for joint 3D hand-object pose estimation". In: *arXiv preprint arXiv:1907.01481* (2019) (cit. on p. 18).

[Hammond 2019] Patrick Douglas Hammond. "Deep Synthetic Noise Generation for RGB-D Data Augmentation". PhD thesis. Brigham Young University, 2019 (cit. on p. 65).

[Hastie *et al.* 2009] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. "Multi-class adaboost". In: *Statistics and its Interface* 2.3 (2009), pp. 349–360 (cit. on p. 48).

[He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity Mappings in Deep Residual Networks". In: *CoRR* abs/1603.05027 (2016). arXiv: 1603.05027. URL: http://arxiv.org/abs/1603.05027 (cit. on p. 18).

[Hinton 1990] Geoffrey E Hinton. "Connectionist learning procedures". In: *Machine learning*. Elsevier, 1990, pp. 555–610 (cit. on p. 47).

[Hintze and Nelson 1998] Jerry L Hintze and Ray D Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184 (cit. on p. 68).

[Kessler *et al.* 1995] G. Drew Kessler, Larry F. Hodges, and Neff Walker. "Evaluation of the CyberGlove As a Whole-hand Input Device". In: *ACM Trans. Comput.-Hum. Interact.* 2.4 (Dec. 1995), pp. 263–283. ISSN: 1073-0516. DOI: 10.1145/212430.212431. URL: http://doi-acm-org.ez67.periodicos.capes.gov.br/10.1145/212430.212431 (cit. on p. 11).

[LeCun *et al.* 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444 (cit. on p. 14).

[Lima *et al.* 2016] LL Lima, JSR Melo, TS Fragoso, TM Vieira, and MC Oliveira. "Fisiomotion: SISTEMA DE AVALIAÇÃO DE PACIENTES PORTADORES DE ARTRITE REUMATOIDE USANDO SENSOR DE MOVIMENTOS". In: *XXV Congresso Brasileiro de Engenharia Biomédica - CBEB 2016*. 2016 (cit. on p. 5).

[Ledoit and Wolf 2004] Olivier Ledoit and Michael Wolf. "Honey, I shrunk the sample covariance matrix". In: *The Journal of Portfolio Management* 30.4 (2004), pp. 110–119 (cit. on p. 48).

[Marques 1997] Amélia Pasqual Marques. *Manual de goniometria*. Editora Manole, 1997 (cit. on pp. 2, 3).

[Moon, Chang, *et al.* 2018] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 15, 16).

[Meals *et al.* 2018] Clifton G. Meals, Rebecca J. Saunders, Sameer Desale, and Jr Kenneth R. Means. "Viability of Hand and Wrist Photogoniometry". In: *HAND* 13.3 (2018). PMID: 28391753, pp. 301–304. DOI: 10.1177/1558944717702471. eprint: https://doi.org/10.1177/1558944717702471. URL: https://doi.org/10.1177/1558944717702471 (cit. on pp. 1, 4, 20).

[Molchanov *et al.* 2015] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. "Hand gesture recognition with 3D convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. 2015, pp. 1–7 (cit. on p. 65).

[Moon, Yu, *et al.* 2020] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. "InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image". In: *European Conference on Computer Vision (ECCV)*. 2020 (cit. on pp. 18, 19).

[Mota *et al.* 2013] Licia Maria Henrique da Mota *et al.* "Guidelines for the diagnosis of rheumatoid arthritis". In: *Revista Brasileira de Reumatologia (English Edition)* 53.2 (2013), pp. 141–157. issn: 2255-5021. doi: https://doi.org/10.1016/S2255-5021(13)70019-1. url: http://www.sciencedirect.com/science/article/pii/S2255502113700191 (cit. on p. 2).

[Mueller *et al.* 2018] Franziska Mueller *et al.* "Ganerated hands for real-time 3D hand tracking from monocular RGB". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 49–59 (cit. on pp. 18, 19).

[Ng *et al.* 2021] Mei-Ying Ng, Chin-Boon Chng, Wai-Kin Koh, Chee-Kong Chui, and Matthew Chin-Heng Chua. "An enhanced self-attention and A2J approach for 3D hand pose estimation". In: *Multimedia Tools and Applications* (2021), pp. 1–16 (cit. on pp. 76, 78).

[Norkin and White 1997] Cynthia C Norkin and D Joyce White. *Medida do movimento articular: manual de goniometria*. Artes médicas, 1997 (cit. on p. 2).

[Newell *et al.* 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European Conference on Computer Vision*. Springer. 2016, pp. 483–499 (cit. on p. 15).

[Oikonomidis *et al.* 2011] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. "Efficient model-based 3D tracking of hand articulations using Kinect." In: *British Machine Vision Conference (BMVC)*. Vol. 1. 2. 2011, p. 3 (cit. on p. 13).

[Oikonomidis *et al.* 2012] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. "Tracking the articulated motion of two strongly interacting hands". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 1862–1869 (cit. on p. 13).

[Oberweger and Lepetit 2017] Markus Oberweger and Vincent Lepetit. "Deepprior++: Improving fast and accurate 3D hand pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 585–594 (cit. on p. 15).

[Orfale *et al.* 2005] Adriana Garcia Orfale, Pola Maria Poli de Araujo, Marcos Bosi Ferraz, and Jamil Natour. "Translation into Brazilian Portuguese, cultural adaptation and evaluation of the reliability of the Disabilities of the Arm, Shoulder and Hand Questionnaire". In: *Brazilian Journal of Medical and Biological Research* 38.2 (2005), pp. 293–302 (cit. on p. 3).

[Oberweger, Wohlhart, *et al.* 2015] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. "Hands Deep in Deep Learning for Hand Pose Estimation". In: *Computer Vision Winter Workshop.* . 2015, pp. 1–10 (cit. on p. 15).

[Panteleris and Argyros 2017] Paschalis Panteleris and Antonis Argyros. "Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 575–584 (cit. on p. 18).

[Pedregosa *et al.* 2011] F. Pedregosa *et al.* "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 47).

[Pereira *et al.* 2017] Luís Carlos Pereira, Sylvia Rwakabayiza, Estelle Lécureux, and Brigitte M Jolles. "Reliability of the knee smartphone-application goniometer in the acute orthopedic setting". In: *The journal of knee surgery* 30.03 (2017), pp. 223–230 (cit. on p. 5).

[Platt 1999] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *ADVANCES IN LARGE MARGIN CLASSIFIERS.* MIT Press, 1999, pp. 61–74 (cit. on p. 47).

[Panteleris, Oikonomidis, *et al.* 2018] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. "Using a single RGB frame for real time 3D hand pose estimation in the wild". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE. 2018, pp. 436–445 (cit. on p. 18).

[Pisharady and Saerbeck 2015] Pramod Kumar Pisharady and Martin Saerbeck. "Recent methods and databases in vision-based hand gesture recognition: A review". In: *Computer Vision and Image Understanding* 141 (2015), pp. 152–165 (cit. on p. 13).

[Poier *et al.* 2018] Georg Poier, David Schinagl, and Horst Bischof. "Learning Pose Specific Representations by Predicting Different Views". In: *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).* (to be published). 2018 (cit. on p. 16).

[Romero *et al.* 2017] Javier Romero, Dimitrios Tzionas, and Michael J. Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* (Nov. 2017). URL: http://doi.acm.org/10.1145/3130800.3130883 (cit. on p. 18).

[Rasmussen and Williams 2005] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005. ISBN: 026218253X (cit. on p. 47).

[Santavas *et al.* 2020] Nicholas Santavas, Ioannis Kansizoglou, Loukas Bampis, Evangelos Karakasis, and Antonios Gasteratos. *Attention! A Lightweight 2D Hand Pose Estimation Approach.* 2020. eprint: 2001.08047 (cit. on pp. 18, 20).

[Sharp *et al.* 2015] Toby Sharp *et al.* "Accurate, robust, and flexible real-time hand tracking". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. 2015, pp. 3633–3642 (cit. on pp. 13, 19).

[Simon *et al.* 2017] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. "Hand keypoint detection in single images using multiview bootstrapping". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 1145–1153 (cit. on p. 18).

[Sridhar *et al.* 2013] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. "Interactive markerless articulated hand motion tracking using RGB and depth data". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2456–2463 (cit. on pp. 13, 19).

[Stenger *et al.* 2006] Björn Stenger, Arasanathan Thayananthan, Philip HS Torr, and Roberto Cipolla. "Model-based hand tracking using a hierarchical Bayesian filter". In: *IEEE transactions on pattern analysis and machine intelligence (PAMI)* 28.9 (2006), pp. 1372–1384 (cit. on pp. 12, 17).

[Sun *et al.* 2015] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. "Cascaded hand pose regression". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 824–832 (cit. on pp. 13, 14, 19).

[Scott *et al.* 2010] David L Scott, Frederick Wolfe, and Tom WJ Huizinga. "Rheumatoid arthritis". In: *The Lancet* 376.9746 (2010), pp. 1094–1108. doi: 10.1016/s0140-6736(10)60826-4 (cit. on p. 1).

[Tagliasacchi *et al.* 2015] Andrea Tagliasacchi *et al.* "Robust Articulated-ICP for Real-Time Hand Tracking". In: *Computer Graphics Forum*. Vol. 34. 5. Wiley Online Library. 2015, pp. 101–114 (cit. on p. 13).

[Tajali *et al.* 2016] Siamak Bashardoust Tajali, Joy C MacDermid, Ruby Grewal, and Chris Young. "Reliability and validity of electro-goniometric range of motion measurements in patients with hand and wrist limitations". In: *The open orthopaedics journal* 10 (2016), p. 190 (cit. on p. 4).

[Tang *et al.* 2015] Danhang Tang *et al.* "Opening the Black Box: Hierarchical Sampling Optimization for Estimating Human Hand Pose". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3325–3333 (cit. on pp. 13, 14, 19).

[Tompson *et al.* 2014] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. "Real-time continuous pose recovery of human hands using convolutional networks". In: *ACM Transactions on Graphics (ToG)* 33.5 (2014), p. 169 (cit. on pp. 14, 19).

[Tkach *et al.* 2016] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. "Sphere-meshes for real-time hand modeling and tracking". In: *ACM Transactions on Graphics (TOG)* 35.6 (2016), p. 222 (cit. on p. 13).

[Wan *et al.* 2017] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. "Crossing nets: Dual generative models with a shared latent space for hand pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 7. 2017 (cit. on p. 15).

[Wan *et al.* 2018] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. "Dense 3D regression for hand pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5147–5156 (cit. on p. 15).

[S.-E. Wei *et al.* 2016] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Convolutional pose machines". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732 (cit. on pp. 15, 18).

[Wu *et al.* 2020] Zhenyu Wu *et al.* "MM-Hand: 3D-Aware Multi-Modal Guided Hand Generation for 3D Hand Pose Synthesis". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 2508–2516 (cit. on p. 65).

[Xiong *et al.* 2019] Fu Xiong *et al.* "A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image". In: *Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV)*. 2019 (cit. on pp. 15–17).

[C. Yan *et al.* 2020] C. Yan *et al.* "3D Room Layout Estimation From a Single RGB Image". In: *IEEE Transactions on Multimedia* 22.11 (2020), pp. 3014–3024. doi: 10.1109/TMM.2020.2967645 (cit. on p. 19).

[Chenggang Yan, Gong, *et al.* 2020] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. "Deep multi-view enhancement hashing for image retrieval". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2020) (cit. on p. 17).

[Chenggang Yan, Li, *et al.* 2020] Chenggang Yan, Zhisheng Li, *et al.* "Depth image denoising using nuclear norm and learning graph model". In: *ACM Transactions on Multimedia Computing Communications and Applications* (2020) (cit. on pp. 17, 65).

[Yuan, Ye, *et al.* 2017] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. "Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4866–4874 (cit. on pp. 16, 17, 19).

[Yuan, Garcia-Hernando, *et al.* 2018a] Shanxin Yuan, Guillermo Garcia-Hernando, *et al.* "Depth-based 3D hand pose estimation: From current

achievements to future goals". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2636–2645 (cit. on p. 16).

[Yuan, Garcia-Hernando, *et al.* 2018b] Shanxin Yuan, Guillermo Garcia-Hernando, *et al.* "Depth-based td3D hand pose estimation: From current achievements to future goals". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 37, 39).

[Zimmermann and Brox 2017] Christian Zimmermann and Thomas Brox. "Learning to estimate 3D hand pose from single RGB images". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4903–4911 (cit. on pp. 18, 19, 26, 27).

[Zhang *et al.* 2020] Zhaohui Zhang, Shipeng Xie, Mingxiu Chen, and Haichao Zhu. "HandAugment: A simple data augmentation method for depth-based 3D hand pose estimation". In: *arXiv preprint arXiv:2001.00702* (2020) (cit. on p. 65).

[Zimmerman *et al.* 1987] Thomas G. Zimmerman, Jaron Lanier, Chuck Blanchard, Steve Bryson, and Young Harvill. "A Hand Gesture Interface Device". In: *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*. CHI '87. Toronto, Ontario, Canada: ACM, 1987, pp. 189–192. isbn: 0-89791-213-6. doi: 10.1145/29933.275628. url: http://doi.acm.org/10.1145/29933.275628 (cit. on p. 11).

[Zhou *et al.* 2016] Yimin Zhou, Guolai Jiang, and Yaorong Lin. "A novel finger and hand pose estimation technique for real-time hand gesture recognition". In: *Pattern Recognition* 49 (2016), pp. 102–114 (cit. on p. 15).