FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Computational models for robot-induced hallucinations in Parkinson's Disease

- Automatized pipeline to extract motion features of the sensorimotor feedback task

Duarte Teixeira Rodrigues

Master in Bioengineering

Supervisor: Bruno Herbelin, PhD

Supervisor: Fosco Bernasconi, PhD

Co-supervisor: João Paulo Cunha, PhD

October 06, 2022

© Duarte Teixeira Rodrigues, 2022

Computational models for robot-induced hallucinations in Parkinson's Disease - Automatized pipeline to extract motion features of the sensorimotor feedback task

Duarte Teixeira Rodrigues

Master in Bioengineering

Approved by . . . :

President: Fernando Jorge Monteiro, PhD Full Professor

Referee: Swann Pichon, PhD

October 06, 2022

Resumo

A doença de Parkinson é uma doença neurodegenerativa progressiva complexa, que segundo a Organização Mundial de Saúde, está a afetar mais de 8.9 milhões de pessoas em todo o mundo [10], sendo esta a segunda doença neurológica mais prevalente. Um dos sintomas mais predominantes nesta doença são as alucinações, das quais as alucinações de presença são um subgrupo, em que se estimam ser preditores para uma evolução grave da doença, tornando-as um fator chave a ser estudado.

Os recentes avanços em neurociência cognitiva realizado pelo laboratório do Professor Olaf Blanke oferecem um novo método baseado em robótica, de forma a induzir uma alucinação de presença (HP) específica e clinicamente relevante em participantes saudáveis [17], através de conflito sensorial motor, bem como em pacientes DP [14]. Na tarefa comportamental, os participantes são solicitados a realizar um movimento repetitivo de toque com o indicador direito, que está preso ao robot frontal. Atrás deles, o robot traseiro reproduz os movimentos deles em tempo real, fornecendo um feedback congruente tátil nas suas costas. A tarefa é dividida em 2 condições, a síncrona (sem atraso), e a assíncrona onde se introduz um pequeno atraso (500ms) entre o movimento dos participantes na frente e o toque do robot nas costas. Embora cada participante possa realizar a tarefa de forma diferente, uma questão é levantada se os controlos saudáveis (HC) e os pacientes PD exibem traços, movimentos ou comportamentos comuns durante a execução da tarefa. Para além disso, esta mesma questão pode ser feita entre os pacientes PD que costumam ter PH durante a sua vida diária (denominadas de PD-PH) ou aqueles que não tem (PD-nPH).

A gravação capturada para registrar a experiência foi analisada com o DeepLabCut, um software de estimativa de pose sem marcadores, que usa a visão computacional para prever as posições de pontos-chave definidos pelo usuário. No caso desta experiência, foram selecionadas as partes do corpo, assim como alguns elementos do robot. Como resultado, um modelo preciso de aprendizagem aprofundado é capaz de prever a trajetória das partes do corpo durante o vídeo. Com base nos dados de movimento, foi desenvolvida uma pipeline para processar e computar características de movimento que descrevem as características comportamentais dos participantes, ao realizar a tarefa. Estas métricas foram então comparadas entre grupos e condições experimentais, revelando alguns comportamentos diferentes influenciados pela manipulação do robot.

Os participantes saudáveis são capazes de manter a mesma postura, com comportamento consistente (mesma amplitude de toque e velocidade) em todas as tentativas, e conseguem seguir as diretrizes da tarefa. Por outro lado, os pacientes com PD, devido à sua rigidez e fraqueza em geral, executam o toque usando todo o tronco ao mesmo tempo, num movimento de balanço para a frente e para trás. Dentro deste grupo, o PD-PH revela-se ser o mais sensível à tarefa, apresentando, em média, maiores amplitudes de movimento e uma maior extensão do braço. No entanto, para obter resultados adequados, é necessário garantir uma configuração prática, que possa atingir determinados requisitos tecnológicos, sendo o mais importante ter um ambiente de gravação padrão entre os participantes.

A partir deste projeto é possível ter uma ideia de quais devem ser as condições tecnológicas

e orientações gerais para que o pessoal não-técnico possa manipular o sistema, caso um sistema robótico e as configurações das câmaras sejam instalados em ambientes clínicos diferentes.

Abstract

Parkinson's disease is a complex progressive neurodegenerative disease, that according to the World Health Organization, is affecting more than 8.9 million people worldwide [10], it being the second most prevalent neurological disease. One of the symptoms most prevalent in the disease are hallucinations, from which presence hallucinations are a subgroup, that has been estimated to be a putative predictor to a severe evolution of the disease, making them a key subject to be studied.

Recent advances in cognitive neuroscience performed by the laboratory of Prof Olaf Blanke offers a new method based on robotics to induce a specific, clinically relevant presence hallucinations (PH) in healthy participants [17], through sensorimotor conflict, as well as in, PD patients [14]. In the behavioral task, the participants are asked to perform a repetitive poking movement with their right index finger, which is attached to the front robot. Behind them, the back robot reproduces their movements in real time, providing a congruent tactile feedback on their back. The task is divided into 2 conditions, the synchronous (no delay) and the asynchronous, where by introducing a short delay (500ms) between the participants' movement on the front and the robot's touch in the back. Although each participant may perform the task differently, a question is raised whether healthy controls (HC) and PD patients exhibit common traits, movements or behaviors while performing the task. Furthermore, the same question can be prompted between those PD patients that usually have PH during their daily lives (denominated as PD-PH) or the ones that do not (PD-nPH).

The recording captured to log the experiment was analyzed with DeepLabCut, a markerless pose estimation software, that uses computer vision to predict the positions of user-defined keypoints. In the case of the experiment, the body parts were selected as well as some elements of the robot. As a result, an accurate deep learning model is able to predict the trajectory of the body parts during the video. Based on the movement data, a pipeline was developed to process and compute motion features that describe the behavioral characteristics of the participants, when performing the task. These metrics were then compared between groups and experimental conditions, revealing some different behaviors influenced by the robot manipulation.

Healthy participants are able to maintain the same posture, with consistent behavior (same amplitude of poking and speed) across trials, following the task guidelines. PD patients, on the other hand, due to their rigidity and overall weakness, do the poke using the whole torso at the same time, in a rocking motion back and forth. Inside this group, the PD-PH reveals to be more sensible to the task, presenting on average higher amplitudes of motion and wider arm extension. However, to achieve proper results, it is necessary to guarantee a practical setup, that fulfills certain technological requirements, being the most important to have a standardized recording environment across participants.

From this project, it is possible to have an idea of what should be the technological conditions and guidelines for non-technical staff to manipulate the system, in case a robotic system and camera setup are installed in different clinical settings.

Acknowledgements

Not only this amazing master's project, but my entire course, was undoubtedly one of the best experiences in my life and it would not have been so, if it wasn't for the support and people I had around me, not only in the last 6 months I spent developing the thesis, but also all the others that stood by me during the times of hardship and worry, intense study and exams or simply in fun and leisure.

When I chose to embark in this adventure in Switzerland, the choice of my supervising team was left to chance, but now, 8 months later, I can say that I was a very lucky man to get both Bruno and Fosco as my supervisors. Either in engineering or science, they both directed me to clear goals, offering guidance when setbacks and inconveniences appear along the way. I could have not asked for a better team. This was proven by the great achievements reached in my project.

On a more personal note, there is no bigger pillar that embraces me than Beatriz. Over all aspects of life, she has been the dearest of people that continued by my side on all highs and lows. This unconditional support and constant encouragement were the key for me to build up and conclude this life chapter with such a happy ending. I know that the story does not finish here, and many more are to come.

I am lucky as the sources of unconditional love do not stop here, as all my family, specially my mother, upon all her sacrifice and patience, was the person that was there every single day tending and caring for me, so that I could "grow" into the best version of myself and have the best life I could think of. My father took the role of counsellor, always available to give me a piece of his wisdom, and as a symbol for me, a beacon of unrelenting force that never stops fighting the problems of life and business, independently of how big and scary they might look.

Finally, I could not dismiss my dear comrades that never abandoned my flanks, that stood back-to-back with me, overcoming all projects, studies, exams, meetings... honestly true brothers and sisters in arms. My right and left arm, Ricardo and Mariana, were the heroes that shared with me all the struggles on the courses and work, but also were the very best company, always present in every lunch, class, metro or car ride, or even a simple call. However, the last battle, the 6 months of thesis abroad, had special characters that never looked away when seeing me in need of help or just to vent about life. The elements of a round table of researchers, warriors, students or fighters, whatever their title, they were the definitely the most incredible friends I could ask for. The 4 knights that stood alongside with me, Neža and her unstoppable caring nature to ensure the well-being of everyone around, either with long meaningful talks or chocolate; Louis, an angel-on-earth that stopped at nothing to guarantee both my best performance in science and that I taste the most unforgettable delicious food; Juliette AKA my coach, that in spite her permanent rushing life, always had the time to check up on me and give thoughtful advice and motivational sessions and finally but not least, Arthur, a partner in crime that helped me keep my sanity with coffee breaks (taking that one time he almost made me crawl up the walls) and games in VR, but always present with the rest as shoulder I could hold on to.

I thank you all as the glue that was able to keep my puzzle pieces together at all times.

"If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat."

Sun Tzu, The Art of War

Contents

Acknowledgements v

1 Introduction 1
- 1.1 Context and Motivation . . . 1
- 1.2 Research Aims . . . 2
- 1.3 Dissertation Structure . . . 3

2 Minor Hallucination in Parkinson's Disease 5
- 2.1 Clinical Presentation and Pathophysiology of PD . . . 5
- 2.2 System to Induce Presence Hallucinations . . . 8
- 2.3 Towards an Automatic Hallucinatory Analysis . . . 12

3 Tracking Technologies for the Clinical Sensorimotor Task 13
- 3.1 Selection of Motion Capture Technology . . . 13
  - 3.1.1 State-of-the-art Comparison . . . 18
- 3.2 DeepLabCut Software . . . 20
  - 3.2.1 DeepLabCut in Parkinson's Disease . . . 22

4 Project Methods 25
- 4.1 DLC Pipeline for Clinical Recordings . . . 25
  - 4.1.1 Project Creation and Labeling . . . 25
  - 4.1.2 Adaptive Cropping Algorithms . . . 27
- 4.2 Extraction of Behavioral Features . . . 31
  - 4.2.1 Database Standardization . . . 31
  - 4.2.2 Data Cleaning and Filtering . . . 32
  - 4.2.3 Feature Computation . . . 34

5 Study A — Controlled Condition Experiment 39
- 5.1 Database . . . 39
  - 5.1.1 Experimental Protocol . . . 40
  - 5.1.2 Sociodemographics . . . 41
- 5.2 Motion tracking Specifications and Model Performance . . . 43
- 5.3 Results . . . 44
- 5.4 Discussion . . . 47
- 5.5 Final Remarks . . . 48

6 Study B — Close to Real Condition Experiment **51**
6.1 Database . . . 51
6.1.1 Experimental Protocol . . . 52
6.1.2 Sociodemographics . . . 53
6.2 Motion tracking Specifications and Model Performance . . . 55
6.2.1 Tracking on Multiple Angles *VS* Full Dataset . . . 55
6.3 Results . . . 57
6.4 Discussion . . . 59
6.5 Final Remarks . . . 60

7 Conclusion and Future Work **63**

References **67**

A Appendix **73**
A.1 Laboratorial Tests for PD diagnosis . . . 73
A.2 PD Treatment Options . . . 73
A.3 Socioeconomic impact . . . 75
A.3.1 Socioemotional Impact on Quality of life . . . 75
A.3.2 Economic burden and cost of PD . . . 76
A.4 MoCap System Descriptions . . . 79
A.5 Complete Step-by-Step Guide of DeepLabCut . . . 80
A.5.1 Create a project — Steps 1–5 . . . 80
A.5.2 Frame Selection and Labeling — Steps 6–9 . . . 81
A.5.3 Neural network training and Evaluation - Steps 10-12 . . . 82
A.5.4 Video Analysis — Steps 13–15 . . . 87
A.5.5 Active Learning — Steps 16–18 . . . 88
A.5.6 Tables of the statistical results . . . 89

List of Figures

2.1 Sketch of Mid-Sagittal cut with highlighted brain areas, especially the midbrain in green [36]. 6

2.2 Histological comparison of the substantia nigra pars compacta (SNpc) in a control (A and B) and in a PD brain (C and D) stained by hematoxylin and eosin. Loss of DA neurons is noticeable [32]. 7

2.3 On the left, a model of how a participant interacts with the robot (Lateral view). On the center, a model of how the design of the robotic system. The participant moves the arm attached to the front robot, while the back robot mimics the movement, giving a tactile feedback on the back. On the right, an example of the back exploration done in this the behavioral experiment. Each dot indicates a point in which the back robot touched on the back of the participant [14, 15] 10

2.4 A model for determining the sensory consequences of a movement (adapted from [16]) . For each motor command, an efferent copy is generated and used to predict the outcome sensory feedback. These sensory feedback predictions are then compared to the actual sensory feedback, thus giving a sensory discrepancy that is used to attribute the origin of the sensory feedback (self-attributed or attributed to an 'external agent'). 11

2.5 riPH in patients with PD. (A) riPH in patients with PD (asynchronous versus synchronous stimulation). Each dot indicates the individual rating of the intensity of the riPH [PD-PH (purple) and PD-nPH (orange)]. The dots with the bar on the left and right sides indicate the mixed-effects linear regression between PD-PH and PD-nPH. Error bar represents 95% confidence interval. **P $\leq$ 0.01. (B) riPH in patients with PD (asynchronous versus synchronous stimulation). Each dot indicates the individual rating of the intensity of the riPH. The dots with the bar on the left and right sides indicate the mixed-effects linear regression between asynchronous (black) and synchronous (gray) sensorimotor stimulation. Error bars represent 95% confidence interval. *P $\leq$ 0.05 [14]. 12

3.1 Number of scientific publications using DeepLabCut, over the last 6 years [1]. . . 19

3.2 DeepLabCut workflow. The diagram delineates the workflow, as well as the directory and file structures. The process of the workflow is color coded to represent the locations of the output of each step or section [40]. 20

3.3 Example of labeled frames, extracted from the smartphone video. 23

4.1 Example frame of the keypoints marked on the participant. Notice that the left elbow is not visible, so it is not labeled. 26

4.2 Cropping pipeline to emphasize the region of interest (ROI), by removing unnecessary background, resulting in a zoom in effect of the participant. 28

4.3 Example of the failure cases of the cropping algorithm. The top row encompasses the case where one of the predictions is out of the ROI. The middle row represent the case when the ROI is too big and goes out of frame. In the bottom row, the participant is so far away, that the ROI is smaller than the expected final resolution. In those cases, the ROI is set to be 512x288, to avert the need of interpolating pixels, trying to upsample the image. 29

4.4 The normalization inside the ROI, allows to get the location of the labeled point in the final "DownCrop" videos. 30

4.5 The highlighted line shows acceptable distance for a prediction to still be considered on the wrist, with a good accuracy. On the DonwCrop version, even if the predictions have jitter, it does not influence as much the accuracy, since it has a larger region within each body part. 30

4.6 Plot showing the movement of the index, front and back. The Y scale is the distance of the finger from the left border and the participant is oriented to the right. This means that he moves his finger forward on the local *maxima* of the signal. Therefore, wavy pattern represents the repetitive pokes done by the participant. The following pipeline tries to build a solution to clean this signal, preparing it to extract motion features that are not affected by noise. 32

4.7 Plot of the movement before and after the data cleaning and filtering steps. 34

4.8 On top, plot of the raw movement of the shoulder of the non-moving arm (same participant as plotted before). The movement is distorted by the noise making it impossible to distinguish what is noise or an actual behavior. On the bottom, the high frequency noise that contaminates the signal. The amplitude of the noise and the raw signal is very similar, indicating that the tracking was not successful. . . . 36

5.1 Example frames of the log videos of different participants, in the database of study A. 39

5.2 General experimental protocol for a participant in study A. In the behavioral task, the back stimulation is randomized, not necessarily being steps 3 and 4. Inside this block, the async and sync conditions are also presented in a random order, between participants. Each condition lasts 2 minutes. The total behavioral task lasts 1 hour, because the participants are interviewed between blocks. 40

5.3 On the left, the plot of the evolution of the cross entropy loss. On the left, the comparison of the train and test error (RMSE), for the 8 snapshots saved during the training process. 44

5.4 Plots of the statistical results of the motion features of the moving finger. The top row is the HC *VS* PD (A to C) and the bottom the PD-PH *VS* PD-nPH (D-F). From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.1. 45

5.5 Plots of the statistical results of the motion features of the elbow (left) and shoulder (right). The top row is the HC *VS* PD (A to D) and the bottom the PD-PH *VS* PD-nPH (E-H). The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.2. 46

6.1 Example frames of the log videos of different participants, in the database of study B. 51

6.2 General experimental protocol for a participant in study B. In the behavioral task, the back stimulation is randomized, not necessarily being the right hand first. Inside this block, the async and sync conditions are also presented in a random order, between participants. Each of the 4 conditions lasts 2 minutes. The total behavioral task lasts 1 hour, because the participants are interviewed between blocks. . 53

6.3 Box plots of the RMSE do the "*robot_top*" keypoint, in all the models. 56

6.4 Plots of the statistical results of the motion features of the moving finger. The top row is the HC *VS* PD (A to C) and the bottom the PD-PH *VS* PD-nPH (D-F). From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.3. 58

6.5 Plots of the statistical results of the motion features of the elbow (left) and shoulder (right). The top row is the HC *VS* PD (A to D) and the bottom the PD-PH *VS* PD-nPH (E-H). The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.4. 59

A.1 Total economic burden of Parkinson's disease in the U.S., divided by components, in 2017 (in billions) [7] 77

A.2 Direct cost of PD in the UK according to the Hoehn and Yahr scale, on the left; and to age, on the right (n = 423). 78

A.3 Methods for frame selection. The toolbox contains three methods for extracting frames, namely, by clustering based on visual content (left image) , by randomly sampling in a uniform way across time (center image) , or by manually grabbing frames of interest using a custom GUI (right image). 81

A.4 Ensemble of image transformation examples done for data augmentation. 83

A.5 On the left, 2 example frames of a project intended to track a mouse paw while it grasps a tube. The image on top, the hand is clearly visible, so the model can identify accurately where joints should be. If a keypoint is selected (double arrow) the correspondent score map is the small square with the dark blue background, since the likelihood of the body part is low in all of the image, except in the red region. If a log-transform is applied and the red spot is zoomed, can clearly be see an even smaller region with higher intensity, that results in predicted position of the body part (red dot). In the bottom image, the paw was tilted sideways, so the model cannot identify any joint, even though one of the fingers is still visible (double arrow). This translates to a scoremap of low confidence, that when transformed has a broader region without any specific intense spot, leading to bad predictions (purple dots). [34] 84

A.6 Both benchmark plots derive from the analysis of the same dataset presented in [33], where more details can be found. On the left, the different models are compared regarding there inference speed, being the MobileNetsV2 faster then ResNets. The EfficinetNets, depending on the depth of the architecture can vary substantially. On the left, the error was measured for the same models. The deeper the network, the better the accuracy, in the case of the MobileNetsV2 and ResNet. In case of the EfficientNet, the faster one has the biggest error between them, followed by the 5. The 6 presents the lowest error of the overall architectures. 86

A.7 On the right, an example of a frame after the video analyzes, displaying the tracked keypoints. On the left, a combination of plots that help understand the analysis. On the top left, each color represents the trajectory of a certain body part, during the video, allowing to understand what area was explored. On the top right, a representation of how the body part moved in the X and Y axis, over time. On the bottom left, is represented the likelihood that a body part is identified in the correct position, over time. Ideally, all body parts would have a straight line on 1, which would mean that it was always identified confidently. On the bottom right, a measure of how much a body part moved in consecutive frames. Most of the values are centered around 0 to 20, which means that the body moved. However, if peaks appear could be an indication of miss predictions [34, 40] 87

List of Tables

3.1 Comparison of state-of-the-art motion capture systems. 18

5.1 Sociodemographic comparison between HC and PD 41
5.2 Sociodemographic comparison between PD-PH and PD-nPH 42

6.1 Sociodemographic comparison between HC and PD 54
6.2 Sociodemographic comparison between PD-PH and PD-nPH 54
6.3 Train and test error for each one the models, related to the number of labeled frames. 56

A.1 Direct medical cost of Parkinson's disease by age, gender, insurance coverage, and types of service (in 2017) [7]. *PWP – Person with Parkinson 78
A.2 Statistical results of the motion features of the moving finger 89
A.3 Statistical results of the motion features of the elbow and shoulder 90
A.4 Statistical results of the motion features of the moving finger 90
A.5 Statistical results of the motion features of the elbow and shoulder 91

Abbreviations and Symbols

ARIMA	AutoRegressive Integrated Moving Average
ANOVA	Analysis of variance
cm	centimeters
CNN	Convolutional Neural Network
DLC	DeepLabCut
FPS	Frame Per Second
GABA	Gamma-AminoButyric Acid
GUI	Guide User Interface
HC	Health Control
HD	High Definition
Hz	Hertz
IMU	Inertial Measurement Unit
LEDD	Levodopa equivalent daily dose
LNCO	Laboratory of Cognitive Neuroscience
maDLC	multi-animal DeepLabCut
MBRS	Modified Bradykinesia Rating Scale
MDS-UPDRS	Modified Disease Scale-Unified Parkinson's Disease Rating Scale
MoCap	Motion Capture
saDLC	single-animal DeepLabCut
PD	Parkinson Disease
PD-PH	Parkinson Disease patient with symptomatic presence hallucinations
PD-nPH	Parkinson Disease patient without symptomatic presence hallucinations
PH	Presence Hallucination
riPH	Robot Induced Presence Hallucination
sPH	Symptomatic Presence Hallucination
PSD	Power Spectral Density
ResNet	Residual Network
RGBD	Red-Green-Blue Depth
RMSE	Root Mean Square Error
SNpc	Substantia Nigra Pars Compacta
SNpr	Substantia Nigra Pars Reticulata
STN	Subthalamic Nucleus

Chapter 1

Introduction

1.1 Context and Motivation

Parkinson's disease (PD) is a neurological disorder that affects more than 1 million people, in the United States of America (USA) alone, and forecasting estimations predict that the affected population will grow 20% in the next 10 years [32]. It is a pressing issue to be researched and state-of-the-art laboratories are investing in multiple solutions to develop new ways to diagnose the disease, on the earliest stage possible and implement various types of innovative treatments, either with antiparkinsonian drugs or surgery, through deep brain stimulation.

A recent study in the USA, estimated that the economic burden of PD was about $51.9 billion, divided in $25.4 billion on direct medical costs, such as drugs, surgery, consultation and in/outpatient visits; and $26.5 billion in indirect and non-medical costs. [7] This adds up to a rise of $2.4 billion in 15 years, only taking in consideration the direct costs [21].

The urgency to tackle the diagnosis of PD as early as possibly is crucial to control the neurodegeneration and, potentially prevent the patient from getting serious symptoms or ever needing surgery.

The onset of PD, is diagnosed when the motor symptoms of the patient start to be noticeable, however, before this stage, there are also common signs that can direct doctors towards the same diagnosis.

Hallucinations in Parkinson disease are a frequent phenomenon, ranging from 40% to 75% of patients depending on the disease evolution period [23, 3]. Most frequent and early type of hallucinations in Parkinson disease are called Minor Hallucinations [23, 42, 6], which encompasses Presence Hallucination (also called Feeling of a Presence) -– "the strange sensation that somebody is nearby when no one is actually present and cannot be seen" [12] –-; passage hallucination -– the brief vision of a person or an animal passing sideways –-; and visual illusions -– the visual misperception of real objects or living entities. These have been estimated to be a putative predictors of major negative clinical outcomes such as delusions, cognitive decline and dementia [14, 15]. Therefore, Minor Hallucinations can be seen as an estimator for a more severe evolution of the disease, making them a key subject to be studied.

Hallucinations are subjective and private experiences, and current methods to investigate and evaluate them are based on self-reports depicted by the patients and subjective interpretations of those reports by doctors or investigators, known to be associated with several limitations.

Recent advances in cognitive neuroscience performed by the laboratory of Prof Olaf Blanke offers a new method based on robotics to induce a specific, clinically relevant presence hallucinations (PH) in healthy participants, through sensorimotor conflict. Such robot-induced PHs (riPHs) were induced repeatedly and in real time in the laboratory or in different clinical settings and were controlled by a robotic device across several experimental conditions. The method for riPH was developed and first investigated in healthy participants and subsequently tested in patients with Parkinson's disease.

In effect, the robot stimulation and its associated induction of PH affects the behavior of the participants in many ways, triggering different reactions, like an increment of the normal speed at which they interact with system, or to increase the tremors with random twitching. As an example, a participant report after the behavioral experiment – "[It was] Like something or someone was behind me.(...) I was scared./ It was like a ghost or a presence" (supplementary material of [14]). Although each participant may have a different experience, a question is raised whether healthy controls (HC) and PD patients exhibit common traits, movements and behaviors while performing the task. Furthermore, the same question can be prompted between the PD patients that usually have PH during their daily lives (denominated as PD-PH) or the ones that do not (PD-nPH).

This would allow to study, and, in the future, diagnose the well-being and health condition of a participant based on their motion pattern in the experiment. A significant difference is expected between the conditions of testing and the mentioned groups.

1.2 Research Aims

The performance and behaviour of each participant in the robotic task is evaluated by subjective metrics, either by the personal description given by the participant or by assessments evaluated by the clinical personnel.

The main objective of this project is thus to provide numerical estimations of motion features that describe and quantify objectively the behavior of the participant in a non-obtrusive way. In a future application of this technique in clinical settings, its results could be integrated in the clinical evaluation of PD patients undergoing the robotic task, for a more precise evaluation.

Therefore, two main research question emerge, one the scientific context of the experiment and another regarding the methodology that should be implemented to achieve it. Thus, the questions are the following:

i) Are the movement components od participants during the sensorimotor robotic task differing between HC and PD or between PD-PH and PD-PH? If so, which are the body parts, and respective motion features, that indicate the difference between groups?

ii) What are the technological requirements for the motion tracking setup to answer the previous question?

1.3 Dissertation Structure

To answer this questions and achieve the main objective of the project, the dissertation starts by an overview of the Parkinson's disease, on Chapter 2, and its associated factors and symptoms. From all of them, the emphasis will fall into hallucinations, and in particular on the research work for systematically inducing hallucinations using the robotic task.

Each experiment done with the system is recorded and the motion of the participant could thus be captured and predicted. For that, an outline of what are the state-of-the-art tracking systems is introduced, on Chapter 3, together with their advantages and setbacks, in order to select which one is the best for the current application.

At this stage, the dissertation can focus on my project where I apply the selected technology to the studies on induced hallucinations. Firstly, I describe the methods developed to optimize the tracking of the participants and, secondly, how to extract motion features from those movements, both presented in chapter 4.

Subsequently, I describe the studies, on chapters 5 and 6, in which I applied the aforementioned pipelines and discuss the results achieved in both of them. In the end, I perform a comparison between both, explaining the reason for the differences, on chapter 7.

On the final chapter 8, I summarize my main conclusions derived from the project, as well as the answers to the posed research questions and what future steps should be applied in the experimental task and in the processing pipelines to improve even more the results.

Chapter 2

Minor Hallucination in Parkinson's Disease

Parkinson's disease is a complex progressive neurodegenerative disease, which prevalence is aggravated in older generations, affecting approximately 0.5–1% among those 65–79 years of age, rising to 1–3% among persons 80 years of age and older [29].

The disease that can be manifested in a multitude of signs, both visibly affecting directly the movement, or under the shadow of the mind, as non-motor symptom that disturbs silently the individual. Many of these symptoms, such as depression, anxiety and hallucinations directly affect their quality of life, even more than the motor impairments [20]. Therefore, is important to establish a generalized method to study and quantify these symptoms that does not depend solely on their descriptions of randomly occurrences in everyday life or psychological questionnaires, in order for doctors to improve their diagnostics procedures and, consequently, to prescribe the correct medication to control the disease.

For that purpose, it is of primary importance to understand the origin pf the disease and how it is generally manifested by the patients.

2.1 Clinical Presentation and Pathophysiology of PD

As a neurological disease, the organ most affect by PD is the brain. Taking a look to the brain anatomy, from a top-bottom approach, the cerebral cortex is the most outside layer of the brain comprising of the 4 main lobes (frontal, temporal, parietal and occipital) responsible for managing the higher-level cognitive thoughts and emotions, as well as processing the sensory stimulus, like auditory and visual cues. The next region after the cortex is the "corpus callosum" responsible to connect and mediate information between the left and right hemispheres. Adjacent to this region is the diencephalon, which consists of the thalamus and hypothalamus. These 2 brain regions are responsible for controlling body temperature, blood sugar levels and also serve as the link between the nervous and endocrine system [41], regulating hormone release through the pituitary gland. In the deeper section of the brain, connecting the *cerebrum* (all previously mention areas) and

cerebellum (balance coordination center) is the midbrain. This region is important, since it is the focus of neurodegeneration that originates PD, in specific one of it's components, the "substantia nigra" (SN), the dopaminergic center of the brain.

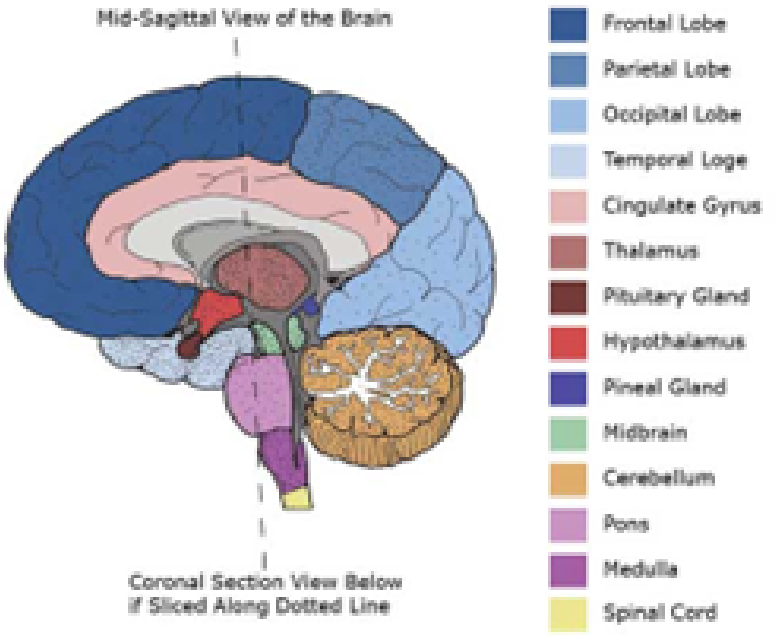

Figure 2.1: Sketch of Mid-Sagittal cut with highlighted brain areas, especially the midbrain in green [36].

The SN is a part of the midbrain, however its function is correlated to the basal ganglia, since it actuates as the main region that controls voluntary movement, cognitive planning, emotions and reward functions, and even cognition and learning [8]. This region is divided into 2 parts:

- *Pars Compacta* (SNpc): Section that contains the dopaminergic neurons
- *pars reticulata* (SNpr): Section that contains the inhibitor gamma-aminobutyric acid-containing (or GABAergic) neurons.

This distinction is important since the only noticeable phenotypical change in idiopathic PD (disease diagnosed from an unknown reason) is seen in transverse section cuts of the brain stem, where almost all cases present loss of the darkly pigmented areas in the *substantia nigra pars compacta*. This pigmentation loss directly correlates with the death of dopaminergic (DA) neuromelanin-containing neurons in the SNpc. Quantitative morphometric studies in *postmortem* PD brains have calculated approximately 30% loss of DA neurons in the SNpc by motor symptom onset (first appearance of motor impairments) and after the motor symptoms appear, nigral DA neuron loss increases up to 60% or higher and is strongly associated with the severity of motor features and disease duration [32].

As the disease progresses, when the cerebellar region starts to be affected, the motor stage begins, and the patients start to demonstrate changes of movement in one of the sides of the body,

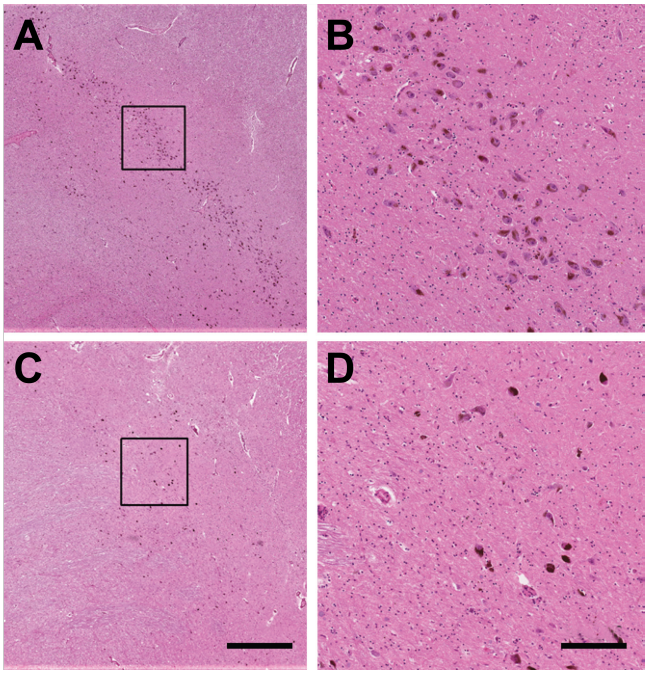

Figure 2.2: Histological comparison of the substantia nigra pars compacta (SNpc) in a control (A and B) and in a PD brain (C and D) stained by hematoxylin and eosin. Loss of DA neurons is noticeable [32].

as PD is classified as an asymmetrical condition. The most common symptoms of this stage are bradykinesia (slowness and lack of spontaneous movement together with a decrement of speed and amplitude), tremors (low frequency shaking of the distal part of the limb), rigidity (uncontrolled *tonus* of the muscle, leading to pain and joint stiffness) and gait imbalance (the equilibrium and balance are affected, making the patient prone to falling) [29, 28].

The clinical diagnostic for PD, since it is considered a movement disorder, is based on these motor impairments shown by patient. The first step to initiate the disease diagnosis is the self-report of some unusual symptoms, that may raise awareness of the doctor to a possible PD diagnostic. The patient is then redirected to a neurologist that assesses several motor skills. These tests are fundamental to detect and evaluate the progress of the most common motor symptoms (mentioned above), being mandatory the presence of bradykinesia [2] as this is the key symptom of PD. The results of the assessments depend greatly on the subjective opinion of the neurologist (being the test commonly performed with more than one doctor present), not existing in the clinical environment any direct quantifiable diagnostic test that indicate the definite presence of PD.

In order to assess the evolution of the disease, multiple rating scales were created to analyze multiple aspects of the patient's life. The most used one is the Hoehn and Yahr Scale, used to rate the progress of PD symptoms as well as their level of disability. The Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is a comprehensive tool to monitor the course and evolution of Parkinson's in multiple aspects, such as, the cognitive

functions, behavior and mood (UPDRS I), the activities of daily living (UPDRS II), the degree of disability and motor impairment (UPDRS III) and the complications of the treament (UPDRS IV) [4, 26]. Another test used to evaluate the cognitive decline of the patient is the Montreal Cognitive Assessment (MOCA), where skills like memory, language and attention are evaluated [39].

Although Parkinson's disease is considered mostly a motor disorded, in the early stages of PD, as early as 12-14 years before the diagnosis, the patient can begin to go through the pre-motor or prodromal phase, where the disease's impact starts from peripheral autonomic nervous system spreading to the central nervous system, creating symptoms like hyposmia (loss of the sense of smell), tendency for anxiety and even minor hallucinations (passage, presence and visual illusions as described in the introduction 1.1). This non-motor semiology starts before the irreversible state of dopaminergic neuron loss [29].

Moreover, there is growing consensus that hallucinations in PD continue to progress throughout the disease [11, 42, 5] with the typical scenario starting with minor hallucinations with preserved cognition, that evolves towards visual hallucinations and cognitive decline, to even, dementia. Because of the earlier onset of minor hallucinations and their likely transition towards visual hallucinations, it has been argued that minor hallucinations, including PH, may be an early indicator of a more severe form of Parkinson's disease [13]. Therefore, the extreme importance of grasping the concepts and correlates from where they stem, to be able to understand them and, consequently, identify them in patients to guarantee they are optimally followed and treated.

2.2 System to Induce Presence Hallucinations

Despite the elevated frequency and clinical importance of hallucinations in several neurological and psychiatric diseases [38, 5], hallucination research is hampered by the low number of devices and experimental approaches for hallucination induction. Previous attempts to elicit hallucinations, such as drug-induced hallucinations (like psychedelics) or those induced by automatized visual stimulation (for instance, flicker-induced hallucinations), failed to induce clinically relevant hallucinations (*i.e.*, resembling those experienced by patients in daily life). Most times, these methods induce complex multisensory hallucinations, instead of specific hallucinations, making it not adapted for conducting studies between participants. Moreover, the intensity of such induced hallucinations may vary and can be associated with more severe alterations of consciousness (i.e., clouding of consciousness, delirium and agitation) that limit their impact for clinical, legal and ethical reasons.

Hence, the clinical presentation of the hallucination intended to be induced needs to be well-defined and of relevance to the research and analysis of Parkinson's disease early stages. The most common type of minor hallucinations are presence hallucinations (PH), present in 25-42% of the PD population, which leads it to be very well investigated and described in scientific literature.

Although this phenomena can be observed in Parkinson's Disease [24, 11, 6], it was first described in psychology in 1902 by William James [44], and later in 1913 introduced in psychiatry by Karl Jaspers who reported this phenomenon in schizophrenic patients [27]. It is known to occur

in others neurological disorders [19, 31] such as epilepsy and Lewy Body Dementia, and can also appear in healthy people facing extreme conditions and/or circumstances [18, 25, 37]. This phenomena are known to occur during a short duration (a few seconds in most of the occurrences). The felt presence usually appears behind or on the side of the subject, and it is often perceived as a shadow that is mimicking posture and gesture [12].

One of the major first insights on the neural correlates of Presence Hallucinations comes from a single case study on a patient who was undergoing evaluation for surgical treatment of epilepsy, where focal electrical stimulation of the left temporoparietal junction repeatedly produced a Feeling of a Presence of another person in her extra-personal space [12]. This illusory presence was closely mimicking the patient's body position and gesture, as well as interfering with some cognitive tasks she was performing. Furthermore, as the temporoparietal junction is known to be involved in self-processing, self-other distinction, the integration of multisensory body-related information, and other illusory own-body perceptions, the authors argued that the Presence Hallucination may have occurred due to a disturbance in the multisensory and sensorimotor processing of the body and the self.

Results from this single case study were later corroborated by a lesion analysis study performed in 12 patients reporting frequent occurrence of Feeling of Presence [17]. This analysis showed that key regions associated with the Feeling of a Presence were the temporoparietal cortex, the frontoparietal cortex and the insular cortex.

Based on these data, a prominent model that describes hallucinations as a disturbance or misattribution of self-related predictive sensory signals [17, 22] developed a robotic stimulation system that exposed participants to temporally and spatially conflicting sensorimotor signals.

At the study, Blanke et al. (2014) [17] reported first results of the novel-designed robot able to induce Feeling of a Presence experimentally in healthy subjects, through sensorimotor conflict (Figure 2.3 A). The robotic system is composed of two parts: a front robot positioned in front of the participant, and a back robot positioned behind the participants. Participants are asked to perform a repetitive poking movement with their right index, which is attached to the front robot. Behind them, the back robot reproduces their movements in real time, providing congruent tactile feedback on their back, while under in a condition of sensory deprivation (to prevent the subject from seeing and hearing the robots, they were blindfolded and used headphones with white noise). Such sensorimotor stimulation includes motor, tactile and proprioceptive signals from the upper limb moving the front robot, and the tactile signals from the back robot. The task is divided into 2 conditions, the synchronous (no delay) and the asynchronous, where by introducing a short delay (500ms) between the participants' movement on the front and the robot's touch in the back. It is in the latter that is possible to robotically induce a presence hallucination (riPH).

A questionnaire was given after the robot manipulation in both conditions, in a randomized order across participants, to measure the feeling of presence ("I felt as if someone was standing close to me (next to me or behind me)"), along with other sensations as self-touch ("I felt as if I was touching my body"), passivity experience ("I felt as if I was touching someone else's body"), and loss of agency ("I felt as if I was not controlling my movements or actions"). In the syn-

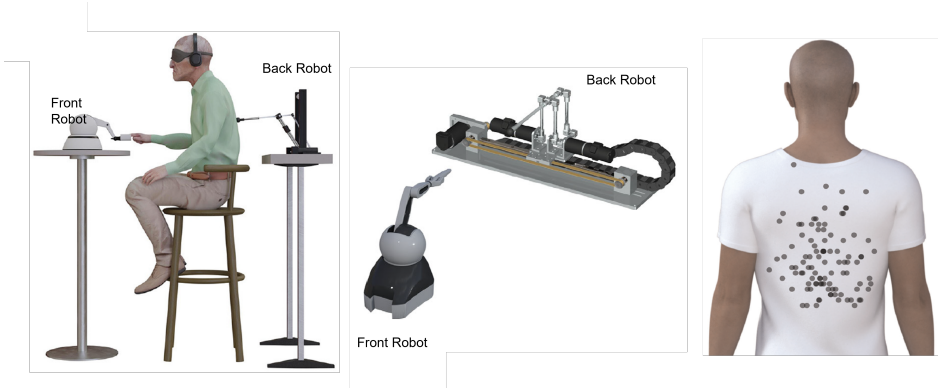

Figure 2.3: On the left, a model of how a participant interacts with the robot (Lateral view). On the center, a model of how the design of the robotic system. The participant moves the arm attached to the front robot, while the back robot mimics the movement, giving a tactile feedback on the back. On the right, an example of the back exploration done in this the behavioral experiment. Each dot indicates a point in which the back robot touched on the back of the participant [14, 15]

chronous condition (spacial conflict between the movement in front and touch feedback delivered on the back) participants did not experience riPH, but instead described control sensations (e.g., illusory self-touch), compatible with previous works on own-body perceptions. The experiment was recorded on video, for the researchers to validate the answers given by the participants and to log the performance of the participant while interacting with the robot.

Participants reported a higher Feeling of a Presence in the asynchronous condition (increased sensorimotor conflict, temporal and spacial, due to the 500ms delay) compared to the synchronous condition, as well as a lower self-touch sensation and a higher passivity experience. These behavioral results demonstrated that a disturbance in the sensorimotor processing of bodily signals on the upper back part of the trunk is sufficient to induce a Feeling of presence, strengthening previous hypotheses.

This robotically induced Feeling of a Presence can be explained by a disruption in the forward model [16] (Figure 2.4). This model hypothesizes that for each motor command, an efferent copy is generated and used to predict the outcome sensory feedback [43, 47, 46]. These sensory predictions are then compared to the actual sensory feedback. If these match, then the sensory feedback is self-attributed, and the predicted effects are canceled from the experience. On the other hand, if a mismatch is introduced by external influences, for example, the combination of a temporal conflict with the spatial conflict, a sufficient discrepancy between the expected sensory feedback and the actual sensory feedback is introduced in the forward model to break the association between the stimulus (touch on the back) and motor command (movement on the front). Hence, the tactile stimulation is perceived as being generated by an 'external agent' reported as a "presence".

More recently, this technology and the procedure were adapted to clinical research to investigate PH in patients with PD [14]. It is shown that patients with Parkinson's disease and preexisting symptomatic PH (PD–PH) have a higher sensitivity to the riPH method, compared with patients with Parkinson's disease, but without preexisting symptomatic PH (PD-nPH).

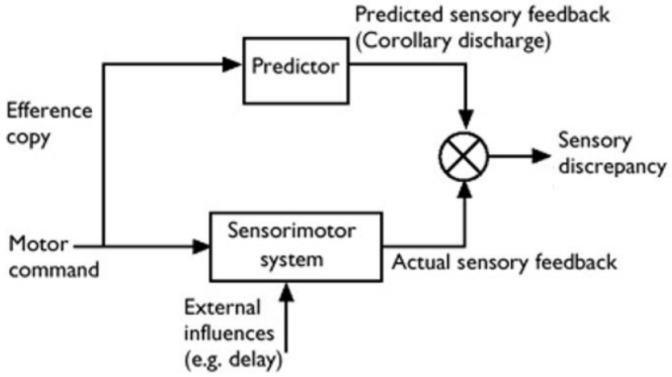

Figure 2.4: A model for determining the sensory consequences of a movement (adapted from [16]) . For each motor command, an efferent copy is generated and used to predict the outcome sensory feedback. These sensory feedback predictions are then compared to the actual sensory feedback, thus giving a sensory discrepancy that is used to attribute the origin of the sensory feedback (self-attributed or attributed to an 'external agent').

The clinical results in patients with PD (summarized in figure 2.5) revealed three major behavioral findings:

- The robotic technology is able to safely induce PH in patients with PD, especially in PD-PH patients
- PD-PH patients showed a stronger sensitivity (*i.e.*, more likely to report riPH) than PD-nPH
- 38% of PD-PH patients reported riPH that were comparable (or even stronger) to their symptomatic PH experienced in daily life

Some patients directly linked the experimentally controlled riPH with their preexisting experiences as symptomatic PH, because they present analogous attributes, for example, they appear in similar locations (on the same side) with similar intensity. These findings reveal that the conscious experience associated with riPH shares characteristics with the spontaneously occurring symptomatic PH [14].

Overall, the results demonstrated that specific sensorimotor conflicts, including bodily signals from the arm and trunk, are sufficient to induce mild to moderate riPH, linking PH to the misperception of the source and identity of sensorimotor signals of one's own body. The data also show that riPH shares similar phenomenological characteristics with symptomatic PH.

Moreover, riPHs were induced safely and repeatedly and were robotically controlled across experimental conditions, which has not been achieved before. This enables the repeated real-time assessment of a specific hallucination compared with standard assessments based on the retrospective evaluation of spontaneously occurring hallucinations, that may have occurred days (or months) before and only a limited number of times.

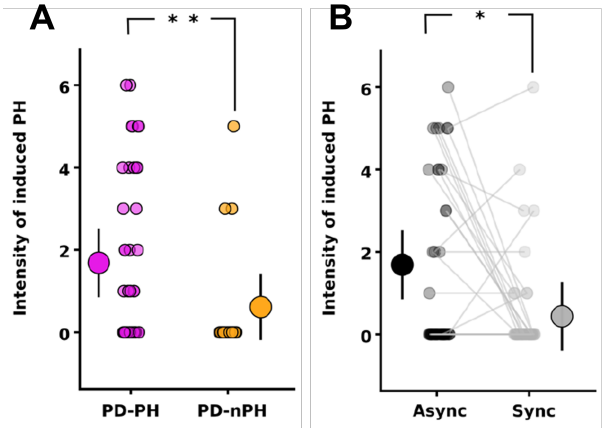

Figure 2.5: riPH in patients with PD. (A) riPH in patients with PD (asynchronous versus synchronous stimulation). Each dot indicates the individual rating of the intensity of the riPH [PD-PH (purple) and PD-nPH (orange)]. The dots with the bar on the left and right sides indicate the mixed-effects linear regression between PD-PH and PD-nPH. Error bar represents 95% confidence interval. **P $\leq$ 0.01. (B) riPH in patients with PD (asynchronous versus synchronous stimulation). Each dot indicates the individual rating of the intensity of the riPH. The dots with the bar on the left and right sides indicate the mixed-effects linear regression between asynchronous (black) and synchronous (gray) sensorimotor stimulation. Error bars represent 95% confidence interval. *P $\leq$ 0.05 [14].

2.3 Towards an Automatic Hallucinatory Analysis

Hallucinations are still unpredictable and private experiences, making their investigation, quantification and assessment highly challenging.

Through the robotic task, specific and short-lasting hallucinations are able to be induced, resembling clinical hallucinations, that can be elicited repeatedly and vary across controlled experimental conditions.

However, the methodology to evaluate them still remains based on the self-centered description of the participant, that may be exacerbated or soothed due to past experiences. Therefore, it is important to develop new procedures to analyze how different participants operate the robotic system, and consequently understand what is the typical motion and behavior of each of the main groups (HC, PD-nPH, PD-PH). This could pave the way for future development of new diagnostic and prognostic tools for disorders characterized by hallucinations, such as PD.

To achieve that goal, a motion capture technology needs to be implemented and used in combination with the robotic task to obtain a computer model of the movement trajectory done by each participant. From that data, motion features can then be computed to describe how the participants interact with the robot and how they react when being stimulated. From it, the behaviors in different groups can be compared, as well as in the different task conditions (synchronous *VS* asynchronous).

Thus, the next step is to realize what is the most appropriate tracking technology to be used in the riPH experiment and for a future potential applications in clinical settings.

Chapter 3

Tracking Technologies for the Clinical Sensorimotor Task

Visually, in the clinical sensorimotor task, participants belonging to different groups clearly reveal distinct reactions to the robot. For instance, a PD patient may start it's poking movements' pattern indistinguishable from a healthy individual but, with time, the conflicting stimulus may lead to tremor on the resting hand, or a slowness in the poking due to bradykinesia. However, the purpose of this project is to look beyond this typical symptoms and verify differences in behavior associated with the riPH task.

These changes in behavior when performing the task are fundamental to be documented and researched. To do so, the most appropriate technology is motion capture. However, different systems offer diverse advantages and setbacks that need to be taken into consideration according to the intended investigation, specially when studying PD patients, since these have a more debilitated posture amongst other limitations.

3.1 Selection of Motion Capture Technology

Motion Capture, also known as MoCap, is a process to digitally scan and record people's or objects' movements. The output of these systems allow to select certain points of interest with the purpose of studying the trajectories they perform in time and space.

Although most of the systems allow tracking inanimate objects, the most common target are humans. To replicate their movement, an array of keypoints is placed on the participants body joints and body parts of interest.The path performed by the multiple joints allows to build a simulation of the behaviour of the person or even to reproduce, in a digital environment, the motion performed with a 3D representation.This is an interesting and useful tool to deliver more natural-looking behaviours in skeleton rigging techniques used in animation applied in the entertainment industry, such as filmmaking or videogame.

MoCap systems can be divided into two main groups: non-optical, where the information is extracted using motion sensors; or optical systems, where the subject is filmed, with one or

multiple cameras, and its position is estimated based on markers or, in more recent versions, through markerless computer vision methods.

I. Non-Optical MoCap systems

The tracking of the human body needs to supplement information regarding the position and angle of the body joints to the computer models, in order for it to be recorded, and later processed, to give a clear estimate of the body stance. In these systems, the information is captured with the help of motion sensors that are attached directly to the joints. These can be of various kinds, such as, inertial, mechanical and magnetic.

i) Inertial Motion Sensors

This system uses gyroscopes combined with accelerometers to create Inertial Measuring Units (IMU). The role of gyroscopes is to gather information on 3 degrees of freedom, regarding the angular velocity of the movement tilt (pitch, roll and yaw, corresponding to the 3 axis of rotation X, Y and Z) and the absolute angles, when in comparison to an initial position. The accelerometer measures the linear acceleration of the device according to a single axis. Therefore, each IMU has 3 accelerometers placed orthogonally to each other, in order to measure the acceleration in the 3 axes (X, Y and Z) to offer an additional 3 degrees of freedom. However, the information given by a single IMU is ambiguous without a context.

Therefore, to get accurate measurements, there were developed software models of the human motion, in order to solve what is the best pre-defined distribution of sensors along the body, to then accurately assess the IMU data. In order to calibrate the trackers, according to the initial stance of the model, the person simply needs to stand still with their arms fully extended, either open laterally, in a T-pose, or resting by the side.

One of the most advanced motion tracking technology that solely focus on inertial sensors are the X-Sens. Some details can be found in table 3.1.

ii) Mechanical Motion Sensors

These type of systems are based on flexible electrogoniometer or strain gauges. This sensor is made of a flexible spring with plastic blocks on each end. The strain gauge mechanism is housed inside the spring, which changes its electrical resistance proportionally to the change in angle (*i.e.*,the force applied to bend the spring) between the plastic end blocks' longitudinal axes, offering a single degree of freedom.Although this is not enough to perform a full tracking, it provides a direct measurement of movement that can be combined with the other sensors, as a complementary data source for the tracking to be more exact.

This kind of sensors can be both wired or wireless, however they require to be tightly fitted to the body, which means that most of these systems are integrated in body suits, enabling the sensor to be close to the joint and in the correct place more easily.

A commercial product based on this mechanical motion sensing is the Gypsy7 exoskeleton, from MetaMotion. Some details can be found in table 3.1.

iii) Magnetic Motion Sensors

Finally, the third system, magnetic MoCap, work as marker-based tracking, where sensors are placed on the body to measure the low-frequency magnetic field generated by a transmitter source. The sensors and source are cabled to an electronic control unit that correlates their reported locations within the field, communicating them to the main software that associates those positions and rotations into 3D space. Most common systems have 6 to 11 sensors attached on the body joints.

This kind of system is not widely used since it suffers from a lot of shortcomings, such as, the readings being highly influenced when two or more sensors are close to each other, since they induce unexpected distortion patterns in the magnetic field; each sensor needs to be coated with a shielded cable, that leads to an encumbrance problem, since heavier sensors are more prone to slip while moving, a problem that also results in an offset when tracking the joints and, consequently, a repeated calibration process; and the recording is also affected by the interference caused by surrounding electronic systems, which means that the system can only be properly used in a controlled in environment.

These setbacks make the tracking seem too impractical and susceptible to noise, when compared to the technologies previously discussed.

As discussed, non-optical systems can offer a very accurate body tracking, based on different technologies and sensors, that need to be directly attached to the subject. An advantage that both inertial and mechanical systems have in common, is that they don't depend on the surrounding environment of the tracking object, *i.e.*, the system works independently of the lighting conditions, the reflectiveness of the surfaces around, the types of electronic equipment in the room, or even if it is done indoors or outdoors.

Finally, the main and obvious disadvantage of all the non-optical systems is the fundamental requirement to attach the sensors directly to the person. When studying the behavior of Parkinson patients is of the utmost importance that the tracking performed matches, as close as possible, the behaviour they elicit during their everyday life or, in case of diagnostic medical exams, such as the finger tapping exam, that the movement captured represents the true physical state of the patient. Hence, when devices or sensors are attached to the debilitated body of a PD patient, it is expected that these devices may disturb the movement in some way, leading to a non-realistic motion and therefore, not good data, which prevents to build reliable analysis tool.

In order to avert this problem, behavioral studies need systems that can track participants without the need of sensors. The solution are tracking systems based on video recordings.

II. Optical MoCap systems

Optical MoCap systems are tracking technologies that use one or multiple cameras surrounding the object(s) or subject(s) of interest, recording their movement in different angles, while maintaining a fixed position. The multiple viewpoints can then be concatenated to form a 3D reconstruction of the body, that can be subsequently analysed and studied. In case a single camera is used, getting a 3D visualization is only possible if a depth camera is used, such as, the RGBD Microsoft Kinect Azure. Otherwise, a single 2D camera can only give information on the frames that are being recorded, in a 2D space.

The tracking based on optical systems can be divided into 2 groups, the ones that need markers to be secured in the body and the ones that are markerless, where the body joints to be tracked are user defined after the recording, without the need to attached anything to the body.

i) Marker-based Optical systems

In these systems, the intended tracking points are equipped with markers that have special properties that can be identified by the surrounding cameras, recording their position, as the person moves. Optimally the cameras should be fixed around the periphery of the volume intended to track, at varying heights, so that each camera captures a distinct view of the volume. In each perspective, the 2D positions are calculated, and the overlapping position data are compared to compute the 3D location, via triangulation. For the tracking, the markers need to be mounted in predetermined places on the subjects' skin or clothing, to have information on the desired body parts.

The markers have different attributes depending on the type of system used and the type of movement that is relevant for the study. Thus, markers can be divided into 2 categories, namely, active markers and passive markers. The former have built-in LED inside that emit light of specific wavelengths, allowing for the cameras to recognize it. Each marker can emit light on a different frequency, making it very useful, since it creates a way for each tracker to be identified independently, being robust enough to not "confuse" the motion capture software, even if it goes momentarily out of the field of view of the camera. The downside for these systems, is that the lighting conditions surrounding the subject that is being tracked need to be adequate to not overexpose or obfuscate the trackers. For instance, infrared markers cannot be used outside, since the sunlight would create too much interference on the tracking. Nevertheless, this tracking technology has been developed and is present in many of the industrial motion recording systems, being the most known state-of-the-art system commercialized by Vicon. Some details can be found in table 3.1.

Regarding the passive markers, the source of light comes directly from the camera. When the light is reflected back, by the retroreflective coating on the marker, it allows to identify the position of the markers. On top of that, the multiple cameras take snapshots at an extremely high frequency, resulting in smooth video footage, that via the triangulation method, can identify the position of the markers in space. The first step to

perform to use a passive marker setup is to calibrate the system. To do so, markers are positioned orthogonally in a known position and orientation, both relevant to themselves and the cameras. This becomes the point of reference (origin) for the 3D reconstruction. A disadvantage of the passive marks is that they cannot be directly identified as different body parts, since the light emitted are all in the same frequency.

One of the most well-known companies that pioneered the MoCap field, using passive markers, largely towards research is Qualisys. Some details can be found in table 3.1.

ii) Markerless Optical systems

Markerless motion capture, as the name suggests, does not require any kind of marker or tracker to be attached, in order to track the body. Instead the system is based on computer vision and pattern recognition algorithms, to identify where is the position of the body, based on its silhouette or using reference points given *a posteriori*. If multiple cameras are used, a 3D estimate of the body can also be predicted, similarly to the technology used on the previous marker systems.

There are multiple advantages on using markerless systems, namely, the camera used doesn't need a high resolution and high frame rate in order to track the subject, working as good if it is recorded with a camera from the phone or webcam; the subject can use ordinary clothes, which allow to perform a more variate range of movements, that are more natural and comfortable for the subject without the risk of damaging any equipment, like sensors. If the algorithms are robust enough, even conditions like the lighting, background or if the subject is momentarily out of the FoV do not disturb the quality of the body tracking.

The leading software companies developing markerless tracking are Noldus, specialized on face tracking; and Theia3D, able to automatically predict the skeleton position in 3D space.

Opposite to all the previous cases, this technology is also available as open-source software, that derive from research institutions that study behavior or perform gait analysis. Two examples of this are OpenPose, from the Perceptual Computing Lab, in CMU; and DeepLabCut, from the Mathis Lab, in EPFL.Some details can be found in table 3.1.

Summarizing, optical systems can offer robust methods that prove to have a great accuracy, allowing to retrieve very precise data, despite that the subject still needs to put on markers for the cameras to identify the points of interest.

This results in a trade-off between excellent precision and accuracy *VS* the impact that the trackers and markers may have by being attached to the body. When studying the behavior or biomechanics of motor impaired people, is fundamental that the system doesn't influence their actions or the way they move, since in a diagnosis exam or in the sensorimotor experiment, the objective is to understand their natural response to the stimulus without being hindered by sensors or markers.

3.1.1 State-of-the-art Comparison

A market analysis was done, regarding the most modern products that implement the technologies described on the previous section 3.1, over various factors, in order to validate which is more appropriate for the clinical sensorimotor task.

Table 3.1: Comparison of state-of-the-art motion capture systems.

Specifications	**Motion Capture System**					
	XSens	Gypsy7	Vicon	Qualysis	OpenPose	DeepLabCut
Type	Non-Optical MoCap systems	Non-Optical MoCap systems	Optical MoCap systems	Optical MoCap systems	Optical MoCap systems	Optical MoCap systems
Sensor (tracker)	Inertial	Mechanical	-	-	-	-
Optical marker	-	-	Active	Seprate systems: Active or Passive	-	-
Calibration	Upright position with software assistance	Factory calibrated joint sensors	Extensive calibration dependant of camera system	Fast calibration wand waving with software assistance	No need	No need
Price Range	3 500 €	-	~17000€ (2 cameras + software)	~15000€ (430€ per active marker)	0 €	0 €
Camera specification	No Camera	No Camera	Specific camera to detect light from trackers	Specific camera to emit and detect Infrared light	General camera	General camera
Number of joints	17	14	Unlimited	Recommended 2-4 in each bodypart, but unlimited	Full Body - 25 Facial - 70	Unlimited, but need to be noticeable on the frame

An advantage that both inertial and mechanical systems have in common, is that they don't depend on the surrounding environment of the tracking object, *i.e.*, the system works independently of the lighting conditions, the reflectiveness of the surfaces around, the types of electronic equipment in the room, or even if it is done indoors or outdoors. This comes from the fact that the tracking doesn't rely on cameras, but only from the measures of the sensors. In consequence, the price of the X-Sens is more than 4x smaller than the Vicon and Qualisys.

These marker-based optical systems require specific and very hardware demanding setups, as the cameras possess a set of characteristics that are perfected for highly accurate measures (as described in the Appendix A.4). Another benefit of using cameras is that the subject can have as many keypoints as need, virtually unlimited, only restricted by the size of the subject. On inertial systems, the number of IMU's is reduced because of their unit price (the most basic IMU from X-Sens costs 115€, while a retroreflective marker from Qualisys costs 5.5€). Another disadvantage of the non-optical tracking system the hardware is wearable, which can easily experience mechanical wear or break down (since they are submitted to multiple movements, sweat, among other environmental conditions).

However, all of them have in common the fact they need to attach something to the subject. In respect of the experiments developed to induce presence hallucinations in PD patients, a markerless system is the best option, since in video recordings to log the experiments, the participants did not have any tracker placed on them, and were simply filmed from a only one angle. Besides, the system must be research oriented, allowing to get the raw position information, in a single 2D video, and offering tools to then analyze the results and get motion metrics. Another aspect that helps research is that the system should be open-source, so there is plenty of study cases based on it and, at the same time, it offers a low-cost solution to research, that is reproducible across the laboratories.

In the future, when applied to a diagnostic tool (for instance, in the analysis pipeline that derives from the robotic system) it leads to lower cost non-copyrighted technology, saving money both in development and licensing, and consequently in the final price of the product.

Thus, to achieve all this specifications, the 2 ideal solutions that fit this description are DeepLabCut and OpenPose. However, concerning the specific case of the robot induced hallucinations experiment, it is of interest to track both the participant and the front robot, since it allows to get features that further improve the analysis.

The OpenPose software is very flexible to track all kinds of postures and angles, but it is very specific for human applications and simply tracking the body. DeepLabCut (DLC), on the other hand, was designed to be as versatile as possible, allowing to generalize and identify any kind of pose, both human and animal, as well as, other objects present in the scene. This makes DeepLabCut ideal for the riPH type of application, since it is resourceful enough to identify the participant on the task, but also elements from the robot.

A proof of the adaptability of DeepLabCut is the common interest surfaced by the scientific community in this software, growing exponentially the number of publications, in a multitude of scenarios, over the last 6 years.

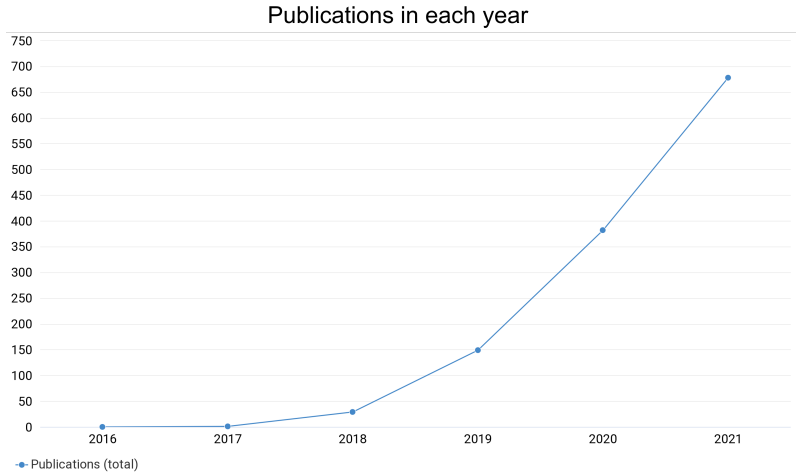

Figure 3.1: Number of scientific publications using DeepLabCut, over the last 6 years [1].

The increased number of publications is a good validation on the reliability of the software, further supporting the choice of being the suitable for this project.

3.2 DeepLabCut Software

DeepLabCut (DLC) is an open-source markerless pose estimation system, that uses computer vision to detect the geometrical configuration of multiple user-defined key points, without the need of adding markers *a priori* or motion capture (MoCap) trackers on the subject. In case of animal or human studies, this key points are the articulations of the body or some specific body part, like the eyes or the corners of the mouth. Since, all points are defined after the video is recorded, the system can be considered non-invasive and non-obtrusive, ideal to study natural movement, that is as similar as possible to the one elicited in natural/quotidian life. From all the published studies done with DeepLabCut in the Mathis Lab, with multiple collaborations worldwide, a library of pre-trained models, called "Zoo Models", was created to showcase the versatility of the system.

Taking a step back, to understand how it is possible to achieve accurate 2D motion capture is important to look into the function of DeepLabCut, the pipeline behind it and the steps needed to go from a new video recording to a generalized model that can track the moving key points and extract motion features from them. In picture 3.2 we have a representation of the main steps in a DeepLabCut workflow.

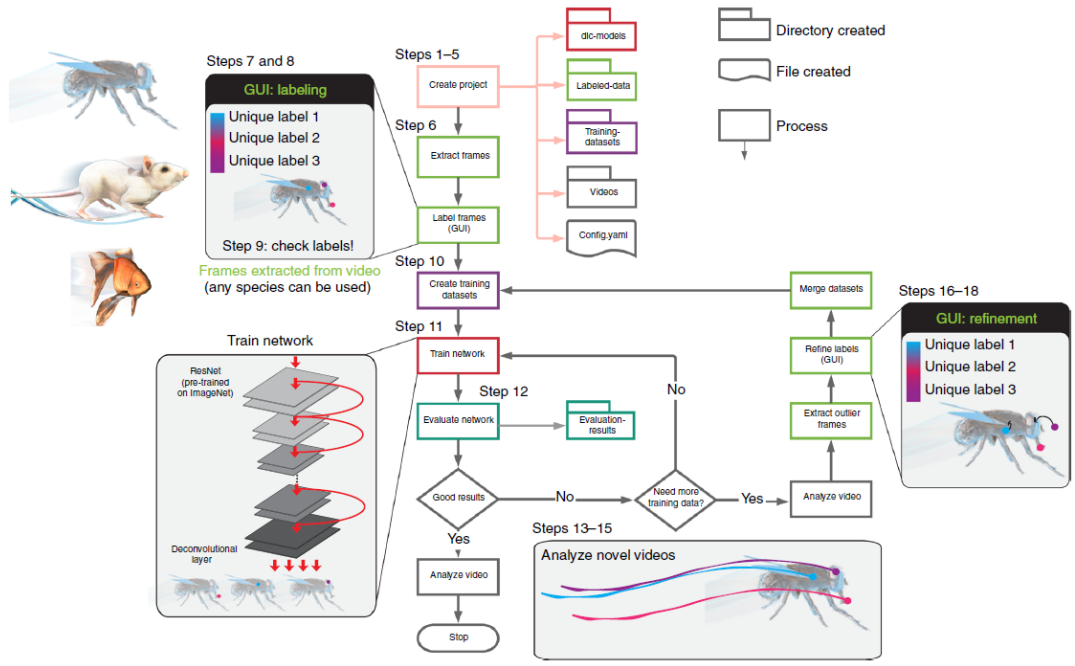

Figure 3.2: DeepLabCut workflow. The diagram delineates the workflow, as well as the directory and file structures. The process of the workflow is color coded to represent the locations of the output of each step or section [40].

DeepLabCut can be divided into two main categories according to the number of subjects that are intended to be tracked in the video. The first release of DLC (ref 2018) focused in detecting keypoints solely belonging to the same subject, therefore being called "Standard" or "Single Animal" (saDLC) project. In 2022, a new version of DLC was released, where for the first time, the models were able to identify, in the video, the same same key point in more than one subject, for example, identify the snout of multiple mice. This opened the door to perform more complex studies regarding the behavioural changes when multiple animals interact with each other. Hence, this new type of project is defined as "Multi-Animal" (maDLC). The workflow described is a summary of the main steps of DLC, however its methods are complex and require a precise fine tuning to be applied efficiently. Therefore, an in-depth explanation of all the steps is given in the Appendix A.5.

Upon deciding which type of project fits best the research purpose, the pipeline starts by selecting the training videos. This should be downsampled (according to the suggestion presented in the appendix), because high-resolution videos lead to very high training times later on (Appendix A.5.1). The complete videos won't be used for training, rather a certain number of frames are extracted from them. An important aspect is that the extracted frames need to reflect the variability of the behavior during the experiment, which means they should be as different as possible (for the model to generalize). This can either be done manually, or through one of the automatic algorithms (temporal uniform distributed way or by using K-means clustering). With the frames extracted, the next step is labeling the body parts of interest. DLC has very large receptive fields (due to deep neural networks), which means that various visual features or geometric layouts can be considered as keypoints, as long as they are salient and identifiable points. To label the images, DLC has a graphical-user interface (GUI) to open the frames and pinpoint where the body joint is located (Appendix A.5.2). At this stage, the training set is complete and ready to be inputted in one of the deep neural network architectures of DLC. The benchmarks and details regarding the training phase, such as, image augmentation, hyperparameters, models and loss functions are presented in the appendix A.5.3, as well as the differences between saDLC and maDLC on this step.

After training, the model is evaluated by computing the root mean square error (RMSE) between the groundtruth label and the prediction. This value is highly dependant on the size of the subject in the frame (if the individual fills a small percentage of the frame, even a low pixel RMSE can translate to a bad accuracy, since consecutive pixels may display different body parts, and *vice-versa*, if the subject fills almost all the frame, the same RMSE may indicate a high accuracy, because with-in the same cluster of pixels the body part is the same), as well as, resolution (Appendix A.5.3). Upon having a trained model that can identify the intended body parts with a relatively acceptable RMSE, is time to apply to unlabeled data. It's important to note that these could be the same videos from which the frames were extracted, since only a small set was selected and, in this phase, the complete video (all frames) are going to be analyzed. Even though the evaluation metrics of the model may be good, the prediction on new frames, may have a lot of noise (jitter) and miss-predictions (Appendix A.5.4).

DLC recommends multiple strategies to improve the model or the project, to enhance the predictions on the new videos. First, the simple extraction of new frames to increase the size of the training set. Secondly, DLC has outlier estimators to detect which of the predictions are indeed faulty. To use it, the noisy video previously analyzed is examined using the outlier algorithms to detect in which situations/movements the likelihood of the prediction is not satisfactory. Those frames can then be refined and corrected using the labeling GUI. Finally, the first training dataset and the new refined one are merged and ready for a new training iteration. Before starting, the weights of the model can be initialized using the first iteration, achieving better accuracies and less computational time to train. Returning to the issue of the size of the subject on frame, another solution can be a dynamic cropping of the subject, it appears zoomed in. Then a new project can be started, using the same labels. Like this, the RMSE has a lesser impact as explained before. All this methods are designated in DLC as active learning (Appendix A.5.5).

Once the tracking, *i.e.* the prediction of the location of the body parts is adequate, new videos can be analyzed using the DLC model. A coordinate for each keypoint is predicted for each frame of the video.

Based on the coordinate data files multiple packages were developed to help on the next steps, namely, for the identification of motion features, to help scale the analysis to multiple projects, to get information of the body part inside a specific ROI or machine learning models to do behavior analysis and classification.

All these methods emerge, as new projects are published, and as DLC applications continue to grow exponentially within the scientific community.

3.2.1 DeepLabCut in Parkinson's Disease

The DeepLabCut framework was firstly developed to study rodents and soon evolved to the study of behavior not only just in animals, but also in humans. Plenty of studies regarding gait analyses, posture or simply developing the most robust method to identify the human skeleton. However, considering the topic of the thesis, the state-of-the-art of DeepLabCut applied to the study of Parkinson's disease arises as a method to obtain quantifiable metrics that describe the behavior of patients, in an accessible, contactless way without the need for any extra devices.

Studies with this aim can already be found in the literature. The following example research case of (Williams et al. - [45]) investigates the symptom of bradykinesia and tests if DLC can effectively be a tool to measure the slowness of movement of PD patients.

In the study, the clinical assessment used to evaluate this symptom was the finger tapping test, where the neurologist observes the patient repeatedly tapping their index finger against the thumb "as quickly and as big as possible". This test can be scored through 2 different clinical scales — the Modified Bradykinesia Rating Scale (MBRS) and Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). In both scales the doctor needs to give a rating from 0 to 4, however in the MBRS the doctor evaluates 3 characteristics of the movement, the speed, the amplitude and the rhythm of 10 seconds of tapping; whereas in the MDS-UPDRS the score is for an overall performance of the patient in the first 10 taps.

The aim of the study is to measure the same 3 characteristics of movement using DeepLabCut on videos that were recorded with a smartphone. This brings a substantial advantage towards previous solutions that are based on wearable technologies, since most of the times, the fact that the PD patient needs to use sensors on the hands disrupts the true performance that the patient could have. Another setback of wearables is that it hinders the pathway to large-scale, routine clinical use, because they need to acquire an extra device.

In this research, the setup was very simple, the participant simply needed to rest their elbow on a chair arm with the forearm lifted to a 45 degree angle, while a camera placed on tripod in front was filming the participants' hand and forearm. A total of 137 videos were recorded and from each one 20 frames were selected to be labeled, using the K-means algorithm. Six joints were labeled in the hand as be seen in figure 3.3. The model based on the ResNet50 was trained for 1,030,000 iterations.

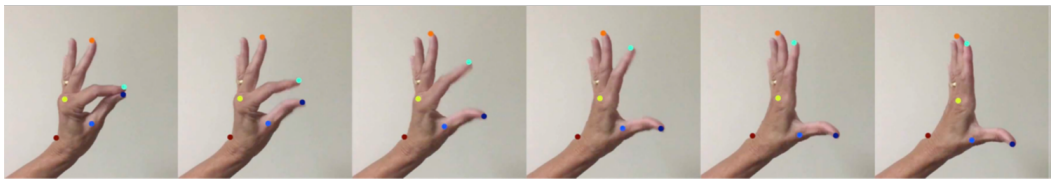

Figure 3.3: Example of labeled frames, extracted from the smartphone video.

After training and analysing the videos (full HD resolution – 1920x1080), the RMSE was of 8.39 pixels, a good accuracy and precision regarding the proximity of the hand to the camera. From the successful tracking, the trajectory of each keypoint could be analyzed and from that the 3 features of movement could be computed, in order to compared to the ratings given in the questionnaires.

To validate the relevance of the results measured with DLC, a Spearman correlation was computed between the ratings of the neurologists and the metrics of DLC, revealing a highly significant correlation between both scales. This means that the results achieved with DLC match the medical examination, supporting the veracity that this system can actually be a tool to assist doctors in their diagnostic procedures.

This proves that DeepLabCut is a very versatile tool, that can be applied to the study of Parkinson's disease, and that their behavior can be successfully tracked and measured. Moreover, DLC as markerless software is an ideal choice for the study of presence hallucinations, since no sensor or tracker should be applied, as it would hinder the true movement of the patient. Therefore, a similar logic to the one described, was applied in the study of the PH symptom.

At last, after discussing the main technologies in the market and their advantages and downsides, it is clear why DLC is the most appropriate technology to be used in my research context, to track the participants. With the DLC models it will be possible to analyze the videos of the experiment, in order to understand if there are significant changes in the movement and behavior, when comparing different groups of individuals, while performing sensorimotor task.

Chapter 4

Project Methods

The clinical sensorimotor task was recorded on video both to verify the performance of the participants when interacting with the robot and to keep a recording of their answers to the multiple questionnaires done after each condition. A full session, with all the conditions lasts around 1 hour, however the tasks performed with the robotic system may differ, depending on the study. Each condition lats 2 minutes, that can either be done synchronous (Sync) or asynchronous (Async). Therefore, the first step before applying DeepLabCut is to extract the 2-min segments that contain the behavioral conditions from the full recording (for each of the studies, in the next chapters, the specific conditions will be explained).

The complete methodology described in the next sections will be available online and open-source in `https://gitlab.epfl.ch/duteixei/msc-thesis-riPH-in-PD`.

4.1 DLC Pipeline for Clinical Recordings

Once the videos are properly extracted for each participant, the dataset should have at least one video in Sync and another Async, both with an approximate duration of 2 minutes.

Different systems were used to record the experiments, but the majority exports videos on High-Definition (HD) resolution (1280x720), with 30 or 60 frames-per-second (FPS). The first step to apply DLC is to reduce the resolution of the videos, to prevent the input of the neural networks, later on, from being too large. Multiple resolutions were tried to verify which was the best fit in the trade-off between lowest possible resolution, but with sufficient perceptible details (*i.e.*, discernible body parts). A constraint for DLC to work is that both the height and width of the frame need to be even numbers. The ideal downsampled resolution observed was 512x288 pixels, reducing 2.5x from HD resolution or, in case the videos is on full HD (1920x1080), the dimensions are reduced by reducing 3.75x.

4.1.1 Project Creation and Labeling

At this moment, a DLC project can be initialized. The "multi-animal" version was chosen, although only a single individual is present on the frame, since it contains more advanced methods

and models to predict the location of the body parts. The videos from which the frames are going to be extracted for labeling, need to be inputted at this stage. On the first study the videos from both conditions from all participants were selected, while on the second, only videos from the sync condition were used.

It is also at this stage that the keypoints of interest are defined. Since sometimes the angles of recording change, it is important to distinguish between the right (side 1) and left side (side 2) of the subject, specially in lateralized disease like PD. Therefore, 13 points of interest were defined, placed both on the participant and the robot, namely, finger1, wrist1, elbow1, shoulder1, finger2, wrist2, elbow2, shoulder2, chin, robot joint, robot top, robot base and the robot logo. During the analysis the reasoning behind tracking the front robot will be more clear. The last thing to be defined is the neural network architecture. The "*DLCRnet*$_5$0" was chosen due to its good performance in previous studies [30].

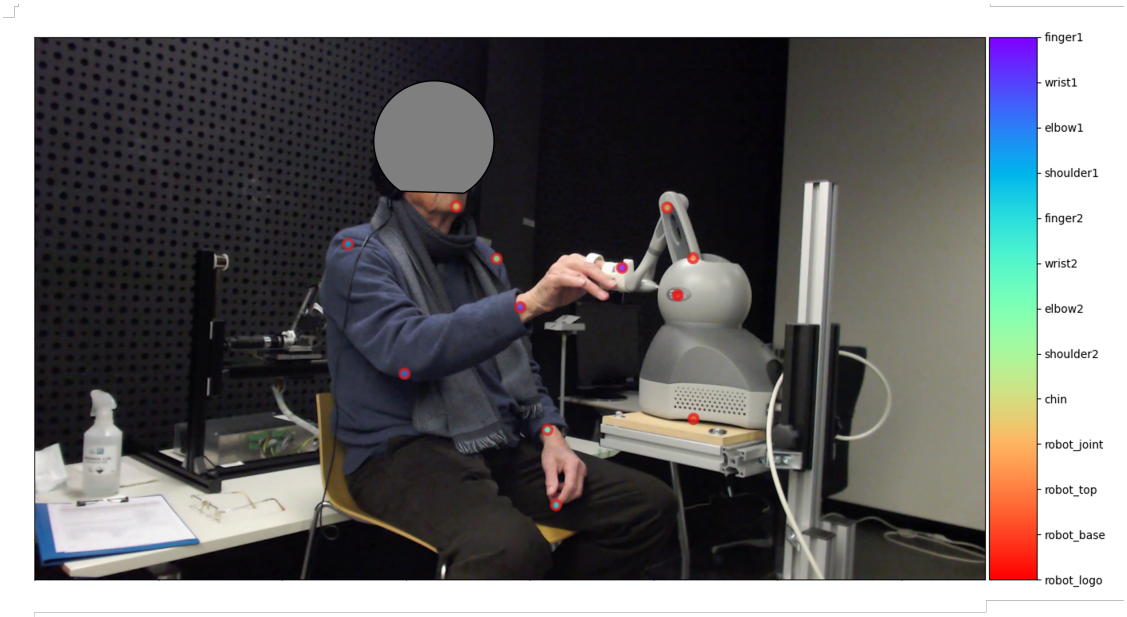

Figure 4.1: Example frame of the keypoints marked on the participant. Notice that the left elbow is not visible, so it is not labeled.

Moving on to the extraction, this was done using the K-means algorithm, since it helps to select frames with different stages of the movement, such as, arm fully extended or retracted (poking movement), or when a researcher intervenes in the task to give assistance. The uniform algorithm was also experimented since the task is based on the repetitive poking movement, but the results were not as good since extracting interesting frames is based on chance.

Regarding the number of frames to be selected the benchmark simply states that it should be enough to be descriptive of the movement in general, but also catch all of the less common scenarios. The published projects implemented on DLC have very different number of labeled frames. Depending on how complex the application is, the projects gather at least hundreds, if not thousands, of labeled frames. In each of the PH studies, the number of frames chosen were a trade-

off between spending a reasonable time in manual labeling and having the best possible accuracy in the model. The labeling was done manually in the GUI according to the recommendations of DLC, only labeling visible parts, as consistent as possible.

During the labeling process, one of the most usual problems among the video recordings is the inconsistent distance between the subject and the camera specially, in study B. This leads to an immense diversity of scenarios and inconsistencies when evaluating the performance of the model in new videos. Therefore, an adaptive cropping step is needed to standardize the database and to enhance the accuracy by focusing only on the region of interest where the participant is.

4.1.2 Adaptive Cropping Algorithms

In the pipeline, there are 2 methods to standardize the database, either a label-based cropping or a prediction-based cropping. In case, it is known *a priori* that the camera doesn't move or zooms-in during the experiment, then from the labels of each video, is possible to know what are the extremes keypoints on top and bottom, usually being, the chin and resting hand. However, if this can't be ensured, then a prediction-based method is needed, in order to know during the whole video what are the extreme points, either top-bottom or left-right. Only one of them is needed to calculate the cropping dimensions, since it is fundamental to keep the aspect ratio, to not introduce video distortions.

The first step in the general cropping pipeline is to know the position of the keypoints. The label-based method uses the labeled frames to check the coordinate of each body part, in order to find to find the maximum and minimum on both the X-axis and Y-axis.

4.1.2.1 Keypoint approximation model

For the prediction-based, as the name suggests, first we need to run model to have predictions in the original, uncropped videos. To do so, the labeled data is split between train and test (95% – 5%), as used in the DLC benchmark papers [40, 30]. The training dataset is then used to train a first DLC model. The specifications and parameters of this model will also be very similar to the one applied later on, to the cropped videos. Hence the details exposed now are valid for both. To run the model, some hyperparameters need to be defined. The batch size used was 1, since the GPU used didn't have enough memory to support more images being used at once. The number of total iterations ("*max_iter*") in the model was set to 500,000 as suggested in DLC guide for this batch [9]. The learning rate, evolves as default, starting with 0.0001 from 0 – 10,000, evolving to 0.00005 from 10,000 – 20,000, and finishing at 0.00001 form 20,000 – 500,000. Usually, 8 snapshots are saved during training, every 50,000 iterations [9].

After training, the evaluation metrics are computed on the 8 snapshot to check both for over-fitting and for which of them generalized best, *i.e.* has the lowest test RMSE. This snapshot is then used to perform the analysis of subsequent videos. However, at this stage, is not expected that the models have high values of accuracy, but they need to be good enough to provide a general idea of where the participant is in the frame. All the videos intended to be cropped can now be analyzed.

As a result of the analysis a coordinate file is created, indicating the position of each body part in each of the frames, with an associated likelihood value of how certain the prediction is.

4.1.2.2 Cropping pipeline

At this stage, independently on the kind of data, the adaptive cropping algorithm can be done. This is concordant with the active learning strategy recommended by DLC guide [40, 9].

As explained, the adaptive cropping, goes through the keypoints in the video and selects which one is the extreme point on the top, bottom, left and right. In the prediction-based method, these values need to be supported by a likelihood higher than 90%, to make sure the extremes points are not miss predictions.

When this coordinates have been found, independently of the method, it is possible to get the height of the frame (a little margin of 10 pixels is added to the top and bottom, to ensure that the extreme body parts are fully intact in the video), and consequently based on the original aspect ratio, it is possible to know the width. By centering the width, on the mid point between the extreme points left and right, we have a bounding box that surrounds the region of interest (ROI) to be tracked (participant + front robot).

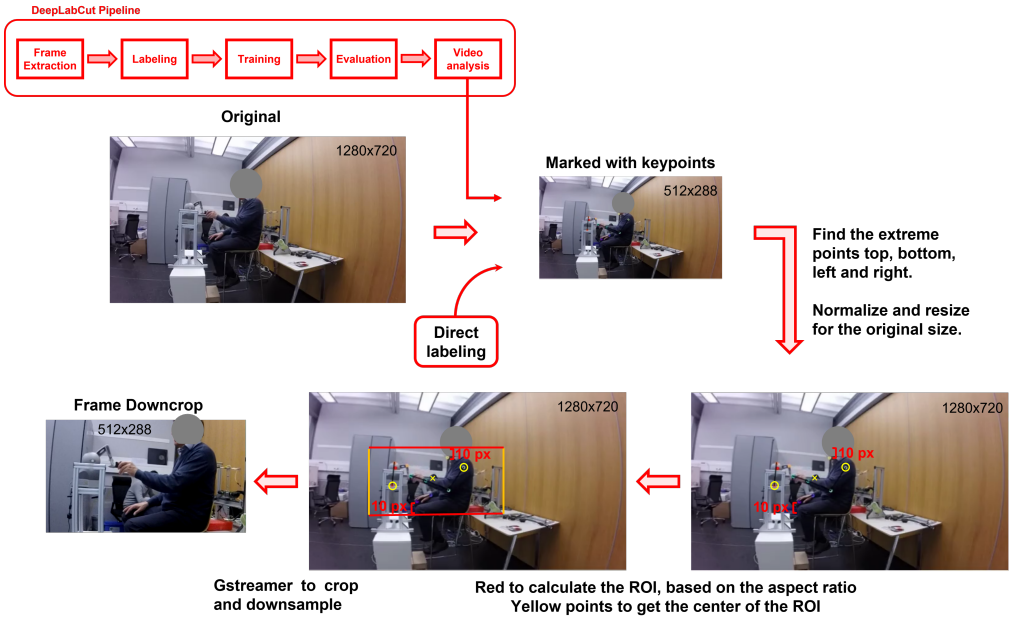

Figure 4.2: Cropping pipeline to emphasize the region of interest (ROI), by removing unnecessary background, resulting in a zoom in effect of the participant.

However, if the database is very diverse between participants, these can be positioned in all manners around the frame. This leads to some cases where the estimation of the initial bounding box may be incorrect. Taking this into consideration, a set of solutions were implemented to rectify and attend to these cases, as is shown in figure 4.3.

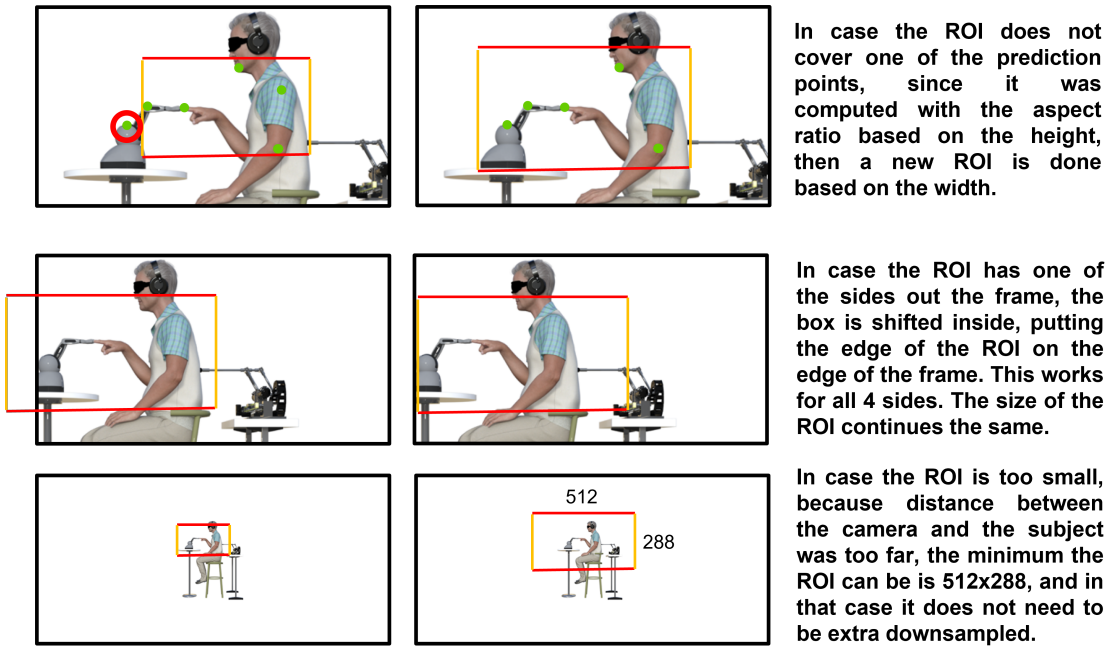

Figure 4.3: Example of the failure cases of the cropping algorithm. The top row encompasses the case where one of the predictions is out of the ROI. The middle row represent the case when the ROI is too big and goes out of frame. In the bottom row, the participant is so far away, that the ROI is smaller than the expected final resolution. In those cases, the ROI is set to be 512x288, to avert the need of interpolating pixels, trying to upsample the image.

The objective with this corrections is to get the ideal bounding box, that is as close as possible to the final resolution (512x288), while keeping the whole participant visible, without compromising the aspect ratio. The solutions implemented in the above figure solve most of the problems. However, when trying to correct one of them, another may happen, for example, if a case in the first row happens, a new ROI is computed; however that one may be outside of the frame now. Therefore, is important to verify all conditions every time again the ROI is adapted. The final ROI is achieved when it respects all the conditions at the same time. There are some exceptions, for example, when the participant is fully removed from the camera view, that an optimal ROI cannot be computed. In this rare cases, the best solution is not to crop the video, maintaining the participant in frame for as much as possible.

The cropping coordinates are in the downsampled resolution, so they were normalized (scale form 0-1) and then multiplied by the high resolution of the original video, meaning they are in the same relative position in the original frame. In this way, the cropping method will zoom in from a high resolution to a size closer to the final downsampled resolution. This is why it is important to keep the aspect ratio constant while cropping, so that the cropping area can easily converted to the final resolution - 512x288.

The GStreamer software (an optimized library to handle media content, enabling GPU computing in decoding and encoding of videos) was used to automatize the cropping of the fullHD original video and the encoding of the new final video. This new type of videos were denominated

as "DownCrop" (downsample + crop).

Important to note that, if the "DownCrop" videos and their respective previously extracted frames want to be used to create a new project, then the labels previously done do not match the position of the body joints. However, with the coordinates of the bounding box, it is possible to normalize the position of the labels inside it, as is explained in figure 4.4. Therefore, after the cropping process, the labels' coordinates can be multiplied by the downsampled resolution, putting them in the correct places.

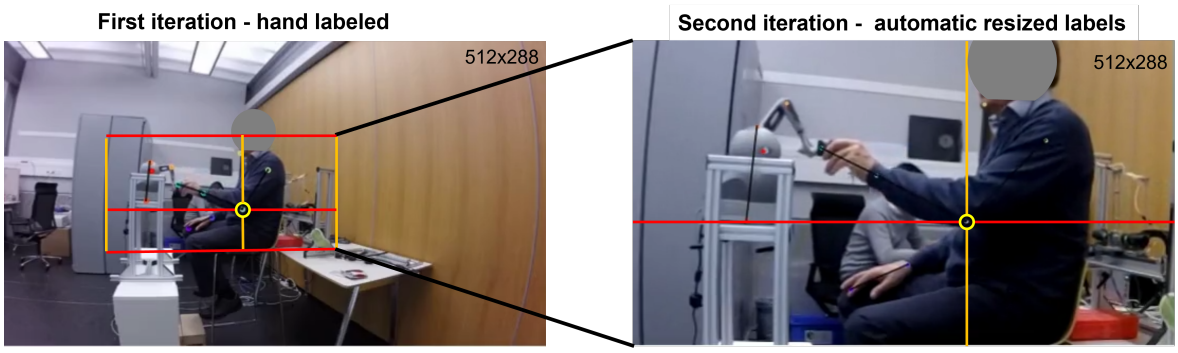

Figure 4.4: The normalization inside the ROI, allows to get the location of the labeled point in the final "DownCrop" videos.

In the "DownCrop" videos, the images are much more standardized, since the participant is always centered in the image, the differences in background are attenuated (since it is not as visible) and, most importantly, the distance between the camera and the subject is much similar. This is relevant, because the jitter has much less influence, as each body parts occupies a bigger portion of the video, as shown in figure 4.5.

Figure 4.5: The highlighted line shows acceptable distance for a prediction to still be considered on the wrist, with a good accuracy. On the DonwCrop version, even if the predictions have jitter, it does not influence as much the accuracy, since it has a larger region within each body part.

4.1.2.3 Keypoint final prediction model

After the cropping process, the data is ready to be used to train the new model. An advantage of prediction-base cropping is that transfer learning can be applied from the weights of the first

iteration training phase. Independently of the methods, the split and hyperparameters were the same as previously described.

While analysing the loss evolution, although it starts to plateau in the initial set of iterations (usually before 100,000), the maDLC guide recommends to train until 500,00 iterations, for a batch size of 1 [9]. Therefore, when selecting the snapshot is important to consider one closer to this limit, rather than closer to the beginning of the loss stabilization. As a rule of thumb, the best model is the one with the lowest test error, after the 300,000 iteration. Each study has details that are specific to them, and will be described in their respective chapters.

The evaluation metrics also need to be substantially better than in the first run. Both the train and test RMSE should be lower on average, meaning the model is more accurate and generalized better to the data. The fact that the video is "DownCrop", also means that the participant is zoomed in. This leads to more accurate predictions (since the body parts are expanded, the keypoint is more easily predicted in the correct position, as seen in figure 4.5), as well as better precision (since the data is more standardized and the model is more robust, jitter is also reduced in the predictions).

From the analysis results a coordinate file per video, from which the behavior of the participant can be studied, by computing motion features that describe his reaction to the sensorimotor task.

4.2 Extraction of Behavioral Features

Once the video is tracked it is possible to understand how the keypoints of the different body parts move during the video. As mentioned, the videos are acquired as a log of the experimental task, being captured in a profile/lateral view of the participant (example in figure 4.5). As such, the X axis of the frame represents the movement back and forwards when moving the front robot, and the Y represents the up and down movement.

For example, if the movement of the tip of the index finger that is attached to the robot is analyzed in the X direction, an harmonic pattern appears, that represents the repetitive poking done by the participants.

4.2.1 Database Standardization

When performing an analysis across participants, is important to make sure that the distances computed are always the same, according to a population standard. The cropping of the videos helps to improve the uniform look, but that procedure is not enough to guarantee that the pixel distance is comparable according participants, since it depends on the original distance from the camera, the angle that it is being filmed from, among other factors.

In order to achieve a true standardization, the distances need to be normalized by a factor that is present in all the videos and never changes size across all the participants. The best choice to achieve this is the front robot, therefore the importance of tracking its top and base, in order to get its height (difference between the robot base Y coordinate and robot top Y coordinate). Since the robot height is known, it's possible to assume that the height in pixels is equivalent to the real

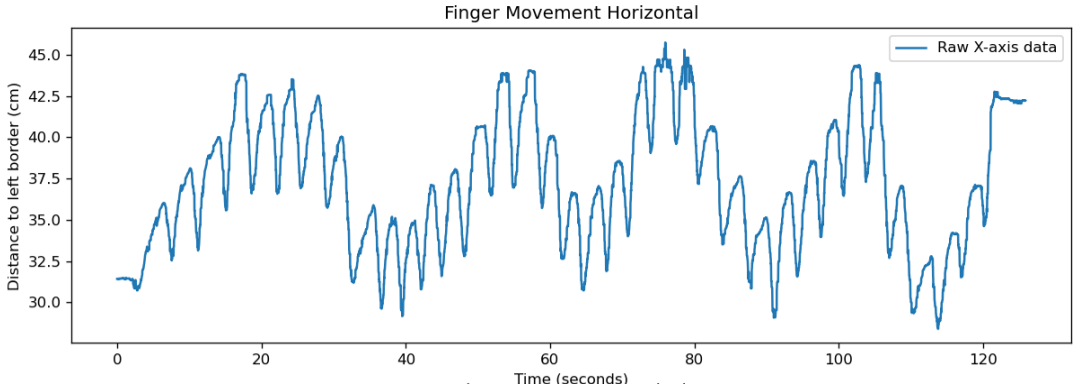

Figure 4.6: Plot showing the movement of the index, front and back. The Y scale is the distance of the finger from the left border and the participant is oriented to the right. This means that he moves his finger forward on the local *maxima* of the signal. Therefore, wavy pattern represents the repetitive pokes done by the participant. The following pipeline tries to build a solution to clean this signal, preparing it to extract motion features that are not affected by noise.

height of the object, equal to 22 cm. Hence, in each video the coordinates of all body parts are normalized based on the robot height, making the measure of each movement comparable across the whole database. Another benefit is that the measures become in a unit close to centimeters, making the data more comprehensible.

To make sure this estimation is viable in a video, a "sanity check" is performed where the variance of the prediction of the top and the base of the robot, should not be more than 3 pixels (considered to be the inter human variability [40]). This test guarantees that the prediction does not move/burst too much at any given point, validating the robot height assumption made before.

At this point, after checking the data and standardizing the prediction coordinates, it is possible to start to extract the actual features of movement. When inspecting the experiment, most of the changes over time are linked to the arm doing the poking. Therefore, as the behavioral pattern changes with poking motion, the trajectory of the joints of the moving arm are the ones of most interest to study, namely, the wrist, elbow and shoulder; as well as the tip of the finger. Hence, the motion features are going to be computed focusing on these keypoints.

The pipeline to calculate features is the same to all body parts, being divided in 2 parts, starting with data cleaning and filtering and then the implementation of the algorithms for the feature calculation.

4.2.2 Data Cleaning and Filtering

Cleaning the data before extracting any feature is fundamental to guarantee that the feature is not being influenced by noise, leading to a false description of the movement. In the first step, the 2 minutes recording is shortened, since the beginning and end of the videos sometimes do not have useful data. As can be seen in figure 4.6, the extremes of the signal do not show the same behavior as the bulk of the signal. This happens because the participant is simply waiting for the task to start

and, in the end, resting the hand on the table or lap after finishing. Hence, two 7-second segments are trimmed, both from the beginning and the end, in all participants.

The second step is to evaluate if a body part is being predicted or not. Sometimes, in the beginning or during the video, the research occludes the participant to assist him on the task. On these moments, the body part is not tracked, leaving "holes" on the signal. Most of the features to be calculated, require a full continuous signal, therefore being necessary to correct these missing segments. In case they appear in the beginning, the missing signal is trimmed out.However, if it appears in the middle of the task, the missing data takes the value of the last recorded point. This makes sense empirically, since the research would have stopped the participant to explain how to properly do the movement, so the arm is assumed to stay stationary.

The third step is related to bursts between consecutive predictions, where the signal has sudden spikes. This may happen due to people moving in the background, lighting conditions that do not allow to identify the arm clearly, having the researcher passing in front of the camera momentarily or the camera itself being moved. A curious note when these bursts happen is that the signal starts in the baseline, then it bursts to an off-set (positive or negative), but the tracking done on that off-set is still describing the movement of the participant, and finally, a few seconds later, it may come back to the original baseline. Hence, an algorithm was develop to clean these miss predictions.

It is based on a sliding window of 280 samples with an overlap of 50 samples (values defined after multiple trials with different windows sizes and overlaps). The mean inside the window is computed and, if 2 consecutive samples have a "jump" bigger than 25% of the mean, then a spike is detected and the difference between this samples is saved (if the spike is a peak, the difference is negative and, if the spike is a valley, the difference is positive).

$$Burst_{diff} = S_i - S_{i+1} \tag{4.1}$$

The sliding window continues to go through the signal and, if another spike is detected in the opposite direction, it means that the keypoint went back to the baseline position, thus, to all the samples between the spikes, the previous difference ($burst_diff$) is added, shifting them back to the baseline. This algorithm works even if the burst is a spike (only 1 sample) or if the second spike overshoots past the baseline (2 spikes in 1).

Lastly, looking back to the signal example of figure 4.6, it's noticeable a slight high frequency noise in the signal. In the tracking result, this is translated in a small jitter of the point around the area of interest. The smaller the jitter the more precise the prediction is, hence, ideally, a perfect tracking would be a smooth curve. Neither of the previous cleaning steps target this problem. The best solution for this is to apply a low pass filter that still includes the principal behaviors (principal low frequency harmonic), but erases all the jitter. To achieve this a butterworth filter was selected, since it has the flattest frequency response, *i.e.* the one that introduces the least temporal distortion/phase shift in the signal. The filter has an order of 5, as a trade-off between a steep cut-off without affecting the phase response. To define the cut-off frequency of the low pass filter, the power spectral density (PSD) of the signal was computed, in order to reveal the most

relevant frequency. Although the signal has a clear peak, it is usually surrounded by peaks with less power, but that also affect the signal shape. Therefore, to the PSD maximum (main frequency that the participant does the poking) a margin of 1.5 Hz is added to encompass the other small frequencies that describe other aspects of the signal.

$$Cutoff_{freq} = PSD_{peak} + 1.5 \tag{4.2}$$

As a result, a clean filtered signal is achieved, that describes the behavior in a smooth continuous way.

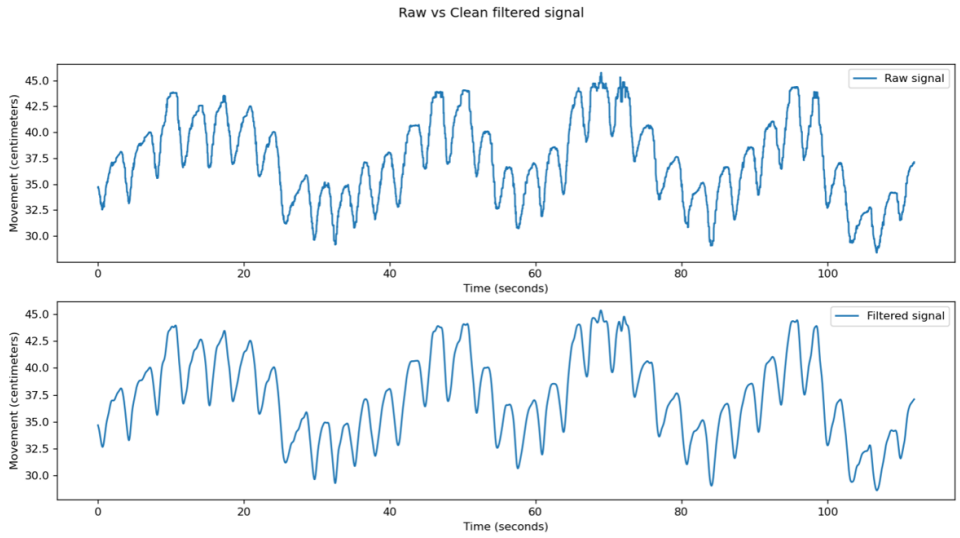

Figure 4.7: Plot of the movement before and after the data cleaning and filtering steps.

This pipeline is applied on the X axis and Y axis to extract multiple motion features.

4.2.3 Feature Computation

Motion features are intended to be quantifiable metrics to describe the overall behavior of a participant. That can be translated as the combination of the individual patterns and attributes that each body part or joint manifests. Thus, the same set of features were computed for the 4 keypoints of interest.

To start, the most basic two features are the mean and standard deviation of the front and back movement (X axis). This is the direction of interest, because in the finger it translates to the poking movement and in the shoulder, for example.

The speed of movement was also extracted, based on the filtered signals, taking in consideration both vectors X and Y. To do it, a package called DLC2kinematics was used, to calculate automatically the speed in every sample frame. From this results a signal, from which the mean and standard deviation can also be extracted. As the coordinate values were standardized by the

height of the robot, the result comes in a measure approximate to centimeters per second. The acceleration was also attempted, but to apply a second derivative to the signal made it indiscernible from noise and spikes, so no feature valuable was extracted.

Regarding the filtered signal, these were the 4 features extracted. Another curious feature is the amplitude of poking. However, this cannot be calculated in the simple filtered signal, since usually the movement has a baseline wander. This can be seen in example figure 4.6, as the pokes are influenced by a underlying very low frequency. This is due to the fact that sometimes the participant does the poking movement closer to the chest and other times closer to the front robot. Therefore, to calculate the amplitude, the filtered movement needs to be flattened. To do so, a the signal of the baseline wander is needed. To compute it, another low pass filter was applied to the signal, with the same filter model and order as before. However, the cutoff this time is a very low frequency (0.1 Hz), to eliminate all the poking and jitter. From it results simply the baseline wander that can be subtracted from the filtered signal, originating a flat version, with the pokes centered in zero.

To calculate the mean amplitude, the first step is to find all the peaks and valleys in the signal, that are interpreted as the position most to the front (close to the front robot) and most to the back (close to the chest), respectively. With all the peaks and valleys, is possible to compute their average values. The difference between these values is the mean amplitude of the signal, that can interpreted as the average amplitude of the poking movement. Besides the average, another possible interesting feature is the max amplitude achieved by the participant during the task, which was calculated as the difference between the highest peak and lowest valley. The spike cleaning step is specially important for this feature, as it would highly influence the values of the local *maxima*.

When the flattened signal is calculated specifically in the moving finger, it is also possible to determine the frequency of pokes during the experiment. The zero crossings are computed and then halved (because to have a poke, the participant needs to pass by the midpoint going front and back), to count the number of pokes. This is divided by the total time (total number of samples divided by the sampling frequency), to obtain the frequency of poking.

$$Frequency of Pokes = \frac{\frac{1}{2} ZeroCrossings}{N/F_S} \tag{4.3}$$

The final two features are not regarding the movement, but instead the quality of the original signal, before removing the spikes and filtering. The point of these features is to have metrics for the jitter to assess the tracking quality and precision of the motion tracking.

Following a similar reasoning as before, if a clean signal is achieved with a low pass filter, then if the same cutoff is used, but instead in a high pass filter, the resulting signal should be mostly the noise induced by the jitter. From this signal, the maximum amplitude is calculated to know what was the biggest disturbance induced by the jitter.

The bad signal quality, shown in figure 4.8, for the shoulder of the resting arm is a common effect seen in all the joints that are farther from the camera or that do not show any ample

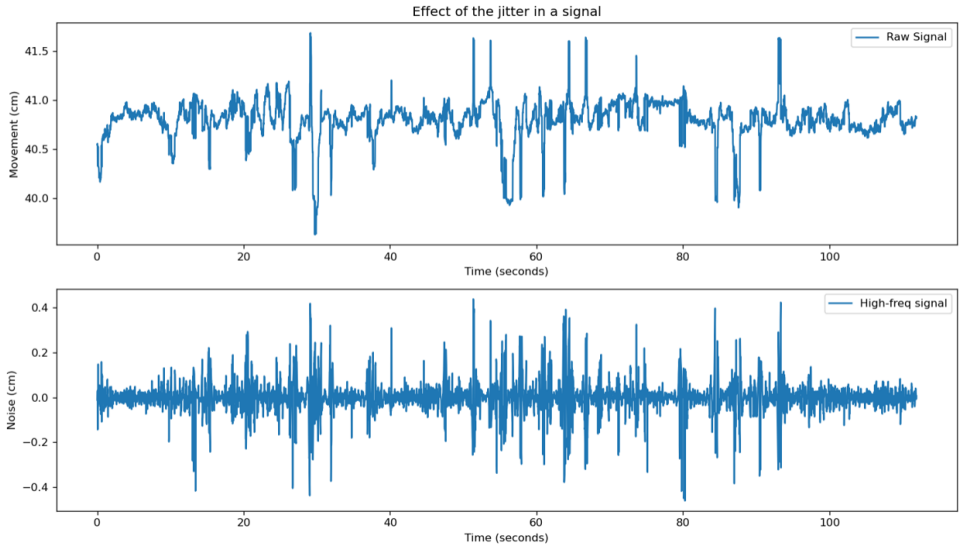

Figure 4.8: On top, plot of the raw movement of the shoulder of the non-moving arm (same participant as plotted before). The movement is distorted by the noise making it impossible to distinguish what is noise or an actual behavior. On the bottom, the high frequency noise that contaminates the signal. The amplitude of the noise and the raw signal is very similar, indicating that the tracking was not successful.

movements. Hence, the body parts of the resting arm, as well as, the chin, were considered to be excessively affected by the jitter noise.

The second feature is a measure of the jitter in the overall signal. This can be achieved by computing the sum of the absolute differences, sample by sample, between the original noisy signal and the low pass filtered one. Once this value is then divided by the sample size to get an average value (to eventually be comparable in case different signals have different frame rates). This precision metric is still experimental, and was not used to evaluate if the movement signal is usable.

Overall, 8 features are computed for all the body parts from which 6 are related to motion and 2 are to assess the signal quality. In case of the finger, an extra feature is computed regarding the frequency of poking.

The motion information gathered from each participant is organized according to their group and condition to perform a statistical evaluation. As the objective is to understand how the movement differs between the groups, a 2-Way ANOVA test was performed, taking the motion feature as the dependent variable of two fixed factors, the group and condition (Sync or Async). If a p-value lower than 0.05 is observed in a fixed effect, a statistical difference can be assumed, revealing a main effect influencing the motion feature.

The methods described were applied to 2 independent studies, the first as the proof of concept that the pipeline is able to analyze the sensorimotor task properly to determine differences between groups of participants; and the second explores the limitations of motion capture, by implementing the pipeline in a study with more diverse conditions.

Chapter 5

Study A — Controlled Condition Experiment

Upon establishing a clear link between the sensorimotor discrepancy triggered by the robotic system and the induction of presence hallucinations, numerous participants started to be tested to further consolidate the knowledge on PH and the neuroscientific, cognitive, and behavioral paradigm behind them.

5.1 Database

Figure 5.1: Example frames of the log videos of different participants, in the database of study A.

On the first study, the intent of the database is to be a structured set of videos filmed in a

controlled environment, to promote a homogeneous look across all participants. The videos logging the experiment, were all filmed in the same room, with the cameras set in the same position for all the participants. The camera is pointing to the participant together with the front robot, at approximately the same distance and angle to capture the same perspective of the task every time. Although the frame of reference is the same, the camera changes sides according to which hand the participant is using (depending on the dominant hand). The dominant hand of most of the participants is the right hand, establishing this as the standard frame of reference, as can be seen in figure 5.1. In case there is a left handed participant, and the viewpoint changes, the frame will be reversed, *i.e.* the front robot is now on the left side of the frame instead of the right. To accommodate for this, the video needs to be flipped on the y axis, inverting the left to the right. This technique standardizes the look of the video, as well as the signals obtained from tracking.

The camera used was always the same, meaning that the aspect ratio and resolution of the videos is consistent (16:9 - 1920x1080).

All these factors are very important, because inter-subject variability is decreased and possible elements that may introduce noise in the DLC tracking performance are eliminated. If this wasn't the case, the model would need more training data to generalize to different scenarios, setup conditions or even technical video differences.

As mentioned in the methodology, the log of each experiment is 1 hour long, so the first step is to understand what are the relevant parts of the behavioral task, that are of interest to be analyzed.

5.1.1 Experimental Protocol

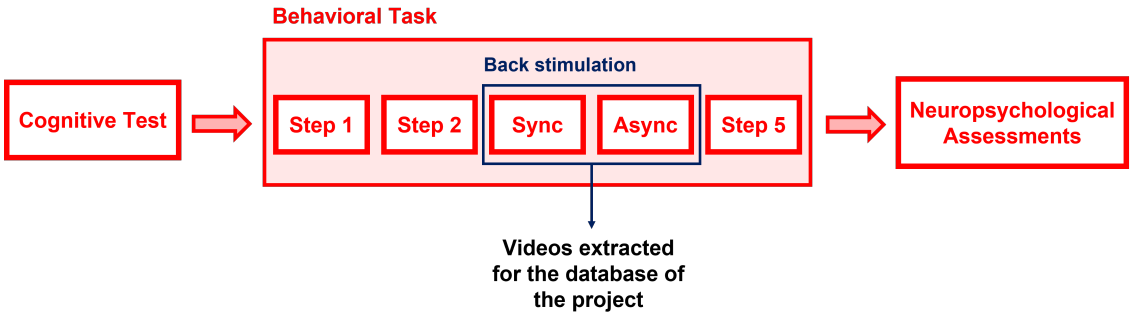

Figure 5.2: General experimental protocol for a participant in study A. In the behavioral task, the back stimulation is randomized, not necessarily being steps 3 and 4. Inside this block, the async and sync conditions are also presented in a random order, between participants. Each condition lasts 2 minutes. The total behavioral task lasts 1 hour, because the participants are interviewed between blocks.

The protocol of the behavioral task (*i.e.* the participant interacting with the robotic system), on this first study, is composed of 5 steps, varying the body part being stimulated by the robot and the synchronicity of the stimulus. From the overall task, the 2 conditions of interest are the ones where the robot is touching the participant's back, one in a synchronous condition, with the movement performed on the front being replicated immediately on the back, and the other with

a 500 ms delay, in an asynchronous condition. Both of them are 2 minutes long periods. These trials appear in random order.

From the total video of the behavioral task, these 2 conditions are the ones extracted, resulting in 2 videos per participant, identified by the subject ID code, condition executed and hand used.

Besides the behavioral task, the study also encompasses other examinations, that despite not being the focus of this work are worth mentioning. The researchers and clinical team perform cognitive, neuropsychological and psychiatric questionnaires, to asses the health and mental state of the participant. From this data, sociodemographic comparisons can be computed taking into account the overall population.

5.1.2 Sociodemographics

As, the objective is to determine the differences in movement between groups of participants, healthy controls (HC) *VS* PD patients and, within the latter, between the ones that have symptomatic PH (PD-PH) *VS* the ones that do not (PD-nPH), the comparison between populations are also based on this distinction. In the study were collected videos of 32 participants in total. 4 of these participants were excluded from analysis as they did not fit the inclusion criteria in the neuropsychological assessments, resulting in a total of 28 participants analysed.

Table 5.1: Sociodemographic comparison between HC and PD

	HC (N=10)	PD (N=18)	p
Gender	7 (M)	15 (M)	0.410 (χ2)
Age (years)	70.30 ± 7.06	68.72 ± 7.29	0.584
MoCA (scale)	27.78 ± 1.78	26.21 ± 2.23	0.077

The p-value of the gender was obtain with a chi-squared test, while the other factors come from an independent student t-test.

From the table 6.1, it is possible to extract some information regarding the study population. Regarding gender, in both groups there is a majority of males, specially in the PD group, reaching 83%. The age factor is very close between the groups, which is fundamental for the healthy participants to be considered viable controls. Lastly, the MoCA assessment measures the cognitive decline of a participant. In this scale, a score of 26 or higher means the person is normal and cognitively capable. Both groups have on average values above 26, but as expected the PD is slightly lower. This is an important verification to make sure the participants can comprehend the task and perform it by following specific guidelines.

Focusing now on the 18 participants that have PD, extra assessments were taken into consideration, to confirm that the potential differences in movement, observed in the behavioral task, do not originate from external factors. Therefore, it is important to ascertain that the overall patient's pattern of motion is not significantly affected by factors like a longer disease duration or a more severe motor impairment, but instead they are due to the robotic stimulation.

Table 5.2: Sociodemographic comparison between PD-PH and PD-nPH

	PD-PH (N=7)	PD-nPH (N=11)	p
Gender	5 (M)	10 (M)	0.210 (χ2)
Age (years)	70.15±7.38	69.00±7.88	0.669
PD Duration (years)	6.92±4.38	8.19±4.16	0.404
UPDRS III (scale)	14.84±8.44	17.53±2.50	0.281
MoCA (scale)	26.14 ± 1.95	26.250 ± 2.45	0.923
LEDD (mg)	1324.13 ± 689.90	871.97 ± 461.76	0.125

The p-value of the gender was obtain with a chi-squared test, while the other factors come from an independent student t-test

From table 6.2, the gender and age distributions remain similar to the ones on first table. Regarding the PD duration, the PD-PH group have have a lower duration. This goes according to the paradigm that PH, and minor hallucinations in general, surface usually as a symptom in an early phase. PD-nPH participants may not have this kind of hallucinations in the first place or they might have evolved to more severe ones, like visual hallucinations, as the disease progresses. This progression may also lead to more prominent motor impairments. This is shown in the results of the UPDRS III scale for PD-nPH, the higher the score the more impaired the patient is. Looking at the cognitive decline, they are similar and both groups are in average cognitively capable. Lastly, the levodopa equivalent daily dose (LEDD) is much lower in the patients without PH, although they have the disease for a smaller period of time. This could be a speculation for the fact that the PH is a sign of an overall worsening of the patients' condition, and therefore is treated with a higher drug dosage.

At this stage, it can be confirmed that any of the factors mentioned present a significant difference between the groups, therefore the distinctions observed on the movements are not explained by any of these aspects.

The database is then composed of 28 participants that have 2 videos each, resulting in 56 videos to be analyzed by the DLC pipeline.

5.2 Motion tracking Specifications and Model Performance

The homogeneous look and constant recording setup simplifies the pipeline, as a label-based adaptive cropping can be applied, because the cameras do not move or change positions throughout each condition. However, there are still sources that highly influence the appearance of the video and that may affect the tracking performance. For instance, the experiment sometimes is done with different lighting conditions, which reduces the visibility of the participant and makes the identification of the body parts a harder task. Another factor that may not seem obvious is the color of the clothing used by the participants. This is a source of noise, since the model needs to interpret that different participants have different color in the same keypoints, specially on the elbow and shoulder. A similar problem was previously reported in [35] with a database of horses with different colored fur. Finally, the fact that the participant is sporadically occluded by the researcher, in case some assistance is needed during the task, brings a big source of miss predictions (in spite, these scenarios being taken into account when selecting and labeling the frames).

Therefore, to ensure that the model recognizes every participant in every condition (sync and async), all the videos that are intended to be analyzed in the end, were inputted in the initial stage to extract frames from them. The 4 participants that were excluded, although not being target of the analysis, were still used in the labeling stage, to include more data in the training set, to generalize the models as much as possible.

As referenced in the methods 4.1, the number of labels extracted is highly dependent on the type of project and videos. The only constraint suggested in [40] is that usually more than 500 frames are needed to get results close to human accuracy. Therefore, as this project is intended to be a proof of concept, is preferably to have a high number of frames that is surely able to converge into accurate predictions. For that, it was estimated that 25 frames would be enough data to describe each video and its different scenarios. The frames were labeled and follow the pipeline (cropping and label relocation) described in the methods. An important note is that the data is balanced, since all the body parts of interest are visible in the all the subjects, which means that each keypoint trains on very similar amounts of data.

In total, 1587 frames were used to train the DLC model (13 frames are missing, since the K-means extraction method could not find, in some of the videos, the 25 different clusters).The training process follows the guidelines described in the methods.

According to figure 5.3, the loss starts to plateau around the $40,000^{t}h$ iteration, although the model only truly starts to get accurate enough in new videos after 300,000 iterations. Therefore, each model is trained until a max of 500,000 iterations as recommend in the DLC guidelines [9] for this batch size. Over time, 8 snapshots are saved and then compared between one another, to find the best one according to the lowest test error, after 300,000 iterations. On this case, looking to the image on the right, the ideal model is the $450,000^{t}h$ snapshot.

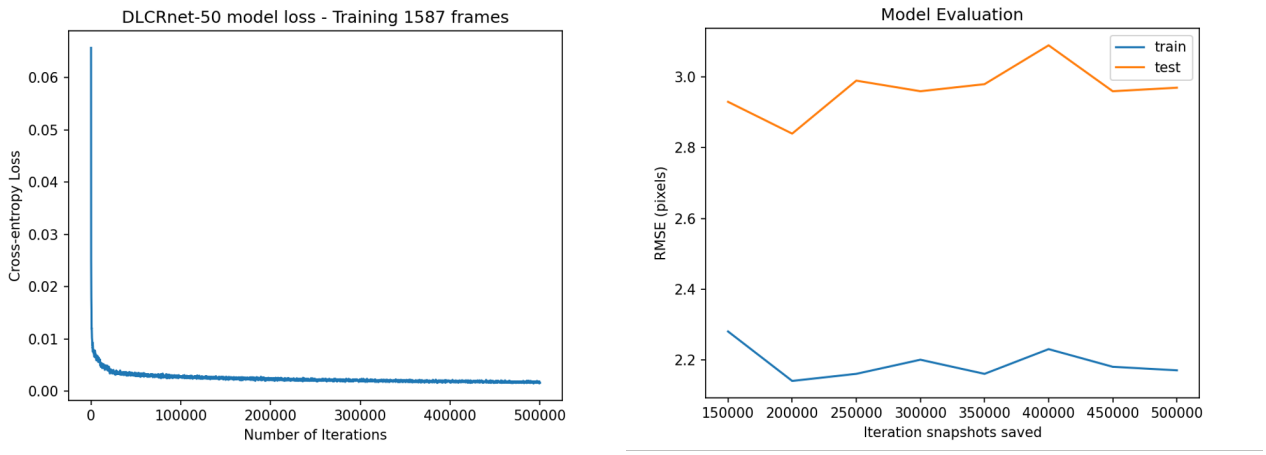

Figure 5.3: On the left, the plot of the evolution of the cross entropy loss. On the left, the comparison of the train and test error (RMSE), for the 8 snapshots saved during the training process.

Based on this model, all the videos from the 28 participants were analyzed. Upon visual inspection of the the tracking of the videos, the keypoints are accurate and can follow the body parts while they perform big movements, for example the tip of the finger while the participant does the poking. However, DeepLabCut was not developed to track very small movements, in the range of millimeters. The jitter influencing the tracking makes the movement of certain body parts absolutely imperceptible, such as the chin or the shoulder from the non-moving side. For this reason, the motion features are mainly extracted from the moving arm. The fact that the camera viewpoint is also focused on this side helps on the tracking performance.

From each of the videos of them, the motion features were extracted. On the next step, these behavior descriptors will be compared to find any significant distinction between groups or the experimental conditions.

5.3 Results

The statistical tests have in consideration both the group and the condition as fixed effects. Using the 2-way ANOVA it is possible to realize if the motion feature data possesses any differences that might be inherent to the fixed effect. Since, the objective is to compare 2 sets of groups (HC/PD and PD-PH/PD-nPH), they will be discussed together, separated by the analysis of body parts. To start, one of the points of most interest is the finger attached to the front robot, since it is the body part that moves the most, and consequently is expected to exhibit most differences. The ANOVA is robust enough to account for the different population size in each group.

The features analyzed to compare participants, based on the moving finger were the frequency of pokes, in Hertz (Hz), the amplitude of the movement front and back (poking), measured based on the height of the robot causing the units to be similar to centimeters, and lastly the speed, taking both X and Y vectors of movement into consideration, in a unit, once again, close to centimeters per second. The mean of the movement was not valid feature to be used, since it is an absolute

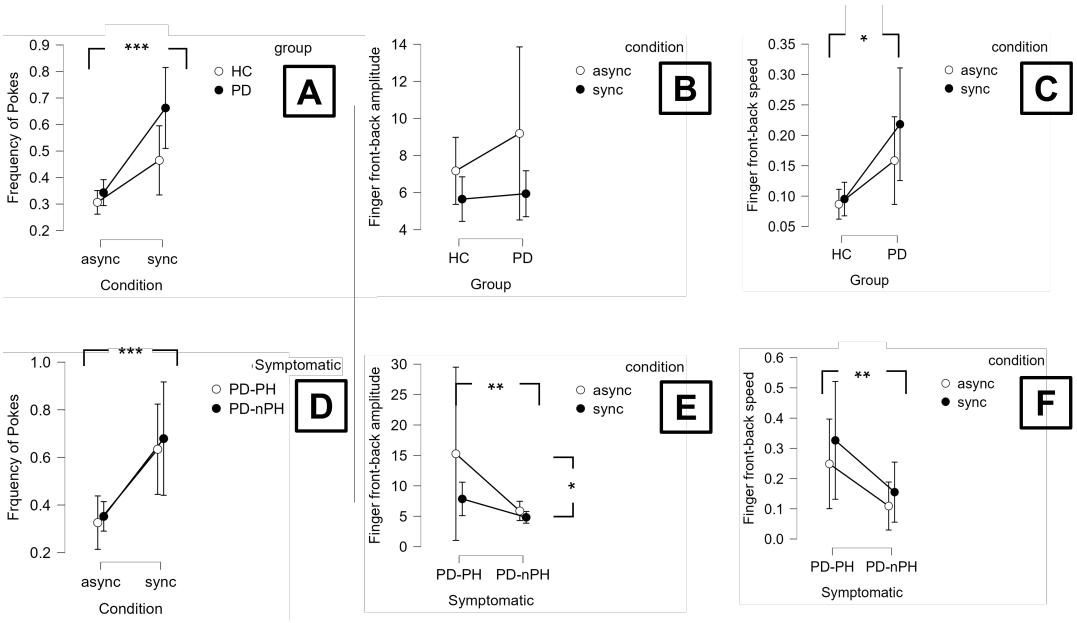

Figure 5.4: Plots of the statistical results of the motion features of the moving finger. The top row is the HC *VS* PD (A to C) and the bottom the PD-PH *VS* PD-nPH (D-F). From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.1.

measure within the frame, and not all the participants are guaranteed to move their arms in the same frame regions, which makes it incorrect to be compared.

Starting on the left, both plots of HC *VS* PD and PD-PH *VS* PD-nPH are computed having the vertical separation being the condition, since there is a strong main effect of synchrony, which means the poking frequency is different between conditions. It's clear that in async, the participants tend to have a much a lower frequency, around 0.35 Hz. On plot A, healthy controls seem to be more consistent, as there frequency does not change as much as PD. Moving on to the other features, the plots are now divided vertically according to the group. On B and C, the speed is significantly different between the groups, as PD patients tend to be faster, on average, specially in sync. Inspecting the last 2 plots, E and F, some interesting results are observed. Regarding the amplitude (plot E) both effects are significant, stronger between groups, as patients with sPH perform higher amplitude movements, on average, specially in async. The PD-nPH are very consistent, since there amplitude remains centered in 5 centimeters, with a low variability interval, on both conditions. The speed once again is statistically different between groups, higher in PD-PH. Interesting to mention that both type of patients respond on the same manner (speed wise) to the synchrony change (similar slope). When looking to both plots (E and F) at the same time is also interesting to see that in async the amplitude is much higher, while their average speed is slower.

However, the model allows to perform the same analysis on other body parts to further evaluate the differences and characteristics of each type of patient. Since the movements of the wrist are usually highly correlated with the finger, only the shoulder and elbow keypoints were taken into

consideration in the analysis.

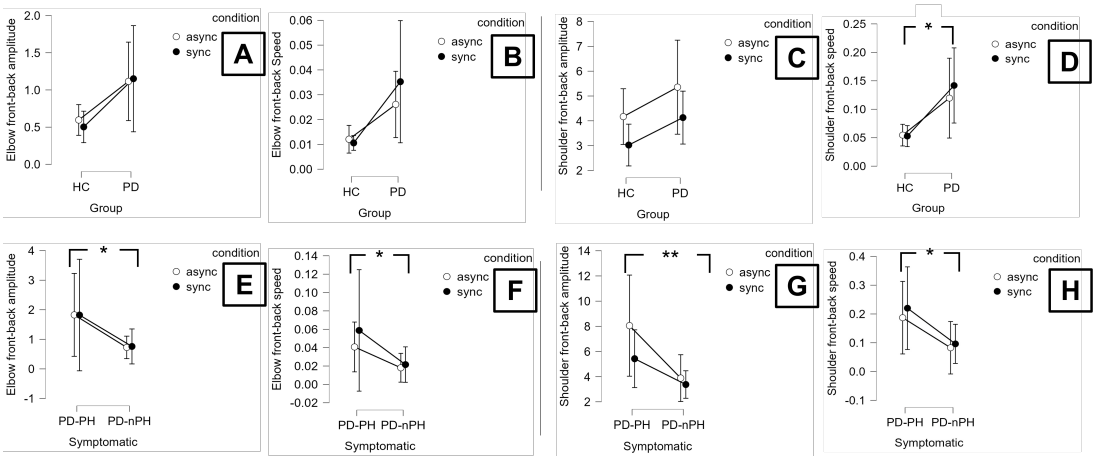

Figure 5.5: Plots of the statistical results of the motion features of the elbow (left) and shoulder (right). The top row is the HC *VS* PD (A to D) and the bottom the PD-PH *VS* PD-nPH (E-H). The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.2.

Starting with the plots A and B, although the averages of the the groups seem clearly different, both the amplitude and speed have scales extremely small, showing that the elbow does not move much during the task, in the front-back axis. Also, the PD patients present a big variance in their motion, having patients ranging as low as the average HC behavior, so no main effect is observed. Between conditions, the movement of the body parts is also very similar in each group. When analysing the shoulder, the amplitude of moving forwards and backwards (plot C), although it's not a main effect, is different between conditions, where the participants tend to move more in async. The only significant difference is on the speed of the shoulder when comparing groups, as the HC do not move much, their speed is also not very high, in opposite of the PD that present a faster movement (with a less variance).

Looking to the plots of the bottom row, all the plots have a main effect on group distinction, which means that the behavior between PD-PH and PD-nPH is significantly different and can be distinguished successfully with these measures. On plot E, the average of amplitude inside each group, in different conditions, is approximately the same despite the variance. The PD-nPH barely move the elbow, with a result very close to 0. Taking this into account, when looking to the speed, it is also very close to zero. Like this, the small movements of the PD-PH helps to differentiate and achieve the main effect on the group. Looking to the shoulder, in plot G, the scale of the amplitude almost doubles when compared to the plot C, specially as PD-PH in async reveals a big amplitude of motion back and forward. However this is not conveyed to the PD-nPH, as the amplitude is fairly low and consistent, in both conditions. Concerning speed, the synchrony affects both groups in the same way, but once again, the speed of PD-PH patients is significantly faster than PD-nPH.

From these results, multiple assumptions can be established to explain the behavioral patterns of the different group according to the robotic tasks condition.

5.4 Discussion

The results observed allow to infer some patterns and typical behaviours specific to each main effect observed.

Starting by the strong effect of synchrony in the frequency of pokes, in the async condition the participants usually get confused by the delay of the sensorimotor stimulation. They try to adapt their movement to create a false sense of synchrony, using strategies such as stopping or slowing down when they extend their arms, waiting for the delayed poke of the back robot. These small periods where the participant waits, lead to a decrease in the frequency, independently of the population, justifying the significance in the plots A and D, of fig 6.4.

In the methods, the figure 4.6 shows the movement of a participant in the async condition. In the small waves, as the signal goes up the participant is extending his arm, doing the poke. As can be seen the peak of the wave is larger, showing that the participant is slowing down, waiting for the poke. However, when the pulling the arm back, the valley of the signal is more sharp/pointy showing that it is done faster, according to the task instructions. This is the typical behavior in response to the temporal conflict.

Another method to follow this strategy is to perform longer movements (bigger amplitude) but slower, to make sure that a new movement towards the front is not done before a poking has been delivered.

However it does not impact as much the speed of the the actual movement back and front of the finger, as can be seen that the PD participants move faster than HC. In the movement plot, this speed is the slope between each local peak and valley. This main effect of the PD patients being faster than the HC, can be due to the fact that they may not be able to control their movements precisely, due to the disease or to the stimulation. This precludes them from performing the task according to the exact instructions given in the beginning by the researchers. In comparison, the healthy participant can follow clearly the guidelines indicated, such as the adequate speed, frequency, area explored, among other factors.

Regarding the main effect of the shoulder speed, in plot D of figure 6.5, the PD patients move faster than HC, which seems counter-intuitive due to symptoms like bradykinesia or muscle rigidity. However, when taking the shoulder into consideration, is important to think that the back robot is poking the participant. Due to the weak posture and rigid stance of the patients, they do not have enough precise control of the front, sometimes pushing forwards more than needed. This may lead to the poke done in the back to be slightly stronger than normal. Due to their weakness and this hard pokes, the patient does not have enough strength to counter the force applied on by the back robot and is pushed forward (increasing their speed). This explains why PD amplitude and speed are so high. The HC can manage the front robot and prevent hard pokes, or in case they

happen, they have the strength to hold against them, remaining in the same in position. This is a possible explanation for the the main effect in plot D fig. 6.5.

Comparing the plot E from fig. 6.4 analysing the amplitude of the finger and the plot G, in fig. 6.5, analysing the amplitude of the shoulder, it is curious to look at the difference between these amplitudes. If the difference is applied between the amplitude of the finger and the shoulder, in the sync condition, the finger just moves around 1 or 2 cm more than the shoulder. From this, it can be assumed that the poking performed is done with the body, being more of a rocking motion back and forward, with the whole torso. In the context of PD, this reaction can be explained by the motor rigidity, which may prevent PD patients to do arm extensions easily.

In the async, the PD-nPH maintain this rocking motion, but the PD-PH, although the amplitudes of both body parts increases a lot, the difference is higher for the finger, meaning they extend the arm more. This can be due to the delay conflict that triggers a different behavior, allowing them to start to move their arms more freely, on average. The big variance in plot E, for the finger, shows how different patients may manifest different responses in this stimulus. As a main effect is observed, it can expected that this reaction is inherent to patients with symptomatic PH, since they may be more sensible to the robot. Besides moving it more freely, they also do it more rapidly, as a main effect is observed in plot F.

The delta of 1 and 2 cm mentioned, is also corroborated by the elbow amplitude, plot front and backward in PD, once again supporting the .

If the task is performed properly, the elbow is not supposed to move much in the X axis, as the arm extension to do the poking is not very wide. This is confirmed by the very small range in HC, on plot A fig. 6.5. However, in PD, the elbow amplitude is around 1-1.5 cm corroborating the theory of the torso swinging motion, because if the elbow is moving, it means the participant is rocking front and back instead of just extending the arm. When looking to plot E, the main effect surfaces again between groups, as the elbow swings much more, getting to an amplitude of 2 cm on average in PD-PH, since the patients are also extending their arms excessively more, as previously seen.

This alteration in behavior is significant enough to implicate all the features of movement, revealing a clear difference between the groups.

5.5 Final Remarks

In summary, the purpose of study A was to validate that DeepLabCut, as a tracking technology could be used to capture the movements of the participants. For that, the acquired database is very controlled, keeping the camera viewpoint, the distance from the robot, among other aspects, standard for all participants. Based on these log videos of the sensorimotor task, the DLC pipeline was implemented, achieving good accuracy results on identifying the position of the body parts. Although the jitter reduces the overall precision, rendering imperceptible the movements of some body parts, the poking movement is clear enough to be filtered by the methods developed. The body joints of this arm were analyzed to extract the motion descriptors.

When analysing and comparing the motion features between groups and task conditions, some behaviours can be speculated. Concerning the poking frequency, it changes significantly between sync and async, since the participants try to compensate the delay with their movement, to achieve a false sense of synchrony. Regarding the distinction between groups, the healthy participants usually maintain the same posture, with consistent behavior (same amplitude of poking and speed) across trials, following the task guidelines. PD patients, on the other hand, due to their rigidity and overall weakness, when are poked by the back robot tend to be pushed to the front. This results in a behavior where the patient does the poke using the whole torso as one, doing a rocking motion back and forward. Inside this group, the PD-PH reveals to be more sensible to the task, since their behavior, most probably due to the induction of PH, is greatly affected. They are shown to be more unimpeded, presenting on average higher amplitudes of motion and arm extension, however being consistently slower in the async condition.

The fact that typical patterns and reactions observed in the task can be verified with the system, as well as, being able to discern new behaviours that are shown to be significantly different between the populations, validates the efficacy and success of the system. However, all the data was acquired under very controlled conditions, that usually cannot be replicated in real life case scenarios. Therefore, the point of study B is to apply the pipeline on videos with very diverse setups and conditions, to verify if the system can achieve the same results. This step is necessary to establish if the methods are robust enough to be applied in clinical translation, in a product or system that would be integrated in different medical environments.

Chapter 6

Study B — Close to Real Condition Experiment

The conclusions achieved in study A, validate the system as proof of concept, however it is fundamental to grasp its limitations. The objective of study B, is to understand if the method can be applied in more diverse cases, that mimic scenarios similar to what real life applications of the robotic system would look like. The study will be successful if the motion features achieve similar results to the ones in study A.

6.1 Database

Figure 6.1: Example frames of the log videos of different participants, in the database of study B.

In this study, the videos of the database were recorded without any predefined consensus, being the full responsibility of the research to log the experiment as seen appropriate or more convenient at that time and place. This of course leads to a very high inter-subject variability in the manner of how the videos are recorded, originating plenty irregularities. When compared to study A, the same sources of inconsistency are still present, such as opposite lighting conditions in environments with plenty of sunlight to almost total darkness; the different colors of the clothes the participants use; and the fact that sometimes the researchers need to intervene to correct the participant.

However, as can be seen in figure 6.1, the videos present a multitude of other different aspects. Starting by the relative position between the camera and the participant, both the recording angle and the distance of filming is never the same, changing the perspective in every participant. Another source of noise is the variable background, since the videos for this study were collected in different locations (in the Campus Biotech, not always in the same room, and in the Sion Hospital). In the background, most of the times, there are researchers moving and people moving, that will introduce noise, since the model could predict their body parts instead of the subject's. Besides this changing environmental conditions, there are also technical issues that surface, since the cameras used to record vary across locations. This results in some of the videos having different resolutions and even different aspect ratios (there cases of videos with 4:3, instead of the typical 16:9). Another inconvenience that is a repercussion of the logging being done as most convenient, is that the camera sometimes moves during the experiment itself. Moving the camera is problematic, because if in the middle of the task the camera gets closer or further away, it changes the standardization applied to the data, because a pixel filmed from far translates into a longer distance in centimeters. The same reasoning can be applied in case the camera is moved and the angle changes, all the computations may also end up wrong or even inverted, if it goes from one side to the other, of the participant.

All these aspects are sources of conflict when trying to compute a model using DeepLabCut. In most of the studies they tend to be uniform like in study A, or not even mentioned. Practically, this means that the training data needs enough information regarding each of these scenarios, for the model to generalize appropriately.

In this study, the log of each experiment is 1 hour long, so the first step is to understand what are the relevant parts of the behavioral task, that are of interest to be analyzed.

6.1.1 Experimental Protocol

The behavioral task in this study also has 5 steps as in study A, however only the first 4 fit the 2-min archetype of back stimulation, that has been used in this project. In this case, the participant experiences the sync and async condition, however it does the poking movement with both the right and left hand instead of just one.

Therefore, from the total video of the behavioral task, these 4 conditions are the ones extracted, resulting ideally in 4 videos per participant, identified by the subject ID code, condition executed and hand used. In this database, some of the experiments only has some of the conditions saved,

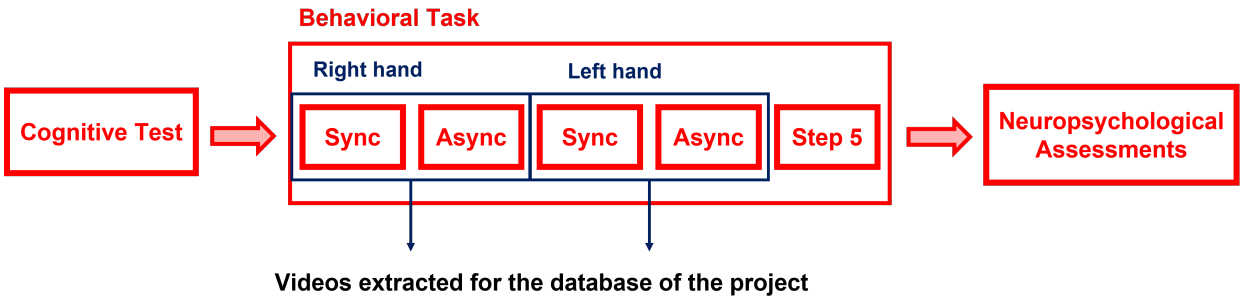

Figure 6.2: General experimental protocol for a participant in study B. In the behavioral task, the back stimulation is randomized, not necessarily being the right hand first. Inside this block, the async and sync conditions are also presented in a random order, between participants. Each of the 4 conditions lasts 2 minutes. The total behavioral task lasts 1 hour, because the participants are interviewed between blocks.

either due to technological constraints or because the participant was not capable of performing the complete task, usually happens more in PD patients.

The same metrics as presented in study A were also collected from the participants in this study, allowing to compute a similar set of demographic comparisons.

6.1.2 Sociodemographics

The sociodemographic comparisons are the same performed in study A. The database of videos of this study has in total 49 participants. However the sociodemographics only presents 33 participants due to various factors. This happens because for a participant to be valid to be analyzed, it has to have logged at least one of the first 2 trials, pass all the cognitive tests and they need to have the clinical data properly compiled, for it to be collected by the pipeline. This is one of the downsides of gathering data from various locations, as the transfer of knowledge some times may be complicated. Nonetheless, the videos from the 49 participants were used in DLC, but the results and motion features were extracted simply from these 33.

The results of the sociodemograpics between HC and PD are very similar to the ones of study A. The only difference is that there more women in the healthy participants. The age matches and the MoCA from both groups is higher than 26.

Regarding the PD-PH and PD-nPH comparison, some differences appear when looking at the population of study A. The PD-PH population as a higher disease duration than the non-PH, which is the opposite of before. The only reason for this, is in case some of the PD patients have cases of lingering PH, *i.e.* even tough the disease progressed the sPH did not disappear. Another opposite comparison, is that although the PD-nPH have a lower disease duration, their motor impairment is more severe. Finally, the last result that also changes is the medication, that is very similar between groups, instead of the big jump presented in study A. An important note is that in these 2 features (UPDRSIII and LEDD), the data of 6 and 10 participants is missing, respectively. Therefore, those 2 scores may not describe the entirety of the population.

Table 6.1: Sociodemographic comparison between HC and PD

	HC (N=14)	PD (N=19)	p
Gender	6 (M)	11 (M)	0.393 (χ2)
Age (years)	66.43 ± 5.96	60.90 ± 10.82	0.094
MoCA (scale)	28.64 ± 1.08	27.84 ± 1.68	0.129

The p-value of the gender was obtain with a chi-squared test, while the other factors come from an independent student t-test.

Table 6.2: Sociodemographic comparison between PD-PH and PD-nPH

	PD-PH (N=8)	PD-nPH (N=11)	p
Gender	7 (M)	4 (M)	0.026 (χ2) *
Age (years)	57.25 ± 14.06	63.55 ± 7.34	0.220
PD Duration (years)	10.13 ± 4.32	8.82 ± 5.86	0.601
UPDRS III (scale)	14.17 ± 7.52	18.71 ± 14.65	0.508
MoCA (scale)	27.50 ± 1.69	28.09 ± 1.70	0.464
LEDD (mg)	818.25 ± 341.61	855.20 ± 903.614	0.941

The p-value of the gender was obtain with a chi-squared test, while the other factors come from an independent student t-test

Nonetheless, and most importantly, none of these features is significantly different between the groups, which once again proves that the influence of the sensorimotor task is causing the movement and behavior to change.

6.2 Motion tracking Specifications and Model Performance

The heterogeneous look and diverse attributes of each video recording requires a 2 step pipeline, since the most reliable option is to apply the prediction-based adaptive cropping. As the starting point, it is important to define how many frames are needed to extract from each video. In study A, the total frame count was 1587. In this study, 49 participants have videos, however not all of them have the 4 videos of all the conditions. Therefore, to not promote data imbalance according to one scenario, only 2 videos per participant were inputted, total of 98. The videos selected were both sync conditions of both hands. Unlike study A, if a different hand is used, the viewpoint does not change. In order to achieve a similar total frame count to study A, but also taking into consideration the trade-off, between time spent on manual labeling and achieving the best accuracy result, it was concluded that 15 frames, per video, would be sufficient. By following the project methods, a total of 1469 frames were extracted and hand-labeled.

At this stage a problem surfaced in the development of the project. The point in this first iteration is to get an approximation of the position of the participant. However, with the diversity of that and high inter-subject variability, mixing all the data "into a big pot" did not seem the best solution. The most problematic factor is the angle from which the participant is filmed, since it varies both the side of the body closest to the camera, hiding and exposing different keypoints, and it also flips the direction and position of the robot, making it closer to one of the sides of the frame. In order to structure the database of videos, 4 main camera angles were defined, depending in which side of the participant the camera was placed (left or right) and whether the camera was placed orthogonally to the robot, *i.e.*, filming laterally the profile view of the participant, or if it was placed diagonally. Hence the 49 patients were divided into these 4 angles.

Based on this angle division, 4 independent datasets were assembled, relying on a more the consistent look within each angle. After running each model, following the specifications and parameters of study A, the tracking results revealed to be satisfactory on identifying the general position of the participant. From this, the "DownCrop" videos were created, standardizing the look of the database.

The next step is to compute the model, with the intention of accurate predictions, to extract realistic motion features. However, with the standardized look, a question is posed whether is better to continue to use different models per angle or if it is better to combine all the new videos into one dataset.

6.2.1 Tracking on Multiple Angles *VS* Full Dataset

To answer this question, the 5 models were computed to verify what is the best solution. A advantage of using 4 separate models is that each set of videos is consistent within itself, and the

Table 6.3: Train and test error for each one the models, related to the number of labeled frames.

	LL	LD	RD	RL	Full
Number of labeled frames	780	165	195	330	1485
Best snapshot	450,000	450,000	300,000	500,000	350,00
Train error (pixels)	2.08	1.88	1.78	1.65	1.84
Test error (pixels)	2.29	2.38	1.79	1.93	2.3

The name of the columns refer to each of the angles. The L_ or R_ is left or right, and _L or _D is lateral or diagonal.

model gets specific for that type of data. The analysis pipeline to later extract the motion features also needs to be tailored to each angle.

The disadvantage of this method, is that there is less data available for the training of the models. In the particular case of the LD angle, there are only 6 participants out of 49 that are recorded in that way. This will lead to low accuracies and high jitter (low precision) when analysing the new videos.

Usually the algorithms and deep learning models used in the back-bone of DLC do not react very well to the lack of data. So, this question is a based on an experimental trade off between how consistent does the training set need to be *VS* how much data is needed to overcome the variability of the of the subjects.

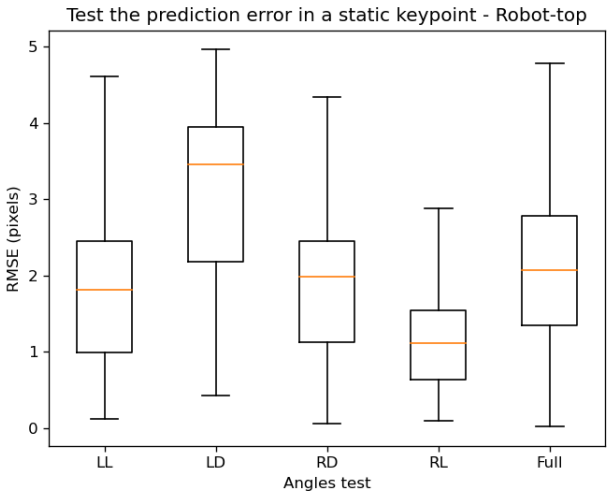

Figure 6.3: Box plots of the RMSE do the "*robot_top*" keypoint, in all the models.

This keypoint was chosen since it should be fixed in one spot and it is labeled every time. So if the RMSE is high, it means that the spot was not very well generalized. If this is the case for a sationary point, the analysis of a moving one is assumed to be even more dispersed, with a higher variability, that is interpreted as as bad accuracy and probably bad precision. What can be seen is that dependant on the angle, the tracking of the point can either be very good, for instance in RL; or bad, in LD. The robot top keypoint is specially problematic, since it changes from the left to the right of the frame, depending on the view angle. Demonstrating a good performance, comparable

to when the models are computed separately, when all the videos are combined together is the proof that the model can generalize to the various angles.

In conclusion, the model based the full dataset was the one used to run the analysis in all the videos. This was the best choice, because if the separate angle strategy would be used, probably even more participants would have needed to be discarded, since the model for the left diagonal angle did not have enough data to generalize all body parts.

After analysis, by visual inspection the model is able to detect every body part and responds very well to the changes of angle, to momentary occlusions, to differences in lighting, among other sources of noise. However, it is also clear that the jitter in the predictions are more intense than the results achieved in study A. Nonetheless, wide movements in frame can be tracked precisely, but in regard of small movements they are indistinguishable from jitter.

After analysing the whole dataset and predict the body part position, is time to extract the motion features out of the videos.

In the behavioral task, although the participants use both hands, the movement of interest comes from the original reaction when experiencing the sensorimotor stimulus and the conflict introduced by delay for the first time. The patterns exhibited are more natural and describe the health state of the participant more clearly, both if healthy or if it is a patient with PD. Hence, the motion features were only extracted from the hand that performed the first 2 trials. In case the participant falls into the category of being filmed in the LL or LD angle, all the signals regarding movement and speed are flipped according to the X axis, since the arm in the video was extending in the direction of the left side of the frame, when it should be towards the right. Like this, the motion signals across all participants.

6.3 Results

In this study, the statistical tests performed are the same and organized identically to the ones in study A. When looking at the results, the features of interest to be analyzed are the ones that arise from a significant difference, presenting main effects, either on group or condition.

The plots A and D, regarding the frequency of pokes, a main effect is revealed once again related to the synchrony of the stimulus independent of the population tested. However, in the finger any other feature reveals a main effect.

If a comparison is done with these results and the ones achieved in the controlled conditions, the orders of magnitude of the plots are similar between each studies. This is a "sanity check" that the results are not senseless. This also proves that the standardization using the robot height is efficient even if the scenarios are highly variate.

In plot C, a difference is seen on the speed of the PD group, since it should be higher than HC, however in async it decreases. Concerning the plots E and F, in the controlled database there were clear differences, however no main effect is shown here. Both the amplitude and speed behave completely different being slightly higher in the PD-nPH instead of in the PD-PH. Since there

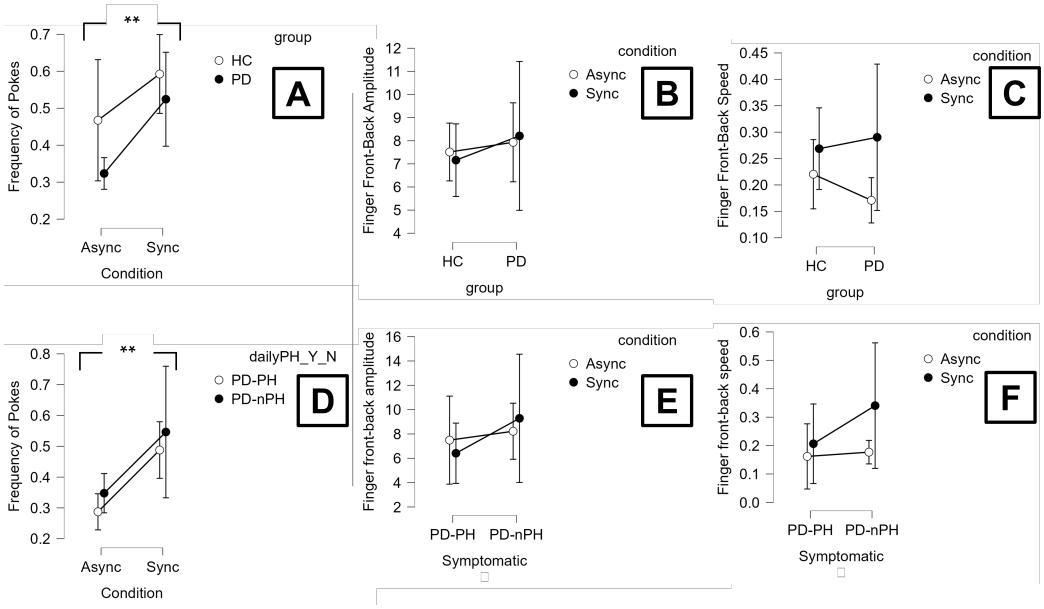

Figure 6.4: Plots of the statistical results of the motion features of the moving finger. The top row is the HC *VS* PD (A to C) and the bottom the PD-PH *VS* PD-nPH (D-F). From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.3.

is not main effect, it can be assumed that these results do not express the intended behavior that differentiates the groups.

When observing the behavior of the elbow and shoulder, most of the plots do not follow the same patterns as seen before in the controlled condition.

When looking to the elbow, specially in plot A, a new main effect surfaces that was not present in the control condition. The average amplitude of the movement of the elbow is much greater in PD, specially in async. These results are in accordance to what is observed in study A, however in a bigger scale, since the difference between HC and PD, respectively goes from 1 to 3.5 cm, where before it simply went form 0.5 to 1.5cm.

If compared with plot E, that amplitude is driven by the PD-PH behavior. In this plot the scale is also bigger, since the async average is around 5 cm, where before, in study A, it was simply 2 cm. Curious enough, the only plots where the slopes and motion features match the previous results, are the ones that exhibit a statistical difference.

Analysing the elbow speed plots, plot B exhibits a complete different behavior and in plot F the sync async lines are inverted, showing even a slight decrease in sync.

Moving on to the shoulder, only the sync movement goes towards what was previously observed, on average, but the variance interval is so big that no special meaning can be derived from it.

In plots G and H, the movement is generally different from study A, showing a similar behavior to what happened in the finger plots, increasing the amplitude and speed on the PD-nPH group.

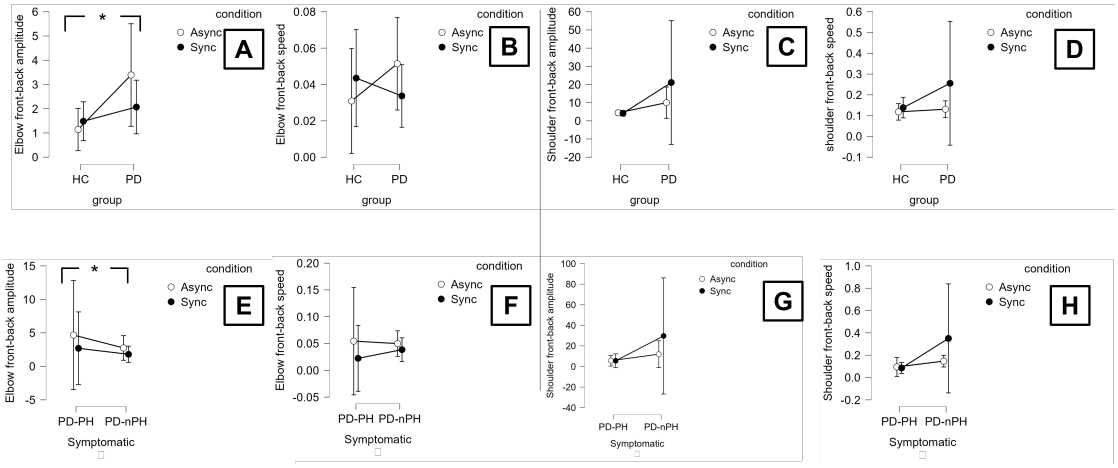

Figure 6.5: Plots of the statistical results of the motion features of the elbow (left) and shoulder (right). The top row is the HC *VS* PD (A to D) and the bottom the PD-PH *VS* PD-nPH (E-H). The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05). Error bar represents 95% confidence interval. The statistical results can be found in Appendix A.5.6.4.

Once again, the sync is higher in this group, but the variability is to wide to explain the behavior.

6.4 Discussion

Knowing the diversity of conditions from which the log videos were acquired, created an opportunity to test the limitations of the methodology towards different sources of noise. As a result, the motion features were extracted and compared to verify if the same patterns emerged in this second study, or if new main effects were revealed.

From the results observed, the methods are still able to verify the difference in the frequency of pokes, highly dependent on the condition. This main effect is explained by the same reason presented in study A, since the frequency decreases a lot in the async.

Another, main effect observed was the significant difference in groups on the elbow amplitude. The slopes follow the same trend as in study A, further supporting the theory that PD patients tend to do a rocking motion with the body, resulting in big shifts of the elbow back and forth.

The plot of amplitude on the PD-PH group is also bigger than in the PD-nPH, further proving that the patients with sPH are more susceptible to the task, moving more widely.

Since, these are the 2 only main effects presented from the results and, coincidentally, they corroborate the conjectures made in study A, it is plausible that the other movements may not be describing the accurately the real behavior of the participants. This could be due to the jitter of the predictions of the body parts or, more probably, due to the dissimilarities that each participant has. This demands for a more robust standardization of each movement and angle, to be certain that the metrics achieved are good descriptors of reality.

As a side note, although non significant in the sociodemographics' table, the average disease duration is inverted in the symptomatic PD groups, when compared to study A. This is could be correlated to the fact that in some plots the behavior between PD-PH and PD-nPH is also inverted, as for example the behavior of the shoulder in plots G and H.

6.5 Final Remarks

The objective of this study was to understand what are the limitations of the methodology pipeline and what standards need to be followed to guarantee a minimal viable set of characteristics to ensure good tracking results and metrics.

To do so, the database acquired consists of multiple setups, that change both in recording environment, such as different backgrounds, lighting conditions or relative position between the camera and the subject, but also technical specs, as different resolutions and aspect ratios. Although the pipeline implemented tries to correct and take as many of this variances into account, there are some factors that cannot be controlled in the current state, such as the influence of the recording angle. Therefore a trade-off was studied, concerning what the best strategy is, either to have a model per angle, or try to combine all angles together in a big dataset. The former has a problem of splitting the data, rendering some angles with too many participants, and other without sufficient data. When analyzing the error of each model, the one based on the complete dataset was able to generalize enough to recognize the different situations, being the solution to follow.

After extracting the motion features, the behaviors were examined and compared with the control study A. The observed main effects between groups were very insufficient, only pointing the change in frequency of pokes between conditions, and higher amplitude of elbow movement, back and forth. The latter supports the theory developed in study A, since the slopes tend in the same direction. As the other features are not significant, it can be assumed that their metrics could be influenced by the heterogeneity of the data.

Hence, a set of conditions needs to be implemented to make sure the participant performing the task can be tracked accurately, and that the motion features extracted are good descriptors of its movement. The most important is that the camera is immobile during the experiment, pointing to the participant and front robot, from a predetermined angle. This should be pointing to the participant from a lateral view, chosen according to the hand that is doing the poking. This will allow for the movement front and back to be tracked properly, in all joints. If possible, a plain background without researchers passing behind would be beneficial (the tracking may get confused and track that person, instead of the participant). Regarding the technical requirements of the camera, it can a single 2D camera, with at least HD resolution, preferably fullHD or higher. The models are adapted for a 16:9 aspect ratio. Being these the main constraints, a simple smartphone attached to a tripod, positioned correctly, is enough to have good log data.

In conclusion, from this study is possible to have an idea of what should be the conditions followed for a multi-centric study or for a product installed in different clinical settings, with non-technical staff manipulating the system. This research case shows some of the technological

challenges for clinical translation, when moving from the controlled environments of a laboratory, to realistic data acquired from multiple locations and circumstances.

Chapter 7

Conclusion and Future Work

According to the World Health Organization, Parkinson's disease is affecting more than 8.9 million people worldwide [10], being the second most prevalent neurological disease. One of the symptoms of the disease are presence hallucinations that have been estimated to be a putative predictors severe evolution of the disease, making them a key subject to be studied. Hence, the sensorimotor feedback task was developed to induce them in a controlled and repeatable manner [14]. The task based on a robotic system that delivers a feedback, with spatial-temporal conflict, according to the poking movement of the participant.

When performing the task, healthy participants and Parkinson's patients, specially the ones with symptomatic presence hallucinations in their daily lives, exhibit different reactions that cannot be explained by external factors, such as reduced cognition, motor impairments, among others. As such, it can be considered that the variance in movement is caused by a different susceptibility to the experimental manipulation. Therefore, with the robotic setup and the behavioral task, it is possible to study if there is a typical behavior inherent to different groups of participants.

To do so, their posture and movements need to be quantified, analyzed and evaluated in a non-intrusive manner to not affect their behavior. Hence, the DeepLabCut software, a markerless pose estimation system based on deep learning, was implemented to track the participants on the log videos of the experiment. The data resulting from the analysis with this system are the coordinates of the body parts during the each sensorimotor condition. This allows to extract motion features that describe each type of group, thus allowing to study whether healthy controls (HC) and PD patients exhibit common traits, movements and behaviors while performing the task. Furthermore, the same question can be prompted between those PD patients that usually have PH during their daily lives (denominated as PD-PH) or the ones that do not (PD-nPH).

When analysing and comparing the motion features some differences were found between the groups. The healthy participants are able to maintain the same posture, with consistent behavior (same amplitude of poking and speed) across trials, following the task guidelines. PD patients, on the other hand, due to their rigidity and overall weakness, when are poked by the back robot, tend to be pushed to the front. This results in a behavior where the patient does the poke using the whole torso at the same time, doing a rocking motion back and forth. Inside this group, the PD-PH reveals

to be more sensible to the task, since their behavior,is greatly affected. They are shown to be more unimpeded, presenting on average higher amplitudes of motion and arm extension, however being consistently slower in the async condition. Thus, it can be considered these behavioral differences are related to the different susceptibility to the manipulation.

However, to achieve these results it is necessary to guarantee a practical setup, to fulfill certain technological requirements, to ensure a good tracking and for the motion feature extraction pipeline to work. Ideally, the patient should be filmed from a predefined angle, with the camera immobile, usually placed laterally orthogonal to the robotic system, capturing both the participant and the front robot in a profile view. The task should be performed in a location where there is not people passing on the background. Ideally, the background should be plain. It is also important to film the participant in a place with good lighting conditions, for all the body parts to be visible. If the task is logged in a setup following these guidelines, the system implemented is able to gather quality data. Consequently, the tracking results will have high accuracy, to analyze the participant's behavior, to quantify it and, ultimately, categorize it as one of the groups.

The conclusions achieved within the presented work are the first steps towards an automatized pipeline, that can capture and translate the behavior into quantifiable metrics. However, new methods could be added as future work. Another interesting body part, in the sensorimotor task, could be the non-moving arm. Sometimes, the dominant hand used in the task is not the same as the side affect by PD, since it is a lateralized condition, where only one side of the body manifests the motor symptoms. So, when doing the poking, the resting hand may exhibit a more intense reaction, such as more intense tremoring or twitching, that are factors that could help identify the PD group. In the context of general presence hallucinations a typical behavior manifested is the searching of the environment around, to check if there is someone close to them, reacting to the feeling of a presence. Therefore, the tracking of other body parts, for example the face, could be a useful indicator of another typical behavior of a PD-PH group, since they would be moving their heads to check. In deep learning, there are plenty of automatic methods to detect faces, and consequently, their movement and facial features.

All of these scientific improvements are the basis to prove that a prototype device, developed in a laboratory environment can evolve to be a commercial product. The need for clinical translation, *i.e.* inputting in the market a minimum viable product, requires a polished pipeline to prove its worth in a medical or clinical setting. At the current stage the system is not yet robust enough to handle the highly variable scenarios, however a first proof of concept regarding the application of motion tracking technology, in controlled conditions, was settled; as well as the critical conditions that may be troublesome. The next phase, involves the application of enhancement steps could also be applied to polish the method pipeline. A good improvement would be related to the standardization of data not just in distance between camera and participant, but also in angle. This would allow different angles of recording to be correctly comparable across participants. To achieve this, in the beginning of the log of each experiment, the researcher would need to show a checkerboard pattern close the robot. Since the measures of each square are known when pointing to the camera and when close to the robot, it is possible to standardize the effect the filming angle.

The opportunity to compute more motion features, to unveil additional patterns and behaviors, could be useful in the design of a diagnosing tool. Finally, all the features could also be applied to an optimized classifier to distinguish the groups automatically, based on machine learning.

This would be the ultimate device, in the future, to help neurologists deliver a more thorough diagnostic, that includes both the analysis of non-motor symptoms, and how prone is a subject to have them, as well as evaluate the patients' motor conditions, by delivering understandable metrics of movement, in an automatized way.

References

[1] Timeline - overview for deeplabcut publication year: 2017, 2018, 2019, 2020, 2021 in publications - dimensions. URL https://app.dimensions.ai/analytics/publication/overview/timeline?search_mode=content&search_text=deeplabcut&search_type=kws&search_field=full_search&order=times_cited&or_facet_year=2017&or_facet_year=2018&or_facet_year=2019&or_facet_year=2020&or_facet_year=2021&year_from=2016&year_to=2021.

[2] Parkinson's disease exam | stanford medicine 25 | stanford medicine. URL https://stanfordmedicine25.stanford.edu/the25/parkinsondisease.html.

[3] Hallucinations in parkinson disease. *Nature reviews. Neurology*, 5:331–342, 2009. ISSN 1759-4766. doi: 10.1038/NRNEUROL.2009.62. URL https://pubmed.ncbi.nlm.nih.gov/19498436/.

[4] King's parkinson's disease pain scale, the first scale for pain in pd: An international validation. *Movement Disorders*, 30:1623–1631, 10 2015. ISSN 15318257. doi: 10.1002/MDS.26270/ABSTRACT.

[5] The psychosis spectrum in parkinson disease. *Nature reviews. Neurology*, 13:81–95, 2 2017. ISSN 1759-4766. doi: 10.1038/NRNEUROL.2016.200. URL https://pubmed.ncbi.nlm.nih.gov/28106066/.

[6] Minor hallucinations in parkinson disease: A subtle symptom with major clinical implications. *Neurology*, 93:259–266, 8 2019. ISSN 1526-632X. doi: 10.1212/WNL.0000000000007913. URL https://pubmed.ncbi.nlm.nih.gov/31289146/.

[7] Current and projected future economic burden of parkinson's disease in the u.s. *NPJ Parkinson's disease*, 6, 12 2020. ISSN 2373-8057. doi: 10.1038/S41531-020-0117-1. URL https://pubmed.ncbi.nlm.nih.gov/32665974/.

[8] Neuroanatomy, substantia nigra. *StatPearls*, 10 2021. URL https://www.ncbi.nlm.nih.gov/books/NBK536995/.

[9] Deeplabcut for multi-animal projects, 2022. URL https://github.com/DeepLabCut/DeepLabCut/blob/master/docs/maDLC_UserGuide.md.

[10] Parkinson disease, 6 2022. URL https://www.who.int/news-room/fact-sheets/detail/parkinson-disease.

[11] Dag Aarsland, Byron Creese, Marios Politis, K. Ray Chaudhuri, Dominic H. Ffytche, Daniel Weintraub, and Clive Ballard. Cognitive decline in parkinson disease. *Nature reviews. Neurology*, 13:217–231, 3 2017. ISSN 1759-4766. doi: 10.1038/NRNEUROL.2017.27. URL https://pubmed.ncbi.nlm.nih.gov/28257128/.

[12] Shahar Arzy, Margitta Seeck, Stephanie Ortigue, Laurent Spinelli, and Olaf Blanke. Induction of an illusory shadow person. *Nature 2006 443:7109*, 443:287–287, 9 2006. ISSN 1476-4687. doi: 10.1038/443287a. URL https://www.nature.com/articles/443287a.

[13] H. Bejr-kasem, F. Sampedro, J. Marín-Lahoz, S. Martínez-Horta, J. Pagonabarraga, and J. Kulisevsky. Minor hallucinations reflect early gray matter loss and predict subjective cognitive decline in parkinson's disease. *European journal of neurology*, 28:438–447, 2 2021. ISSN 1468-1331. doi: 10.1111/ENE.14576. URL https://pubmed.ncbi.nlm.nih.gov/33032389/.

[14] Fosco Bernasconi, Eva Blondiaux, Jevita Potheegadoo, Giedre Stripeikyte, Javier Pagonabarraga, Helena Bejr-Kasem, Michela Bassolino, Michel Akselrod, Saul Martinez-Horta, Frederic Sampedro, Masayuki Hara, Judit Horvath, Matteo Franza, Stéphanie Konik, Matthieu Bereau, Joseph André Ghika, Pierre R. Burkhard, Dimitri Van de Ville, Nathan Faivre, Giulio Rognini, Paul Krack, Jaime Kulisevsky, and Olaf Blanke. Robot-induced hallucinations in parkinson's disease depend on altered sensorimotor processing in frontotemporal network. *Science translational medicine*, 13, 4 2021. ISSN 1946-6242. doi: 10.1126/SCITRANSLMED.ABC8362. URL https://pubmed.ncbi.nlm.nih.gov/33910980/.

[15] Fosco Bernasconi, Eva Blondiaux, Giulio Rognini, Herberto Dhanis, Laurent Jenni, Jevita Potheegadoo, Masayuki Hara, and Olaf Blanke. Neuroscience robotics for controlled induction and real-time assessment of hallucinations. *Nature Protocols 2022*, pages 1–24, 9 2022. ISSN 1750-2799. doi: 10.1038/s41596-022-00737-z. URL https://www.nature.com/articles/s41596-022-00737-z.

[16] Sarah Jayne Blakemore, Daniel Wolpert, and Chris Frith. Why can't you tickle yourself? *Neuroreport*, 11:R11–R16, 2000. ISSN 0959-4965. doi: 10.1097/00001756-200008030-00002. URL https://pubmed.ncbi.nlm.nih.gov/10943682/.

[17] Olaf Blanke, Polona Pozeg, Masayuki Hara, Lukas Heydrich, Andrea Serino, Akio Yamamoto, Toshiro Higuchi, Roy Salomon, Margitta Seeck, Theodor Landis, Shahar Arzy, Bruno Herbelin, Hannes Bleuler, and Giulio Rognini. Neurological and robot-controlled induction of an apparition. *Current biology : CB*, 24:2681–2686, 2014. ISSN 1879-0445. doi: 10.1016/J.CUB.2014.09.049. URL https://pubmed.ncbi.nlm.nih.gov/25447995/.

[18] Peter Brugger. One hundred years of an ambiguous figure: Happy birthday, duck/rabbit! *Perceptual and Motor Skills*, 89:973–977, 8 1999. ISSN 00315125. doi: 10.2466/pms.1999.89.3.973. URL https://journals.sagepub.com/doi/10.2466/pms.1999.89.3.973.

[19] Peter Brugger, Marianne Regard, and Theodor Landis. Unilaterally felt "presences": The neuropsychiatry of one's invisible doppelganger. *Cognitive and Behavioral Neurology*, 9, 04 1996.

[20] Paulo Bugalho, Tânia Lampreia, Rita Miguel, Marcelo D. Mendonça, André Caetano, and Raquel Barbosa. Non-motor symptoms in portuguese parkinson's disease patients: correlation and impact on quality of life and activities of daily living. *Scientific Reports 2016 6:1*, 6:1–9, 8 2016. ISSN 2045-2322. doi: 10.1038/srep32267. URL https://www.nature.com/articles/srep32267.

[21] Leslie J. Findley. The economic impact of parkinson's disease. *Parkinsonism related disorders*, 13 Suppl, 9 2007. ISSN 1353-8020. doi: 10.1016/J.PARKRELDIS.2007.06.003. URL https://pubmed.ncbi.nlm.nih.gov/17702630/.

[22] Paul C. Fletcher and Chris D. Frith. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience 2008 10:1*, 10: 48–58, 12 2008. ISSN 1471-0048. doi: 10.1038/nrn2536. URL https://www.nature.com/articles/nrn2536.

[23] Gilles Fénelon, Florence Mahieux, Renaud Huon, and Marc Ziégler. Hallucinations in parkinson's disease: prevalence, phenomenology and risk factors. *Brain : a journal of neurology*, 123 (Pt 4):733–745, 2000. ISSN 0006-8950. doi: 10.1093/BRAIN/123.4.733. URL https://pubmed.ncbi.nlm.nih.gov/10734005/.

[24] Gilles Fénelon, Thierry Soulas, Laurent Cleret De Langavant, Iris Trinkler, and Anne Catherine Bachoud-Lévi. Feeling of presence in parkinson's disease. *Journal of neurology, neurosurgery, and psychiatry*, 82:1219–1224, 2011. ISSN 1468-330X. doi: 10.1136/JNNP.2010.234799. URL https://pubmed.ncbi.nlm.nih.gov/21551471/.

[25] Jonh Geiger. *The Third Man Factor - Surviving the Impossible by John Geiger – Canongate Books*. Canon Gate, 2009. URL https://canongate.co.uk/books/1173-the-third-man-factor-surviving-the-impossible/.

[26] Christopher G. Goetz, Barbara C. Tilley, Stephanie R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B. Stern, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E. Lang, Andrew Lees, Sue Leurgans, Peter A. LeWitt, David Nyenhuis, C. Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A. Teresi, Jacobus J. van Hilten, Nancy LaPelle, Pinky Agarwal, Saima Athar, Yvette Bordelan, Helen M. Bronte-Stewart, Richard Camicioli, Kelvin Chou, Wendy Cole, Arif Dalvi, Holly Delgado, Alan Diamond, Jeremy P. Dick, John Duda, Rodger J. Elble, Carol Evans, Virgilio G. Evidente, Hubert H. Fernandez, Susan Fox, Joseph H. Friedman, Robin D. Fross, David Gallagher, Christopher G. Goetz, Deborah Hall, Neal Hermanowicz, Vanessa Hinson, Stacy Horn, Howard Hurtig, Un Jung Kang, Galit Kleiner-Fisman, Olga Klepitskaya, Katie Kompoliti, Eugene C. Lai, Maureen L. Leehey, Iracema Leroi, Kelly E. Lyons, Terry McClain, Steven W. Metzer, Janis Miyasaki, John C. Morgan, Martha Nance, Joanne Nemeth, Rajesh Pahwa, Sotirios A. Parashos, J. S.Jay S. Schneider, Anette Schrag, Kapil Sethi, Lisa M. Shulman, Andrew Siderowf, Monty Silverdale, Tanya Simuni, Mark Stacy, Matthew B. Stern, Robert Malcolm Stewart, Kelly Sullivan, David M. Swope, Pettaruse M. Wadia, Richard W. Walker, Ruth Walker, William J. Weiner, Jill Wiener, Jayne Wilkinson, Joanna M. Wojcieszek, Summer Wolfrath, Frederick Wooten, Allen Wu, Theresa A. Zesiewicz, and Richard M. Zweig. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders : official journal of the Movement Disorder Society*, 23:2129–2170, 11 2008. ISSN 1531-8257. doi: 10.1002/MDS.22340. URL https://pubmed.ncbi.nlm.nih.gov/19025984/.

[27] Karl Jaspers. Über leibhaftige bewusstheiten (bewusstheitstäuschungen), ein psychopathologisches elementarsymptom. zeitschrift pathopsychol. 2, 150–161. 1913. URL https://philpapers.org/rec/JASGP.

[28] Greenland JC and Barker RA. The differential diagnosis of parkinson's disease. *Neurology*, 43:S1–11, 2018. ISSN 1526632X. doi: 10.15586/CODONPUBLICATIONS.PARKINSONSDISEASE.2018.CH6. URL https://pubmed.ncbi.nlm.nih.gov/30702839/.

[29] Antonina Kouli, Kelli M. Torsney, and Wei-Li Kuan. Parkinson's disease: Etiology, neuropathology, and pathogenesis. *Parkinson's Disease: Pathogenesis and Clinical Aspects*, pages 3–26, 12 2018. doi: 10.15586/CODONPUBLICATIONS.PARKINSONSDISEASE.2018.CH1. URL https://pubmed.ncbi.nlm.nih.gov/30702842/.

[30] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, and Alexander Mathis. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods 2022 19:4*, 19:496–504, 4 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01443-0. URL https://www.nature.com/articles/s41592-022-01443-0.

[31] P. M. Llorca, B. Pereira, R. Jardri, I. Chereau-Boudet, G. Brousse, D. Misdrahi, G. Fénelon, A. M. Tronche, R. Schwan, C. Lançon, A. Marques, M. Ulla, P. Derost, B. Debilly, F. Durif, and I. De Chazeron. Hallucinations in schizophrenia and parkinson's disease: an analysis of sensory modalities involved and the repercussion on patients. *Scientific reports*, 6, 12 2016. ISSN 2045-2322. doi: 10.1038/SREP38152. URL https://pubmed.ncbi.nlm.nih.gov/27905557/.

[32] C. Marras, J. C. Beck, J. H. Bower, E. Roberts, B. Ritz, G. W. Ross, R. D. Abbott, R. Savica, S. K. Van Den Eeden, A. W. Willis, and Cm Tanner. Prevalence of parkinson's disease across north america. *npj Parkinson's Disease 2018 4:1*, 4:1–7, 7 2018. ISSN 2373-8057. doi: 10.1038/s41531-018-0058-0. URL https://www.nature.com/articles/s41531-018-0058-0.

[33] Alexander Mathis and Richard Warren. On the inference speed and video-compression robustness of deeplabcut. *bioRxiv*, page 457242, 10 2018. doi: 10.1101/457242. URL https://www.biorxiv.org/content/10.1101/457242v1https://www.biorxiv.org/content/10.1101/457242v1.abstract.

[34] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience 2018 21:9*, 21:1281–1289, 8 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0209-y. URL https://www.nature.com/articles/s41593-018-0209-y.

[35] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pages 1858–1867, 1 2021. doi: 10.1109/WACV48630.2021.00190. URL https://www.researchgate.net/publication/352389879_Pretraining_boosts_out-of-domain_robustness_for_pose_estimation.

[36] Dr. Matthew and E. Bain. Brain anatomy | white matter, cerebellum, cerebral cortex, medulla, 2011. URL https://www.healthpages.org/anatomy-function/brain-anatomy/.

[37] Reinhold. Messner. *The naked mountain*. Crowood, 2003. ISBN 1861266332. URL https://books.google.com/books/about/The_Naked_Mountain.html?hl=pt-PT&id=HbdEPwAACAAJ.

[38] Mark J. Millan, Annie Andrieux, George Bartzokis, Kristin Cadenhead, Paola Dazzan, Paolo Fusar-Poli, Jürgen Gallinat, Jay Giedd, Dennis R. Grayson, Markus Heinrichs, René Kahn, Marie Odile Krebs, Marion Leboyer, David Lewis, Oscar Marin, Philippe Marin, Andreas Meyer-Lindenberg, Patrick McGorry, Philip McGuire, Michael J. Owen, Paul Patterson, Akira Sawa, Michael Spedding, Peter Uhlhaas, Flora Vaccarino, Claes Wahlestedt, and Daniel Weinberger. Altering the course of schizophrenia: progress and perspectives. *Nature reviews. Drug discovery*, 15:485–515, 6 2016. ISSN 1474-1784. doi: 10.1038/NRD.2016.28. URL https://pubmed.ncbi.nlm.nih.gov/26939910/.

[39] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005. doi: https://doi.org/10.1111/j.1532-5415.2005.53221.x. URL https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53221.x.

[40] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature Protocols 2019 14:7*, 14:2152–2176, 6 2019. ISSN 1750-2799. doi: 10.1038/s41596-019-0176-0. URL https://www.nature.com/articles/s41596-019-0176-0.

[41] Miana Gabriela Pop, Carmen Crivii, and Iulian Opincariu. Anatomy and function of the hypothalamus. *Hypothalamus in Health and Diseases*, 11 2018. doi: 10.5772/INTECHOPEN.80728. URL undefined/state.item.id.

[42] B. Ravina, R. Camicioli, P. G. Como, L. Marsh, J. Jankovic, D. Weintraub, and J. Elm. The impact of depressive symptoms in early parkinson disease. *Neurology*, 69:342, 7 2007. ISSN 00283878. doi: 10.1212/01.WNL.0000268695.63392.10. URL /pmc/articles/PMC2031220//pmc/articles/PMC2031220/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2031220/.

[43] L. Weiskrantz, J. Elliott, and C. Darlington. Preliminary observations on tickling oneself. *Nature 1971 230:5296*, 230:598–599, 1971. ISSN 1476-4687. doi: 10.1038/230598a0. URL https://www.nature.com/articles/230598a0.

[44] James William. *The Varieties of Religious Experience: A Study in Human Nature*. NY: Longmans, Green, and CO, 1st edn edition, 1902. URL https://upload.wikimedia.org/wikipedia/commons/d/df/William_James_The_Varieties_of_Religious_Experience.pdf.

[45] Stefan Williams, Zhibin Zhao, Awais Hafeez, David C. Wong, Samuel D. Relton, Hui Fang, and Jane E. Alty. The discerning eye of computer vision: Can it measure parkinson's finger

tap bradykinesia? *Journal of the Neurological Sciences*, 416:117003, 9 2020. ISSN 0022-510X. doi: 10.1016/J.JNS.2020.117003.

[46] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural networks : the official journal of the International Neural Network Society*, 11:1317–1329, 1998. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00066-5. URL https://pubmed.ncbi.nlm.nih.gov/12662752/.

[47] Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science (New York, N.Y.)*, 269:1880–1882, 9 1995. ISSN 0036-8075. doi: 10.1126/SCIENCE.7569931. URL https://pubmed.ncbi.nlm.nih.gov/7569931/.

Appendix A

Appendix

A.1 Laboratorial Tests for PD diagnosis

To further characterize the disorder and to deliver a more concrete diagnostic, some laboratory tests can be performed, for example, a DATscan, which is a dopamine transporter scan using a single-photon emission computerized tomography (SPECT) scan. This test does not indicate if a person has PD, but it helps the doctor deliver a more complete diagnostic, since it highlights in the scan image the areas where there is a high concentration of dopamine. Consequently, if the area represented is small, then there is a deficit of dopamine, which is the main reason behind Parkinson.

Another imaging tests are Magnetic Resonance Imaging (MRI) and positron emission tomography (PET) scan, that can also be used for the same purpose. In the former, on recent methods, the neural network related to the PH induction was discovered (Bernasconi 2021), being a potential tool to characterize the disease in it's early stages.

A.2 PD Treatment Options

After a clear set of clinical points that lead to a convincing diagnosis of PD, a convenient treatment is outlined according to each patient. The ideal situation in PD treatment is to start before the loss of the DA cells in the SN region, however current state-of-the-art studies that can infer an early diagnosis, even 2 years before the patient starts to severely decline, are not predicted early enough [2].

The first approach to treat PD is through drug administration, to rebalance the concentration of dopamine in the brain, in order to control and optimally eliminate all of the motor symptoms. The most common medication prescribed is:

- Carbidopa-levodopa: this chemical compound is transformed by the body and turned into dopamine, directly correcting the low concentration that induces the symptoms. Unfortunately, the biggest downside of this drug is that during the years of treatment, the body

slowly gets less sensitive to it and higher doses are incrementally needed. When the maximum concentration is reached new treatments have to be implemented, most commonly, surgery

- Dopamine agonists: This are drugs that have the same effect as dopamine in the body, so their use can also reverse some of the PD symptoms. This method of medication is not usually prescribed and can produce more secondary effects than carbidopa-levodopa

- MAO B inhibitors: as the name suggests, this drug inhibits the brain enzyme monoamine oxidase B (MAO B), that is the protein responsible for the breakdown of dopamine, making its concentration last longer in the brain;

- Amantadine: this drug is mostly used in the beginning of the PD treatment, not to make the symptoms disappear, but to control some of the secondary effects of the carbidopa-levodopa drug.

As referred in the description of the first drug, this path of treatment does wear off during the extensive years of treatment and a new approach based on surgery was developed. Another scenario where surgery is an option is when the medication induces dyskinesia (involuntary, erratic, writhing movements of the face, arms, legs or trunk) or too many non-motor symptoms as a secondary effect, since high levels of dopamine can lead to major hallucinations and delusions, that can greatly affect the patient's quality of life, as will be seen ahead.

Surgery for PD treatment is based on deep brain stimulation, a technique where the neurosurgeon inserts electrodes (leads) through a small opening in the skull. The exact place where stimulation will occur is personalized for each patient, but it is usually focused on the subthalamic nucleus (STN). The surgery is conducted with the patients awake and in OFF state (without medication), for the doctor to evaluate what is the best spot for stimulation and to establish the best parameters on the neurostimulator. In order to conduct the electrical impulse between the lead and the neurostimulator, a subcutaneous insulated wire connects these two points going from the brain to the lower chest area. Subcutaneously, in the abdomen, the battery power supply and circuitry are inserted to modulate the signal sent to the brain. This is the preferred place for the neurostimulator to be installed since it is easy to access, in case of some update or battery change is needed. After surgery, the patient continues to take a low dosage of medication, in order to achieve the best possible result in symptom nullification.

Besides the aforementioned treatments, doctors also recommend occupational therapy with aerobic exercises, to promote active training on the control of balance and posture. It also should include exercises to improve muscle strength of the affected limbs, in order to prevent atrophy and to fight the bradykinesia [7,9]. Normally, the tremors also get attenuated when leading a healthy lifestyle. Having these kinds of activities helps not only in physical terms, but also in mental issues.

Many PD patients suffer from mental illnesses due to the slow withdraw from society, as it will be discussed in the next section. Doctors also prescribe antidepressant drugs to help them cope with the everyday life challenges and maintain emotional stability.

A.3 Socioeconomic impact

Parkinson disease is the second most prevalent neurodegenerative disease in world, estimating more than 10 million diagnosed patients worldwide. In this impact study most of the data was collected from the USA or Europe, as these are the regions with the most developed healthcare systems and available reports.

A detailed analysis on how the patient's life is influenced and altered by the disease, as well as the costs behind it, both direct (medication or surgery) and indirect (lost income and social security payments).

A.3.1 Socioemotional Impact on Quality of life

As previously seen, PD has both a motor and non-motor component that influence different aspects of the lives of patients, most of the times in a negative and unhealthy way.

A Portuguese study assessed the impact of motor (MS) and non-motor symptoms (NMS) on scales of Heath Related Quality of Life (HRQL) and Activities of Daily Living (ADL). The study began by analyzing the origin of the multiple non-motor symptoms and how they affected the normal daily life of PD patients, for example, if it provoked trouble sleeping, mood swings, abnormal fatigue, memory loss or confusion. Other psychological symptoms were evaluated and correlated with the previous ones, such as, visual hallucinations (seeing oneiric images while awake), apathy, anxiety and depression. To evaluate the impact of this symptoms in the HRQL and ADL two questionnaires were applied:

- Euro-QOL: This assessment is comprised of 2 sections. The first is the EQ-5D, which consists of five questions (graded in three levels of severity: 1 = no problem, 2 = moderate problem, 3 = severe problem) related to five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, resulting in the EQ-Index. The second section is the EQ-VAS, which is a vertical visual analogue scale ranging from 0 (worst imaginable health state) to 100 (best imaginable health state).

- Unified Parkinson's Disease Rating Scale Part II (UPDRSII): this test consists of a self-evaluation scale, which includes questions regarding the impact of motor dysfunction on daily activities, such as speech, swallowing, handwriting, dressing, hygiene, falling, salivating, turning in bed, walking, and cutting food.

The study collected and processed the responses of these tests, and the main results presented were that apathy and depression are deeply correlated and may have a common etiology, which can also lead to a possible common therapeutic process, and that these 2 NMS have a bigger impact

in HRQL than non-tremor MS. Sleep disturbances, at night, also have a high impact in the EQ-Index, showing an influence in the HRQL. Regarding the ADL, the non-tremor motor dysfunction revealed a more negative influence, followed by the memory complaints and hallucinations. All in all, the study concludes that NMS have a bigger impact on the quality of life and that the motor impairments affect more the daily activities [12].

Indeed, this is proved by other studies, especially the effect of bradykinesia on social communication. Most patients have plenty of issues in the progressive stages of PD due to conditions like facial masking, where the person has a reduced ability to spontaneously display emotional facial expressions making intent harder to transmit; and dysarthria, where the patient produces speech, which is grammatically and syntactically correct, but that is marked by an abnormal rhythm, harsh voice and inappropriate pauses. Together with these, patients also present emotion recognition problems, where the deficit of cognitive power makes it harder for them to interpret the other person.

As society is deeply based on inter-human interactions where clear speech is a key factor, the mentioned conditions lead to stigma directed to PD patients, where people from their everyday lives start to avoid talking to them leading to a sense of loneliness and dehumanization, as most of the times, a speaker will redirect questions and requests to the person in care and not to the patient, making them fall into self-isolation and a depression state [13].

This mental instability does not only affect the patient himself, but also the closest person in care. A study on the impact of PD in spouses' lives show that a super dependence and feeling of constant care is created as the disease progresses, which leads to states of chronic anxiety and burden. Besides, most of the households have a one-sided income, which adds to the worry and frustration, not only for the spouse but also for the patient [14]. In order to solve this, physicians and neurologist suggest an active lifestyle, for both house members, in order to engage in social life without the guilt of being present. PD patients can enroll in occupational therapy as the spouses have social meetings to keep the mental conditions of both persons healthy.

A.3.2 Economic burden and cost of PD

As mentioned in the beginning PD has a high prevalence in population and, with an aging society the incidence numbers are estimated to grow in the next decades.

From studies focused on the USA, the first important comparison is between the annual medical cost of normal healthy people and PD patients. In 2002, the total annual direct cost per patient was $11,247 for those without PD and $23,101 for those with PD. This represents an incremental cost of $10,349 annually due to the disease. The largest component of the direct cost was inpatient care and nursing home costs. This sums up to an estimated total annual burden (direct costs) of $23 billion in the USA, in 2002 [15]. A more recent study from 2017, estimated that the economic burden of PD was $51.9 billion, including direct medical costs of $25.4 billion and $26.5 billion in indirect and non-medical costs. This adds up to a rise of $2.4 billion in 15 years, only taking in consideration the direct costs. This second study has more precise data demonstrated on figure A.1.

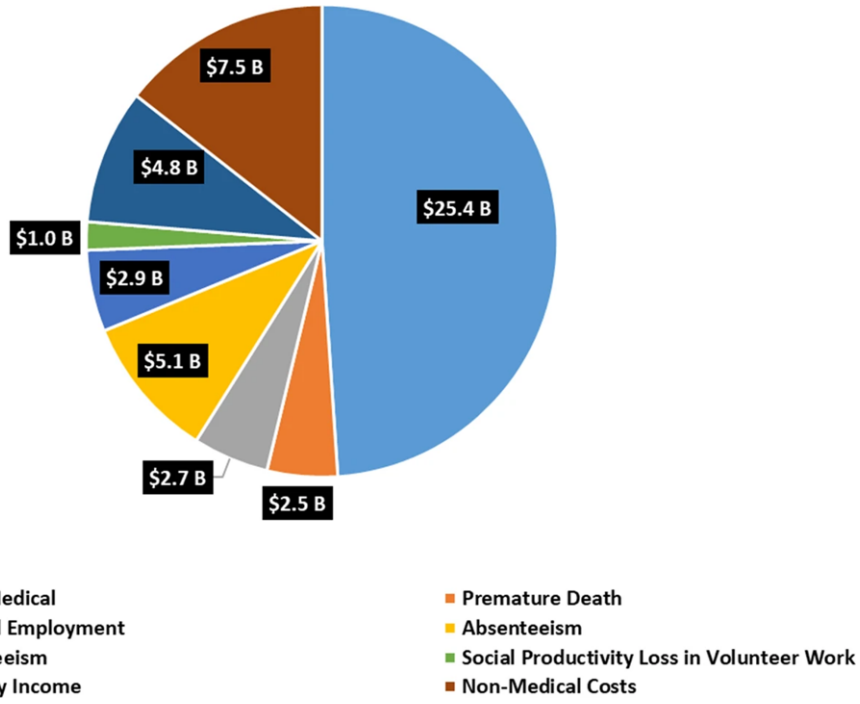

Figure A.1: Total economic burden of Parkinson's disease in the U.S., divided by components, in 2017 (in billions) [7]

The percentages demonstrated in graph are sustained by the Tables II and III of direct and indirect costs, respectively, presented in the Appendix.

This study also refers as an important aspect that more than 90% of the people surveyed was a candidate for Medicare Health Insurance, which helps support most of the annual costs of the disease.

This study also estimates that the annual burden of PD will continue to increase and a prediction for the next 2 decades, by 2037, is that the total economic impact of PD, in USA, will reach the $79.1 billion [7].

Analyzing the market and economic situation of Parkinson in Europe, the UK was also analyzed in the 2002 study, where a more detailed distinction of the costs between different classes of patients was performed. The average direct cost of a PD patient in UK was £5993 per year. However, this cost can suffer a lot of discrepancies depending on the age and stage of neurodegeneration of the patient. The following figure presents the annual costs dividing patients according to their age gap (figure A.2, on the right) and Hoehn and Yahr scale, to distinguish patients according to the severity and progress of their disabilities (figure A.2, on the left). Each bar is divided into 3 sections, related to the percentages of direct costs, namely the money spent in the National Health System (NHS), social services and in private expenditures [15].

Focusing on the indirect costs of UK, the loss of productivity was not very detrimental, since most of the patients were near to or on retirement, however the person responsible for caring for the patient, in case it is a familiar has to quit their professional career for a more flexible job. This adds up to an annual indirect cost over £27,000.

Table A.1: Direct medical cost of Parkinson's disease by age, gender, insurance coverage, and types of service (in 2017) [7]. *PWP – Person with Parkinson

Category	Studied Items
NHS (Primary care)	Drugs, GP visits, home visits by other health professionals
NHS (Secondary care)	Hospital inpatient and outpatient visits
Social services	Home help/support, formal home care, meals on wheels, nursing homes, sitting services, day centers
Private expenditure	Private residential/nursing home; private care at home; equipment; travel expenses; additional laundry, heating or lighting; incontinence aids; extra help/home care; miscellaneous specified PD-related costs

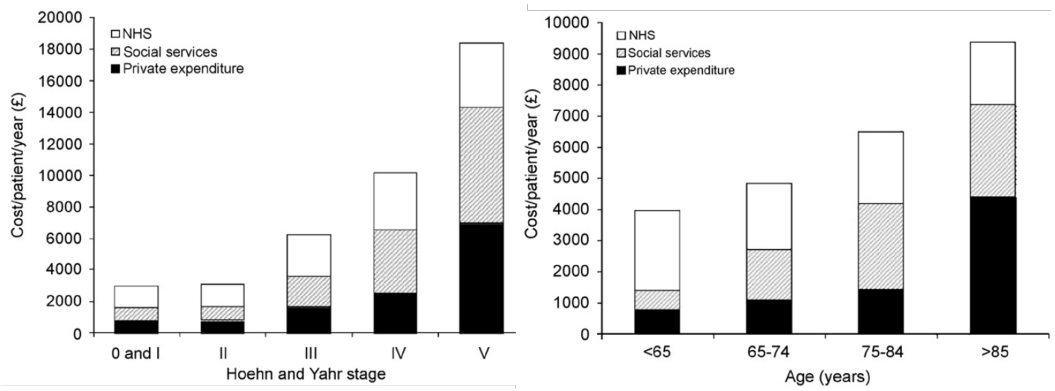

Figure A.2: Direct cost of PD in the UK according to the Hoehn and Yahr scale, on the left; and to age, on the right (n = 423).

When these costs are brought together, the model applied in the study inferred an economic burden of a total of £3,3 billion.

Regarding other countries in Europe, the PD impact is summarized in Table IV, in the appendix, where information from France, Sweden, Germany and Finland is presented, also divided according to disease severity, both for direct and indirect costs.

A.4 MoCap System Descriptions

One of the most advanced motion tracking technology that solely focus on inertial sensors are the X-Sens. Their system is equipped with 17 tracker units, that consist of wireless IMUs, that are strapped on specific parts of body joints or attached on their full-body suit to be instantly in the correct place. One advantage is that the systems is easy to calibrate after the sensors are in place. To have accurate measures, the data is converted through their motion fusion algorithms and then fitted to a biomechanical model of the human body, in order to have accurate measurements, whether it is a small vibration or a high dynamic movement. Another advantage of the X-sens is that the communication is via radio or Wi-fi, meaning that it can be done independently of the external setting. Regarding the software, they also provide their own proprietary motion analysis program for research in sports, ergonomics, motion analysis, human-machine interaction (HMI) or gait analysis.

A commercial suit based on this mechanical motion sensing is the Gypsy7 exo-skeleton, from MetaMotion, launched on 2011. This motion capture system offered various kinds of advantages like event control, specialized for game developing settings, or haptic feedback, by preventing joint movement, delivering more realistic reactions during tracking. The measurements were all based on the 14 joint sensors spread around the body, with a resolution of ±0.125 degrees. The wireless system also allowed for an extended range of motion capture with reduced latency. The data extracted was then analyzed on the company software to perform data cleaning and create a correct stick figure for the actor. Although outdated and discontinued, it is a proof of concept to have a system that can estimate the body position based solely on mechanical sensors.

The Vicon hardware has been improved to provide various advantages make the acquisition process smoother. The main focus of enhancement was on the camera systems building specification to adapt to any environment. According to the application, the camera offers a narrower field of view to highly increase fidelity, up to 16.8 MegaPixels (4 times the normal mode) while tracking high-velocity subjects such as drones and athletes; or the camera can be set to high-speed mode, increasing frame rate up to 500 frames per second (also approximately 4 times the usual 120 FPS in standard settings), while maintaining the same FoV, but reducing selective pixel count. This translates to a claimed accuracy when compared to a reference value of 0.017mm (mean), a fifth of the width of a human hair. Besides improving tracking accuracy, the cameras were also perfected to have bump detection, in other words, the cameras have built-in inertial sensors, all around, to monitor the camera position and rotation, in order to detect if it needs a recalibration, and in case it does, it resets to the correct position without operator input.

Virtual simulations based on motion capture are a relatively novel technology, however one of the most well-known companies that pioneered this field, approximately at the same time as Vicon, is Qualisys. Their system is a well-recognized high fidelity MoCap used in diverse applications, from human biomechanics, sports or engineering. The motion capture achieved is of an extreme accuracy, able to achieve a resolution of 26 Megapixels at 290 frames per second. The strobe is created by 24 high-power LEDs, allowing to detect a passive marker as far as 50 meters away. Contrary to the Vicon, Qualisys intends to be used in any environment, both inside or outside under direct sunlight, due to the use of an embedded Active filtering technology plus a "Sun" filter together with the high intensity IR strobe. Besides this, the cameras also have a special protective housing to make them waterproof, dustproof and resistant to an extended temperature range without damaging the accuracy. Furthermore, the cameras also have extra perks to help management, such as, less cabling and an integrated quick release plate. This set of advantages enables behavioral studies, like gait analysis, on patients with medical disabilities, like cerebral palsy, stroke or Parkinson's disease.

A.5 Complete Step-by-Step Guide of DeepLabCut

This series of steps follow figure 3.2 and intend to support the summarized description given in 3.2.

A.5.1 Create a project — Steps 1–5

DeepLabCut is a software based on python that can be acquired in `https://github.com/DeepLabCut/DeepLabCut` where all the installation process and related information is available.

After defining which type of project best suits the application, the training database, *i.e.* the videos from which the frames are going to be extracted, to then be labeled, needs to be specified. The benchmarks behind DLC give diverse suggestions to enhance the computational time, as well as, to increase the probability of having good results in the end of pipeline. First, is fundamental that the resolution and aspect ratio of the videos used in the training step is the same as the ones given in the testing phase (where the predictions will be made). The lower the resolution the faster the computational times, however it is still very important for the key points intended to be tracked are clearly identifiable. From the "Zoo Models" library examples, the recommendation to get good results is for the videos to have between 300–600 pixels on one edge.

In case the subject only occupies a small area of the frame, when the resolution of the video is downsampled, the subject may get blurry making its body parts indistinguishable. In that case, it is also recommended to apply an adptative cropping centered on the subject intended to track. DLC offers a tool during the labeling stage to perform the cropping, however this may bring disadvantages as will be discussed.

Finally, after the decision of which project to create and making sure that the video is properly formatted to be used, a project can be created. A configuration file is automatically created to

hold the characteristics of the project, such as, which body parts are intended to be tracked, how to connect them to make a skeleton, the neural network to build the prediction model and the probability threshold to consider that a point is predicting correctly a body part.

With the project correctly initialized, the next step is to select the frames from each of the videos in the training database.

A.5.2 Frame Selection and Labeling — Steps 6–9

Once the training videos are inputted, the frames that are going to be labeled need to be selected and extracted from these videos. The number of frames to be selected is arbitrary and depends on the application, the complexity of the video and on how many key points are going to be tracked.

After defining the number of extracted frames, DeepLabCut offers 3 methods to select which frames to extract. An important aspect is that the labeled frames need to reflect the variability of experiments, *i.e*, needs to show the situations that represent both the standard behavior that happens in the majority of the video, but also, the more sparse situations, such as, occlusions, extreme points of movement, sudden changes of camera angle or lighting, or simply when two body parts change their typical path, for example, if they get closer.

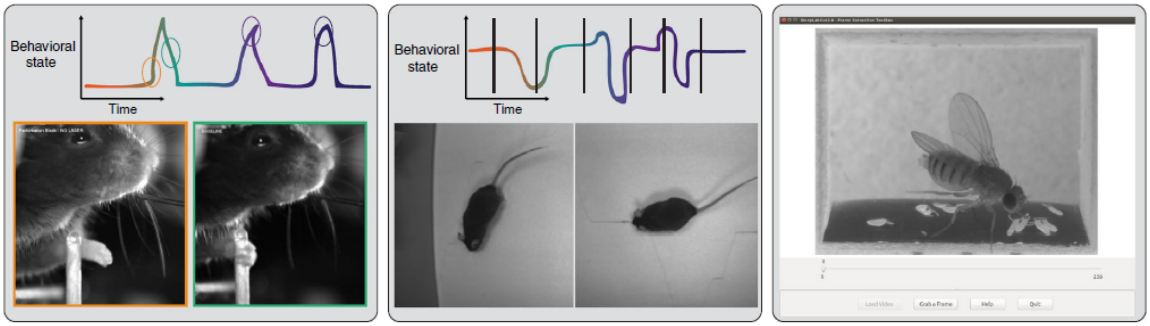

Figure A.3: Methods for frame selection. The toolbox contains three methods for extracting frames, namely, by clustering based on visual content (left image) , by randomly sampling in a uniform way across time (center image) , or by manually grabbing frames of interest using a custom GUI (right image).

The first method is to manually select which frames better describe the normal behavior, as well as which frames include more atypical movements. This process may take some time, since the whole video needs to be inspected. Therefore, an easy solution to automate this process, is simply to extract the frames in a temporal uniform distributed way (since the number of frames to be obtained was defined previously). This raises a problem regarding the interest of the frames picked, because they may demonstrate very similar poses, instead of the desired variety. Finally, the last method is based on K-means clustering, where the algorithm downsamples the video and clusters the frames. Frames from different clusters are then selected. This procedure ensures that the frames look different and is generally preferable, although it has the downside of being more computationally demanding, requiring more time to select the frames, especially if the videos are long.

After this step, DeepLabCut offers a graphical-user interface (GUI) to import the frames. Each frame is presented and the list of body parts defined in the configuration file (pints of interest) are shown to be pinned in the correct place of the frame. Each label/point has a unique color representative of that body part. In this step, is important to be consistent in the marking of the points, for instance, in a broad are like the shoulder, is important to always label the same spot. In case, a body part is occluded it is also recommended not to label it.

When the labeling of all the frames is concluded, the training set is prepared and can be now used as the input for the deep learning models to do the classification the location of each body part.

A.5.3 Neural network training and Evaluation - Steps 10-12

Regarding the data, before entering the training phase, it needs to be split between train and test sets. This division is specified by default by DeepLabCut as 95% - 5% (based on the benchamarks in [34, 30]), because the program needs most of the data to generalize the estimation of the multiple poses, and the test is simply to calculate some evaluation metrics, as will be seen ahead. An important process before heading to the application of the model is data augmentation. This technique introduces transformations in the images, while maintaining the labels in the correct position. The altered images create more variability and flexibility in data, increasing the generalization ability of the models, while preventing the model from overfitting to the data. The transformations induced in the images are summarized in figure A.4. In DeepLabCut, the augmentation is done online while training, which means that in each iteration/batch, the images have a certain probability to be transformed (usually, set to default, as 50%). The transformation applied varies randomly, however, this can also be specified in the project's configuration file.

After the training set is ready, a few hyper-parameters need to be defined, mostly related to computation time and efficiency, such as batch size, number of iterations, and learning rate. The batch size describes the number of images that are fed to the neural network, in each iteration. This is limited by the GPU memory available when running the model. Hence, a smaller batch size would require a higher number of iterations, for the model to train on all the images enough times, to be able to generalize for each pose. Consequently, this would take more time for the model's loss to stabilize. The learning rate is defined by default, according to the model selected to train the project, and adapted according to the number of iterations.

In fact, all these parameters are designed to enhance the performance of the model, in order for it to fulfill its ultimate objective to predict, using the images and the correspondent labeled coordinates, the position of the body parts in the new images/videos. To do so, the most advised architectures are Convolutional Neural Networks (CNN). These are deep learning algorithms which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to classify one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. The algorithm uses filters or kernels with specific configurations to downsample the image according to a certain pattern. This leads to a reduction of the number of input parameters, making it easier to process, without loosing relevant

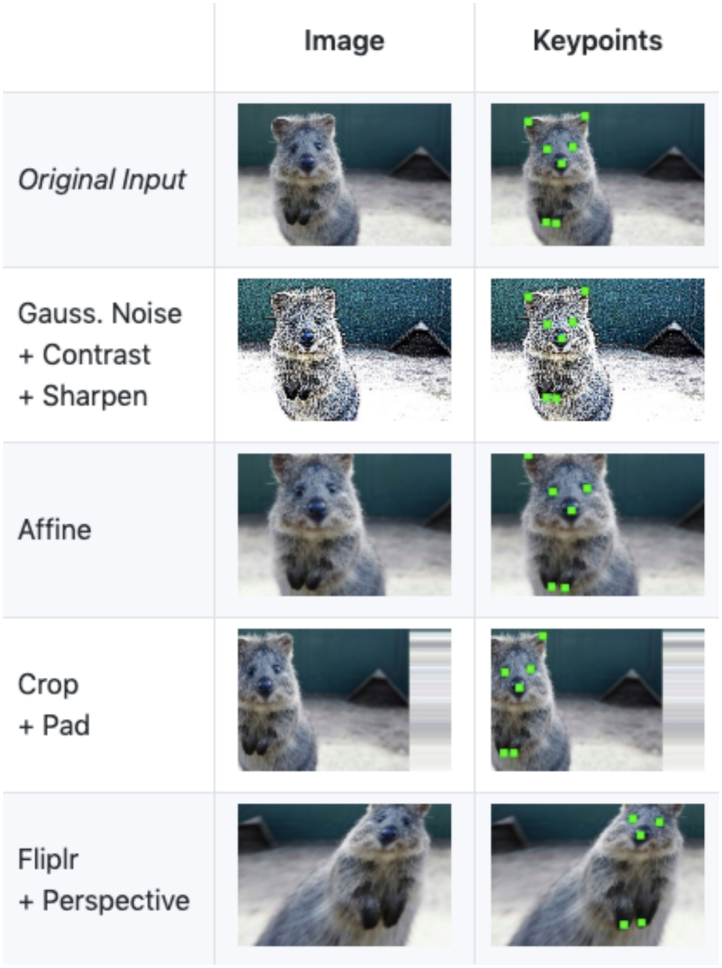

Figure A.4: Ensemble of image transformation examples done for data augmentation.

features. The network is then able to successfully capture the spatial and temporal information and recognize the content of the image. This type of networks are usually applied on classification or recognition of full images, however in the application of DeepLabCut, the objective is to identify the position of the body parts. To do so, the last classification layer of a typical architecture is replaced with deconvolutional layers, one per each desired body part. Each of this layers will up-sample the visual information and produce spatial probability densities, *i.e.* a scoremap where the area where the keypoint of interest is most probably located is highlighted. This also translates into a confidence estimator of the prediction, for instance, if the scoremap is broad and not very intense, the likelihood would be smaller than if the scoremap is more precise and compact. For a better visualization, a log transform is applied to the scoremap, as can be seen in figure A.5.

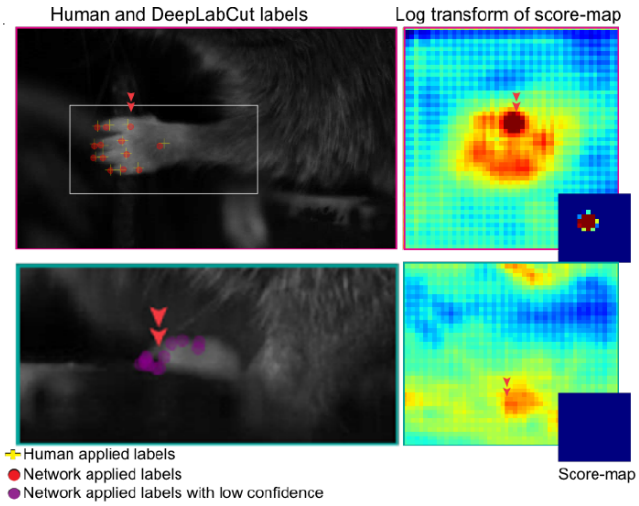

Figure A.5: On the left, 2 example frames of a project intended to track a mouse paw while it grasps a tube. The image on top, the hand is clearly visible, so the model can identify accurately where joints should be. If a keypoint is selected (double arrow) the correspondent score map is the small square with the dark blue background, since the likelihood of the body part is low in all of the image, except in the red region. If a log-transform is applied and the red spot is zoomed, can clearly be see an even smaller region with higher intensity, that results in predicted position of the body part (red dot). In the bottom image, the paw was tilted sideways, so the model cannot identify any joint, even though one of the fingers is still visible (double arrow). This translates to a scoremap of low confidence, that when transformed has a broader region without any specific intense spot, leading to bad predictions (purple dots). [34]

An important step before starting to consider the neural networks that DeepLabCut cut has to offer, is the paradigm of transfer learning, where a network that has been previously trained on a certain animal, could serve as the basis for a new project based in the same species. Thus, before starting the project, is important to check if there is a pre-trained model that can be used to initialize the weights of the new model. This method has been proven to give very satisfying results in DLC as stated in [40, 33]. As mentioned before, DeepLabCut has the "Zoo Models" library, where official pre-trained projects can be exported and be foundation to new research (this can only be achieved in single animal projects, since the "multi-animal" update is very recent). Another

possibility to use transfer learning is to build upon previous iterations on the same project, for example, if a new model wants to be computed after new labeled data is added. The only setback of transfer learning is that it limits the type of neural networks, since the weight configuration of the new model needs to be the same as the pre-trained.

Therefore, when starting a novel project is important to take into consideration the most ideal type of network for your application. As mentioned before, in the DLC workflow, the model is defined in the first steps of the project creation, so it is important to give an overview of all the models offered in DeepLabCut and what are the advantages and disadvantages of each. In DLC, the most important trade-off to be taken into consideration is performance *vs* speed. In case the data is collected and the analysis is done offline, without any time restriction, while focusing on having robust models with high performance, the best network type are Residual Neural Networks — ResNets [34]. The "single animal" projects includes 3 types of architectures, the ResNet50, ResNet101 and ResNet152. In the "multi-animal", a new network called DLCRNet was developed following the same concepts behind the ResNet, however they divided it into 2 parts, firstly a multi-fusion module, where the input goes through parallel convolutions with different steps and "conv" sizes, that are rearranged in order to be fused together; and secondly a multi-stage decoder, where the deconvolutional layers not only upsample the result of the previous fusion to calculate the scoremaps, but also calculates image features and "part affinity fields" (method to identify to which individual displayed in the image does the body part belong to) which then are used to compute a "local refinement field" to predict the position of the body parts in the correspondent subject. More details of the methodology behind the computation of this models can be found in [30]. Concerning, fast response applications, that are more suitable for dynamic environments, where time efficiency is a requirement, DLC has implemented MobileNetsV2. These models were created to be ideal for mobile applications, so the network is less intricate, which results in a drop of overall accuracy, but also inference time. The type MobileNetsV2 network varies according to the value of a hyperparameter alpha, that controls the width of the network, *i.e.* the number of filters applied in each layer. DLC has incorporated 4 different versions (alpha being 0.35, 0.50, 0.75, 1.00). Besides this 2 main back bone architectures, DLC also integrates a type of architecture that is both high in performance and low in inference time, based on the EfficientNet [34, 35] . This model has a scaling method to uniformly define the network's width, depth, and resolution with a set of fixed scaling coefficients, that when they are given the same magnitude can be defined by the hyperparameter Theta. In DLC, this coefficients varies from 0 to 7, originating different models that can even achieve better results, in case of the maDLC, than the DLCRNet and the ResNet depending on the dataset used. A benchmark to compare the models architectures both regarding inference time and accuracy can be found in picture A.6.

At this stage, the training set can start, where a batch of images are inputted in the model in each iteration, and the loss metrics are computed based on the coordinates. In this step, the saDLC simply presents a cross-entropy loss metric, however the maDLC calculates 4 different kinds of loss, namely, cross-entropy loss metric, scoremap loss, location refinement and limb loss. The model trains for the number of iterations set in function's hyperparameters, however it saves

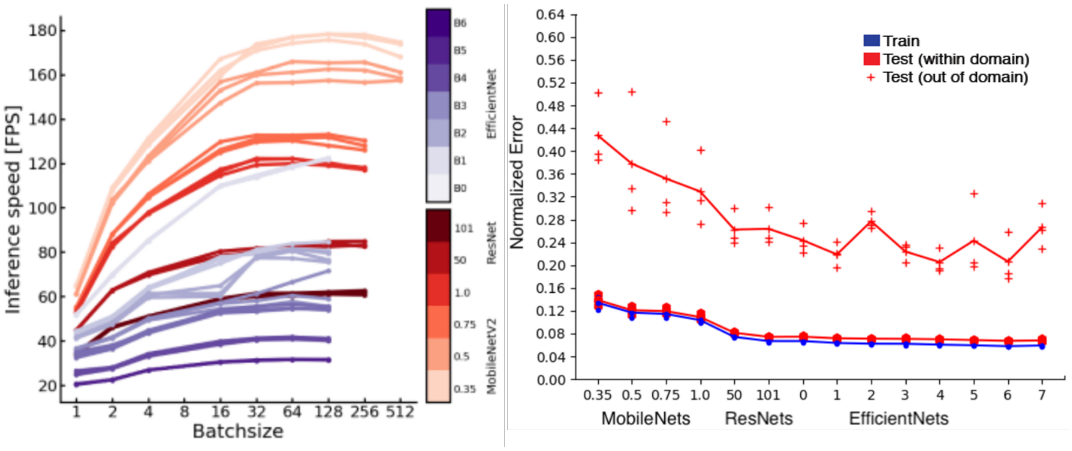

Figure A.6: Both benchmark plots derive from the analysis of the same dataset presented in [33], where more details can be found. On the left, the different models are compared regarding there inference speed, being the MobileNetsV2 faster then ResNets. The EfficinetNets, depending on the depth of the architecture can vary substantially. On the left, the error was measured for the same models. The deeper the network, the better the accuracy, in the case of the MobileNetsV2 and ResNet. In case of the EfficientNet, the faster one has the biggest error between them, followed by the 5. The 6 presents the lowest error of the overall architectures.

periodic snapshots of the weights.

When all the iterations are complete, the evaluation metrics can be computed. The multiple snapshots are useful to compare the accuracy of the model during the training, by computing the root mean square error between the prediction and groundtruth label, based on the euclidean distance (in pixels), as presented in A.1.

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{n}(x_{i_{label}} - x_{i_{pred}})^2 + (y_{i_{label}} - y_{i_{pred}})^2} \tag{A.1}$$

In case of overfitting, the model with more iterations would present a close to perfection prediction in the training set, but lower accuracy in the test set, when compared to the performance of models with less iterations. This should be taken into account, and the snapshot that proves to have the lower test error should be used for detecting the body parts in the new videos. In this step, it is already possible to grasp if the model was able to generalize correctly the positions of the body parts, because if the RMSE should be low enough to be within the margin of error that 2 different researchers labeling data would do. If the pixel error is out of proportions, then the more data should be labeled or a new model should be applied. However, is only possible to truly understand the prediction error and inconsistency, when evaluating a new video, where thousands of new frames are given

A.5.4 Video Analysis — Steps 13–15

In terms of the DLC workflow, this step is straightforward, the videos need simply to be given to the system, and the DLC analyses it, outputting a data file with the coordinates of all body parts, in all the frames of the video. In addition, it also allows to create a new video with same format as the original with the the keypoints displayed, based on the prediction file, as well as, a combination of plots that describe pose estimation, such as the likelihood of the prediction being correct and the trajectory of each body part, as can be seen in figure A.7.

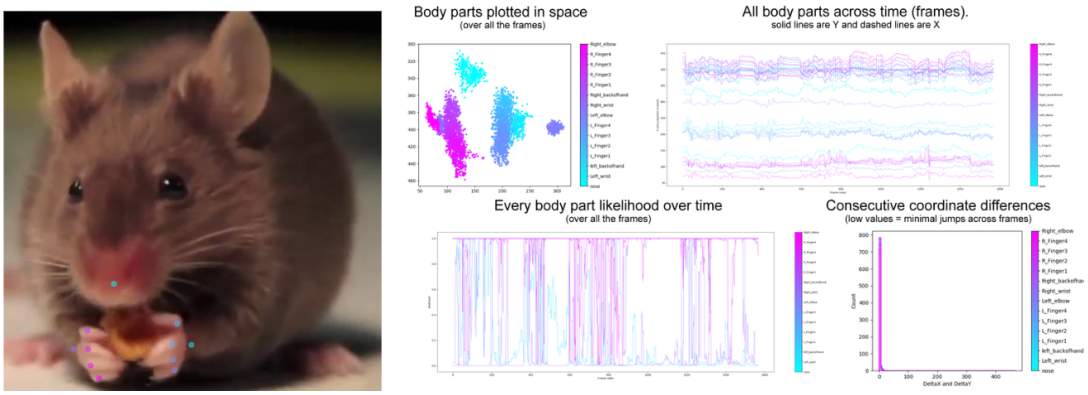

Figure A.7: On the right, an example of a frame after the video analyzes, displaying the tracked keypoints. On the left, a combination of plots that help understand the analysis. On the top left, each color represents the trajectory of a certain body part, during the video, allowing to understand what area was explored. On the top right, a representation of how the body part moved in the X and Y axis, over time. On the bottom left, is represented the likelihood that a body part is identified in the correct position, over time. Ideally, all body parts would have a straight line on 1, which would mean that it was always identified confidently. On the bottom right, a measure of how much a body part moved in consecutive frames. Most of the values are centered around 0 to 20, which means that the body moved. However, if peaks appear could be an indication of miss predictions [34, 40]

After analysing the plots, is possible to have an idea of the behaviour of the subject, with the top plots and how successful the prediction process was with the bottom ones, however the true way to verify it if is well tracked is to review the video and compare the keypoint with the position of the actual subject. A good tracking presents a smooth movement accompanying the subject, however this is not the case most of the times. DeepLabCut is highly sensible to changes in lighting and angle, as well as, to movements that did not appear in the labeled dataset. This results in various issues, such as, total miss identifications, whether DLC does not track the body part at all, even though it is present in the frame, or if it does not recognize it as the body part it should be, marking the keypoint in other limb. Usually, in the latter, the keypoint changes between the true and the incorrect body part, which leads to burst in the predictions. This can clearly be seen in the difference of consecutive predictions plot. Another very common problem in predictions is jitter, *i.e.*, a small vibration like movement of the keypoint, that highly reduces the tracking precision, specially noticeable if a body part is stopped. This issue precludes the tracking of small movements, since they are indistinguishable from the jitter noise.

If all these issues appear after the analysis, the results are not satisfactory, and the labeled data is clearly not enough. Therefore, a second iteration of the pipeline will be needed where new refined data is added.

A.5.5 Active Learning — Steps 16–18

Upon bad results in the video analysis process, there are a few steps that can be done to improve the tracking quality. This steps are designated as active learning, since the user is the one that needs to correct the miss predictions and add new labeled data, to then reiterate on top of the same model.

Evaluating each situation, firstly, in case the subject is only occupying a small amount of the frame, even if the predictions do not have a lot of jitter or bursts, it will create a big impact on the accuracy of tracking. To solve this, a new project should be created where an dynamic cropping on the training data set is applied, in order to zoom in the subject, making its limbs more evident, as well as their movement. If a new model is trained, even tho the magnitude of the jitter is probably similar, it would have a smaller impact on the quality of the tracking, since its accuracy has increased.

However, if the tracking has a lot of miss predictions or confusion in the identification of the different body parts, then more labeled data is needed to optimize the model's performance. For generalization to large data sets, images with insufficient labeling performance can be extracted, manually corrected by adjusting the labels to increase the training set and iteratively improve the feature detectors. One does not necessarily need to correct all errors, as common miss predictions could be eliminated by relabeling a few examples and then re-training. *A priori*, given that there is no ground truth data for the new analyzed videos, it is challenging to find putative "outlier frames". However, one can use heuristics such as the continuity of body part trajectories, to identify images where the decoder might make large errors. To select the outliers, DLC presents various algorithms, for instance, based on the confidence estimator of each prediction, if it his below a certain threshold then is a miss prediction; in case the keypoint jumps more than a certain number of pixels in consecutive frames; or by using a Auto Regressive Integrated Moving Average (ARIMA) model in the time series of the trajectory of the body part, is also possible to identify if a certain frame as an uncommon behaviour [40]. Besides this method, DLC also has a GUI to rewatch the videos and select manually the frames with miss predictions.

After the selection process, the feedback loop starts, where these frames can be imported in the labeling GUI, to be correctly labeled to augment the training set. Note, that here the model trained the first time can be used to perform weight initiation, reducing the training iterations on the new model (since the corrected outlier frames were added to the frames used on the first training, hence both datasets were merged).

With this refinement process, the analysis of the videos should not only be more accurate in recognizing the position of the body parts, but also more precise (less jitter).

Finally, with the high-performance tracked videos, the behaviour and poses of the subject can be studied, in order to understand if certain patterns of movement surface when the individual is in a certain condition.

A.5.6 Tables of the statistical results

The tables presented in the next sections follow the same organization as the plots presented in the results section of each of the studies. As can be seen, independently of the study and of the body part or motion feature studied, the interaction between group and synchronicity was never significant. The statistical test applied was a 2-Way ANOVA test, with no parametric correction.

A.5.6.1 Study A - Motion features of the finger

Table A.2: Statistical results of the motion features of the moving finger

Frequency of Pokes HC vs PD

Effect	F	p
Group	3.607	0.063
Condition	14.988	< .001
Group ✳ Condition	1.702	0.198

Finger front-back amplitude HC vs PD

Effect	F	p
Group	0.502	0.482
Condition	2.148	0.149
Group ✳ Condition	0.281	0.598

Finger front-back speed HC vs PD

Effect	F	p
Group	5.375	0.025
Condition	0.661	0.420
Group ✳ Condition	0.373	0.544

Frequency of Pokes PD-PH vs PD-nPH

Effect	F	p
Group	0.172	0.681
Condition	13.752	< .001
Group ✳ Condition	0.012	0.915

Finger front-back amplitude PD-PH vs PD-nPH

Effect	F	p
Group	9.677	0.004
Condition	4.499	0.042
Group ✳ Condition	2.542	0.121

Finger front-back speed PD-PH vs PD-nPH

Effect	F	p
Group	8.250	0.007
Condition	1.307	0.261
Group ✳ Condition	0.087	0.770

The top row is the HC VS PD and the bottom the PD-PH VS PD-nPH. From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When p < 0.05 is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05).

A.5.6.2 Study A - Motion features of the elbow and shoulder

A.5.6.3 Study B - Motion features of the finger

A.5.6.4 Study B - Motion features of the elbow and shoulder

Table A.3: Statistical results of the motion features of the elbow and shoulder

Elbow front-back amplitude HC vs PD

Effect	F	p
Group	3.352	0.073
Condition	0.008	0.928
Group ✳ Condition	0.041	0.840

Elbow front-back speed HC vs PD

Effect	F	p
Group	3.626	0.063
Condition	0.142	0.708
Group ✳ Condition	0.275	0.603

Shoulder front-back amplitude HC vs PD

Effect	F	p
Group	2.405	0.141
Condition	2.244	0.127
Group ✳ Condition	0.003	0.959

Shoulder front-back speed HC vs PD

Effect	F	p
Group	0.091	0.027
Condition	5.192	0.764
Group ✳ Condition	0.127	0.724

Elbow front-back amplitude PD-PH vs PD-nPH

Effect	F	p
Group	6.635	0.015
Condition	0.001	0.974
Group ✳ Condition	0.002	0.964

Elbow front-back speed PD-PH vs PD-nPH

Effect	F	p
Group	4.780	0.036
Condition	0.613	0.439
Group ✳ Condition	0.282	0.599

Shoulder front-back amplitude PD-PH vs PD-nPH

Effect	F	p
Group	11.670	0.002
Condition	2.969	0.095
Group ✳ Condition	1.354	0.253

Shoulder front-back speed PD-PH vs PD-nPH

Effect	F	p
Group	6.553	0.015
Condition	0.271	0.606
Group ✳ Condition	0.048	0.828

The elbow is presented on the left and the shoulder on the right. The top row is the HC VS PD and the bottom the PD-PH VS PD-nPH. The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05).

Table A.4: Statistical results of the motion features of the moving finger

Frequency of Pokes HC vs PD

Effect	F	p
Group	3.774	0.057
Condition	8.902	0.004
Group ✳ Condition	0.481	0.491

Finger front-back amplitude HC vs PD

Effect	F	p
Group	0.536	0.467
Condition	0.001	0.970
Group ✳ Condition	0.099	0.754

Finger front-back speed HC vs PD

Effect	F	p
Group	0.106	0.745
Condition	3.858	0.054
Group ✳ Condition	0.690	0.410

Frequency of Pokes PD-PH vs PD-nPH

Effect	F	p
Group	0.769	0.388
Condition	8.665	0.007
Group ✳ Condition	2.023e-4	0.989

Finger front-back amplitude PD-PH vs PD-nPH

Effect	F	p
Group	0.994	0.328
Condition	1.849e-5	0.997
Group ✳ Condition	0.351	0.559

Finger front-back speed PD-PH vs PD-nPH

Effect	F	p
Group	1.072	0.310
Condition	2.088	0.160
Group ✳ Condition	0.684	0.415

The top row is the HC VS PD and the bottom the PD-PH VS PD-nPH. From left to right, the features analyzed were the frequency of pokes (Hz), amplitude of movement and speed. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05).

Table A.5: Statistical results of the motion features of the elbow and shoulder

Frequency of Pokes HC vs PD

Effect	F	p
Group	3.774	0.057
Condition	8.902	0.004
Group ✳ Condition	0.481	0.491

Finger front-back amplitude HC vs PD

Effect	F	p
Group	0.536	0.467
Condition	0.001	0.970
Group ✳ Condition	0.099	0.754

Finger front-back speed HC vs PD

Effect	F	p
Group	0.106	0.745
Condition	3.858	0.054
Group ✳ Condition	0.690	0.410

Frequency of Pokes PD-PH vs PD-nPH

Effect	F	p
Group	0.769	0.388
Condition	8.665	0.007
Group ✳ Condition	2.023e-4	0.989

Finger front-back amplitude PD-PH vs PD-nPH

Effect	F	p
Group	0.994	0.328
Condition	1.849e-5	0.997
Group ✳ Condition	0.351	0.559

Finger front-back speed PD-PH vs PD-nPH

Effect	F	p
Group	1.072	0.310
Condition	2.088	0.160
Group ✳ Condition	0.684	0.415

The elbow is presented on the left and the shoulder on the right. The top row is the HC VS PD and the bottom the PD-PH VS PD-nPH. The features analyzed were the amplitude of movement and speed for both body parts. When $p < 0.05$ is considered a main effect ('***' - <0.001 | '**' - 0.01 | '*' - 0.05).