UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Activity recognition and bioinspired approaches for robotics in intelligent environments

Caetano Mazzoni Ranieri

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Caetano Mazzoni Ranieri

Activity recognition and bioinspired approaches for robotics in intelligent environments

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Roseli Aparecida Francelin Romero
Co-advisor: Profa. Dra. Patrícia Amâncio Vargas

USP – São Carlos
July 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R197a

Ranieri, Caetano Mazzoni
Activity recognition and bioinspired approaches for robotics in intelligent environments / Caetano Mazzoni Ranieri; orientadora Roseli Aparecida Francelin Romero; coorientadora Patrícia Amâncio Vargas. -- São Carlos, 2021.
152 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. INTELIGÊNCIA ARTIFICIAL. 2. REDES NEURAIS. 3. ROBÓTICA. 4. COMPUTAÇÃO BIOINSPIRADA. I. Romero, Roseli Aparecida Francelin, orient. II. Vargas, Patrícia Amâncio, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Caetano Mazzoni Ranieri

Reconhecimento de atividades e abordagens bioinspiradas para robótica em ambientes inteligentes

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Roseli Aparecida Francelin Romero
Co-orientadora: Profa. Dra. Patrícia Amâncio Vargas

USP – São Carlos
Julho de 2021

In memory of the COVID-19 victims.

ACKNOWLEDGEMENTS

This work was funded by the Sao Paulo Research Foundation (FAPESP), grants 2017/02377-5, 2018/25902-0 and 2017/01687-0. The project also counted on a two months support by the Coordination for the Improvement of Higher Education Personnel (CAPES), grant PROEX-7153000/D, and other two months by the National Council for Scientific and Technological Development (CNPq), grant 140161/2017-1.

The research was carried out using the computational resources of the Centre for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP, grant 2013/07375-0, and the Robotics Lab within the Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, United Kingdom. Additional resources were provided by the Nvidia Grants program, and the Neuro4PD project, funded by Royal Society and Newton Fund (NAF\R2\180773).

After the institutional thanks, I proceed with those who directly made possible the development of this work. First, I thank my supervisor, professor Roseli Aparecida Francelin Romero, for the guidance and the readiness for solving any academic issues that arrived during this Doctorate. Second, my co-supervisor, professor Patricia Amancio Vargas, for receiving me at the Heriot-Watt University, UK, and contributing to this work with valuable ideas regarding the neurorobotics aspects of the work. Third, my professors and administrative staff of the Institute of Mathematical and Computer Sciences at the University of Sao Paulo (ICMC-USP), which provided infrastructure and administrative support. Within this context, I also thank the staff of the Heriot-Watt University, in which I spent a 10-months internship, with resources by FAPESP. Finally, I thank the examination board, which kindly agree in revising and assessing this work.

Other researchers provided important guidance for this project, especially Renan Cipriano Moioli, for all aspects of this work that involved neuroscience and signal processing for neurophysiological data; Mauro Dragone, for providing resources for the collection of the HWU-USP dataset, and guidance for analyses resulting from that; and Vitor Campanholo Guizilini, for the introduction to deep learning research.

The researchers that carried out their research at the Robots Learning Laboratory (LAR), and related laboratories at our room at ICMC-USP, were of paramount importance. Special thanks are addressed to those who worked with me, providing direct contributions to this work: Daniel Carnieto Tozadore, Guilherme Vicentim Nardari, Murillo Rehder Batista, and Adam Henrique Moreira Pinto. Within this same context, I thank the participants of the Warthog Robotics team, especially Gustavo Stefano and Guilherme Acra, for contributions regarding the LARa robot and software, and the colleagues at Heriot-Watt University, especially Jhielson

Montino Pimentel, for important contributions during the course of my internship at the Heriot-Watt University. Scott MacLeod also contributed directly to this thesis, by participating on the data collection for the HWU-USP activities dataset. Hugo Sardinha and Siobhan Yasmin Duncan also contributed to the outcomes of this work.

For the previous steps, which took place during the scientific initiation and Master's projects that I have developed before the beginning of this Doctorate and have prepared the foundations of my research abilities, I thank Humberto Ferasoli Filho and Silas Franco dos Reis Alves, which are responsible for important steps towards my scientific maturity, as well as all my undergraduate professors at the Sao Paulo State University (UNESP).

Very special thanks are addressed to my family,for all support that made it possible for my academic career from the very beginning, especially my father, Servio Tulio Vieira Ranieri, my mother, Vera Lucia Andrade Mazzoni; and my brothers, Jhonathan Mazzoni Busato and Erick Andrade Busato. In the same context, I thank my girlfried, Aline Cristina Marques, for all support during the last three years. Last, but not least, I would like to thank all my friends, who directly or indirectly contributed to the completion of this work.

“A process cannot be understood by stopping it.
Understanding must move with the flow of the process,
must join it and flow with it.” (Frank Herbert)

RESUMO

RANIERI, C. M. **Reconhecimento de atividades e abordagens bioinspiradas para robótica em ambientes inteligentes**. 2021. 152 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Projetos de automação residencial têm sido desenvolvidos há algum tempo, tendo evoluído para os chamados ambientes inteligentes. Esses ambientes são caracterizados pela presença de conjuntos de sensores e atuadores, conectados de forma a responder adequada e proativamente a diferentes situações. A integração de ambientes inteligentes com robôs permite a introdução de capacidades adicionais de sensoriamento, além da realização de tarefas com maior flexibilidade e menor complexidade mecânica do que os robôs monolíticos tradicionais. Para dotar tais ambientes de comportamentos verdadeiramente autônomos, algoritmos devem extrair informações semanticamente significativas de quaisquer dados sensoriais disponíveis. Reconhecimento de atividade humana é um dos campos de pesquisa mais ativos dentro deste contexto. Neste projeto, foi abordado o projeto e avaliação de técnicas de aprendizado para reconhecimento da atividade humana, considerando diferentes modalidades de sensores. Dois tipos de redes neurais, baseadas em combinações de Redes Neurais Convolucionais com Redes Recorrentes com Memória de Curto e Longo Prazo ou Redes Convolucionais Temporais, foram propostas e avaliadas em duas bases de dados públicas para reconhecimento de atividade multimodal de vídeos e sensores inerciais. A estrutura resultante foi então empregada a um novo conjunto de dados, o HWU-USP activities dataset, coletado como parte deste trabalho, em um ambiente real dotado de vídeos, unidades inerciais e sensores ambientais. Foi avaliada a influência dos sensores ambientais, sincronizados aos dados inerciais e de vídeo, na acurácia dos resultados, tendo se mostrado uma abordagem promissora. Além disso, o novo conjunto de dados foi provido de atividades complexas com dependências de longo prazo, avaliadas por meio de classificadores baseados em segmentos de comprimento limitado, simulando os resultados para aplicações de tempo real. Em um segundo momento, foram desenvolvidos trabalhos sobre dados neurofisiológicos de primatas induzidos à doença de Parkinson, indo de análises e classificação dos dados, com uso de redes neurais, até a construção de um modelo computacional das estruturas acometidas dentro do cérebro. Embora distinta dos estudos sobre reconhecimento de atividades e tecnologias assistivas, focos desta tese, esses trabalhos foram relacionados na natureza das técnicas empregadas, e seus resultados fizeram parte do cenário de aplicação desenvolvido em seguida. Por fim, foi projetado e implementado um cenário de aplicação na forma de simulação robótica, de modo que o módulo desenvolvido pudesse ser avaliado em situações práticas. Para o mecanismo de seleção de comportamento, uma abordagem bioinspirada baseada em modelos computacionais do circuito núcleos da base-tálamo-córtex foi avaliada e comparada a abordagens não bioinspiradas baseadas em heurísticas simples.

Palavras-chave: reconhecimento de atividade humana, base de dados de atividades, aprendizado profundo, modelo computacional bioinspirado, neurorrobótica.

ABSTRACT

RANIERI, C. M. **Activity recognition and bioinspired approaches for robotics in intelligent environments**. 2021. 152 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Home automation projects have been developed for some time, having evolved into the so-called smart environments. These environments are characterised by the presence of sets of sensors and actuators, connected in order to respond appropriately and proactively to different situations. The integration of intelligent environments with robots allows for the introduction of additional sensing capabilities, besides performing tasks with greater flexibility and less mechanical complexity than traditional monolithic robots. To endow such environments with truly autonomous behaviours, algorithms must extract semantically meaningful information from whichever sensor data is available. Human activity recognition is one of the most active fields of research within this context. In this project, the design and evaluations of learning techniques for human activity recognition was addressed, considering different sensor modalities. Two types of neural networks, based on combinations of Convolutional Neural Networks to Recurrent Networks with Long Short-Term Memory or Temporal Convolutional Networks, were proposed and evaluated on two public datasets for multimodal activity recognition from videos and inertial sensors. The resulting framework was then introduced to a new dataset, the HWU-USP activities dataset, collected as part of this work, in an actual environment endowed with videos, inertial units, and ambient sensors. This design allowed for assessing the influence of ambient sensors, synchronised to the inertial and video data, to the accuracy of the results, which has proven to be a promising approach. Also, the new dataset provided complex activities with long-term dependencies, evaluated through segment-wise classifiers simulating the results for real-time applications. In a second moment, works were developed on neurophysiological data from primates induced to Parkinson's disease. Those studies ranged from data analysis and classification, using neural networks, to the construction of a computational model of the affected structures within the brain. Although different from the studies on activity recognition and assistive technologies, which were the focus of this thesis, these works were related in the nature of the techniques used, and their results were part of the application scenario developed next. Finally, an application scenario was designed and implemented as a robot simulation, so that the developed module could be evaluated in practical situations. For the behaviour selection mechanism, a bioinspired approach based on computational models of the basal ganglia-thalamus-cortex circuit was evaluated and compared to non-bioinspired approaches based on simple heuristics.

Keywords: human activity recognition, activities dataset, deep learning, bioinspired computa-

tional model, neurorobotics.

CONTENTS

1 INTRODUCTION 17
1.1 Hypothesis and Contributions 19
1.2 Objectives 21
1.2.1 Specific Objectives *22*
1.3 Organisation of the Work 22

2 MODELS FOR HUMAN ACTIVITY RECOGNITION 25

3 HWU-USP DATASET AND FRAMEWORK 35

4 APPROACHES ON COMPUTATIONAL NEUROSCIENCE 69

5 APPLICATION SCENARIO 105

6 CONCLUSION 127
6.1 Final Considerations 128
6.2 Future Research 132
6.2.1 Activity Recognition Methods *132*
6.2.2 Application Scenarios *133*
6.2.3 Neurorobotics *134*

BIBLIOGRAPHY 135

APPENDIX A ETHICS CLEARANCE FOR THE HWU-USP ACTIVITIES DATASET 141

APPENDIX B PUBLICATIONS 147

APPENDIX C OTHER ACTIVITIES 151

CHAPTER

1

INTRODUCTION

It has been an established fact that the world population is ageing, with a progressively larger proportion of elderly people with respect to younger demographic strata. According to projections by the Department of Economic and Social Affairs of the United Nations, the worldwide proportion of citizens aged between 15 and 64, with respect to those aged over 65 years old, is expected to drop from about 7:1 in 2020 to approximately 4:1 in 2050 (United Nations, 2019). Different challenges have been addressed during this transition (MCCANN, 2017; MELO *et al.*, 2017), one of them being the need for alternatives to assist the elderly in face of a decreasing availability of workforce concurrent to an increasing demand for carers, nurses, and other professionals (KHAN, 2019).

The research community on Ambient Assisted Living (AAL) have proposed and evaluated different automatised solutions that address this challenge, with the objective to support the elderly or people with special needs (CALVARESI *et al.*, 2017). Sets of sensors and actuators within an intelligent environment may help its inhabitants by providing services to assist their daily activities (PATEL; SHAH, 2021) or monitoring their health status (SANDEEPA *et al.*, 2020). The field of Human-Robot Interaction (HRI) also presented contributions to these environments. Robots may contribute either by introducing additional capabilities for actuating in the ambient (GOMEZ-DONOSO *et al.*, 2019), or presenting a friendly, more engaging social interface that may increase the acceptance of the assistive technologies by the inhabitants of an intelligent environment (IGLESIAS *et al.*, 2020).

The design of sophisticated AAL solutions, especially when endowed with robotic devices, requires a reasonable understanding, by the automated system, of the current context of the environment with respect to its inhabitants, which might be inferred based on information gathered by different types of sensors. To this aim, research has been performed within the field of human activity recognition (MOJARAD *et al.*, 2018).

According to Chaaraoui, Climent-Pérez and Flórez-Revuelta (2012), an activity is a

sequence of semantically meaningful actions involving interactions between humans and environment, composed of sequences of basic human motions, called action primitives. Different modalities of data from a sensed environment may be used to provide automatic activity recognition. Most benchmarks provided for this task are based on regular videos (KUEHNE *et al.*, 2011; SOOMRO; ZAMIR; SHAH, 2012; CARREIRA *et al.*, 2019), videos from RGB + depth (RGBD) cameras (NI; WANG; MOULIN, 2011; LIU *et al.*, 2019), inertial sensors (REISS; STRICKER, 2012; CHAVARRIAGA *et al.*, 2013), or ambient sensors from smart environments (COOK *et al.*, 2013; BAKAR *et al.*, 2016).

Of particular interest to the purposes of this thesis are the the multimodal benchmarks, which provide more than one modality being recorded simultaneously (CHEN; JAFARI; KEHTARNAVAZ, 2015; SONG *et al.*, 2016). Multimodal scenarios allow not only for the design of more accurate methods (WEI; JAFARI; KEHTARNAVAZ, 2019), but also for the compensation of missing information when a set of modalities is inaccessible. For example, in a situation in which video data cannot be accessed because it relies on the camera of a robot that is not placed at the same room as the user being monitored, information from wearables of ambient sensors may be gathered and analysed to infer his activity.

Even though activity recognition has been a fertile field of research, approaches that connected recognised activities to actual response behaviours from an artificial agent usually consisted of direct associations (GEORGIEVSKI *et al.*, 2017; LI *et al.*, 2019b; LERA *et al.*, 2020), with few quantitative analyses on the quality of the responses. Even within the HRI context, research is usually focused on enhancing the ability of the machine learning predictors employed for classifying the activities under certain conditions (MANZI *et al.*, 2018), rather than evaluating the suitability of the response behaviours from the robot.

Besides providing suitable behaviours for HRI (PETRICK; FOSTER, 2020; FOSTER *et al.*, 2020), behaviour selection mechanisms for autonomous agents may integrate bioinspired architectures. Different approaches have been proposed within this context, one of them being the simulation of neurophysiological properties of living beings. Li *et al.* (2019a) provided a comprehensive survey on neurorobotics systems (NRS), and the different components that may integrate them. According to the authors, a generalised framework can be depicted for most NRSs in the literature, composed of a simulated brain, which is fed with sensory signals from a body and turns them into control signals for a hierarchical controller, responsible for decoding these signals into control commands for the body, which actuates and senses an external environment.

This thesis was centred in the possibilities for human activity recognition in intelligent environments, which has shown to be of paramount importance in the context of AAL. Deep neural networks, techniques that could provide accurate results for different modalities of data (i.e., videos, inertial units or ambient sensors), were designed and evaluated in different scenarios. Combinations of convolutional (CNN) and recurrent neural networks (RNN) were proposed, including a novel combination between a regular CNN and temporal convolutional networks

(TCN). A dataset for daily activities in a smart home was built, which allowed for experiments in a different setting, resembling aspects that might be important for actual real-time systems, and which has not been implemented previously in other datasets. An application scenario was proposed, and decision-making mechanisms were designed to provide responses for a robotic agent in such a home environment. The central hypothesis of this thesis, the research questions addressed, and the contributions made are presented in the next section.

1.1 Hypothesis and Contributions

The central hypothesis of this thesis is as follows:

> It is possible to enhance the understanding of human behaviour and facilitate human-robot interaction that demands rapid and real-time responses by using deep neural network models of human activity recognition in intelligent environments. These deep learning models applied to a plethora of data from synchronised videos, inertial units, and ambient sensors will produce accurate results based on limited-length segments of data through time, hence generating outputs suitable for the fulfilment of long-term robot tasks focused on time-localised decision-making.

In order to validate this hypothesis, three research questions were formulated, and each of them led to one contribution.

These are enumerated as follows:

1. Are convolutional and recurrent neural networks suitable for human activity recognition based on multimodal sensors, particularly videos and inertial units, and can late fusion and feature-level fusion enhance the results, when compared to single-modality models?
2. Considering activities of daily living of different complexity levels, including long-term dependencies between primitive actions, can a deep learning-based framework provide accurate and consistent classification results by applying a multimodal framework which relies on videos, inertial units, and ambient sensors?
3. Can the time-localised outputs of a multimodal activity recogniser be employed to different decision-making mechanisms for social robots in a real-time application scenario, producing a reliable response for completing long-term tasks that rely on the human activities being performed?

Research question 1 was addressed by the first contribution of this thesis: a set of techniques for multimodal activity recognition were designed and evaluated in order to contribute to the state-of-the-art, particularly based on videos and inertial sensors, the most ubiquitous type

of data available currently (LIMA *et al.*, 2019). To this aim, public available datasets (CHEN; JAFARI; KEHTARNAVAZ, 2015; SONG *et al.*, 2016) were gathered, preprocessed and employed as inputs to a deep learning framework that provided feature extraction and classification for each modality separately, which could be fused according to a proposed feature-level fusion or to a late fusion based on the prediction vectors. More specifically, these techniques were based on combinations between Convolutional Neural Networks (CNN) (ZEILER; FERGUS, 2014), Long Short-Term Memory (LSTM) (HOCHREITER; SCHMIDHUBER, 1997), and Temporal Convolutional Networks (TCN) (BAI; KOLTER; KOLTUN, 2018).

Research question 2 was addressed by the second contribution of this thesis: the introduction of data from ambient sensors to this framework, which was made through the proposition of a new dataset, the HWU-USP Activities dataset, collected in the Robotic Assisted Living Testbed (RALT), at Heriot-Watt University (HWU), Edinburgh, Scotland, UK. The resulting dataset was composed of data from: i) the RGBD camera of a social robot (i.e., the TIAGo robot (PAL Robotics, 2017)), ii) inertial units attached to the users' waist and wrist of the dominant arm, and iii) ambient sensors from the smart home. The most successful methods observed for video and inertial data were combined, and an additional model that took into account both inertial and ambient sensors was proposed, with feature-level fusion within a neural network.

Research question 3 involved a multidisciplinary effort, which provided a more solid framework for its assessment. The contributions of this thesis were enhanced by the results of an approximately 10-months-long scholarship experienced by the author at Edinburgh Centre for Robotics in Heriot-Watt University, under supervision of Professor Dr. Patrícia Amâncio Vargas. In this participation, the author studied, designed and evaluated models in the field of computational neuroscience in the context of Parkinson's Disease (PD), as part of the Neuro4PD project [1]. PD is characterised by a dopaminergic neuronal loss within the substantia nigra pars compacta (SNc), which leads to a dysfunction of the basal ganglia-thalamus-cortex (BG-T-C) circuit. The BG-T-C circuit is a neuronal network with parallel loops that are involved in motor control, cognition, and processing of rewards and emotions (OBESO *et al.*, 2009). The research carried out within this context focuses on studying and modelling such a circuit.

The first outcome of this participation was the assessment of a deep learning framework based on CNN or LSTM, similar to the one employed for the activity recognition tasks reported here, as a technique to distinguish between healthy or PD-induced individuals from a database of marmoset monkeys, collected during a previous study (SANTANA *et al.*, 2014). This dataset consisted of recordings of Local Field Potentials (LFP) within the brain structures of the BG-T-C circuit in marmoset monkeys, either healthy or with lesions provoked by 6-hydroxidopamine (6-OHDA) and alpha-methyl-p-tyrosine (AMPT) injections, measured by electrodes surgically implanted. Not only performance results were reported, but also an analysis based on explainable features learned by the neural networks. More specifically, this study analysed the adherence to

[1] <http://www.macs.hw.ac.uk/neuro4pd/>, Royal Society and Newton Fund (NAF\R2\180773)

the spectral signature expected, for healthy and PD models, of the high-attribution segments of the input signals and of the internal representations of the intermediate convolutional layers.

The second outcome was the design of computational models that resemble neurobiological aspects found in primates, built upon a consolidated computational rat model of the BG-T-C circuit (KUMARAVELU; BROCKER; GRILL, 2016), capable of mimicking both healthy and PD conditions of primate models. To this aim, a data-driven approach was proposed, in which a set of biologically constrained parameters was determined using differential evolution to optimise a fitness function based on the LFP data of the marmosets' dataset already mentioned. This model was fully validated and it is capable of simulating the brain activity at the BG-T-C circuit ehibited by the animal models, with respect to the spectral signature, to the spike dynamics and to the coherence between spike trains in the different brain regions simulated.

Both of these outcomes were combined to a simulated environment to compose the third contribution of this thesis, which addresses the research question 3: the design of an application scenario for a mobile robot within a home environment, and the evaluation of behaviour selection techniques in response to the predictions from the activity recognition framework applied to the HWU-USP Activities dataset. A robot simulation, built in the Gazebo platform (KOENIG; HOWARD, 2004) as part of the LARa framework (RANIERI *et al.*, 2018), was adopted for a scenario in which, according to the activity being performed by the user of the environment, a particular response behaviour may be required from a mobile social robot (i.e., a simulated Pioneer P3-DX platform). To accomplish this based on the predictions made by the activity recogniser, two approaches were considered: the heuristics and the neurorobotics approaches. The heuristics approach consisted of a couple simple heuristics that connected the predictions to the response behaviours without any sophisticated reasoning. The neurorobotics approach was built on the neuroscience-based computational models, given that the BG-T-C circuit is involved in the process of decision-making in mammals (MARKOWITZ *et al.*, 2018), and is commonly adopted in neurorobotics systems to this aim (BARISELLI *et al.*, 2019; BAHUGUNA; WEIDEL; MORRISON, 2018; PRONIN *et al.*, 2021). This was accomplished by introducing different stimuli to different channels of the computational model, producing different response behaviours associated to each of those channels (MULCAHY; ATWOOD; KUZNETSOV, 2020).

1.2 Objectives

Given the hypothesis and research questions presented on the previous section, the main objective of this thesis is to design and evaluate an activity recognition framework based on deep neural networks, fed by multiple modalities within a richly sensed scenario, in particular, videos, inertial units and ambient sensors, and analyse behaviour selection strategies for a social robot within such an environment.

1.2.1 Specific Objectives

In order to address the this main objective four specific objectives were designed. The first and second specific objectives relate to research questions 1 and 2, respectively. The remaining specific objectives, i.e., the third and fourth, both address research question 3. The division of research question 3 into two specific objectives was made to separate the implementation of the application scenario and the design of the behaviour selection strategies, since these steps were developed in different stages of the research.

- Employ publicly available datasets to implement, evaluate, and evolve machine learning algorithms, particularly artificial neural networks, for activity recognition in videos and inertial units, modalities that have been provided simultaneously in some datasets.
- Design and collect a dataset of activities with synchronised information not only from videos and inertial units, but also from ambient sensors within a smart home, and adapt the machine learning techniques explored to this new scenario.
- Implement a simulated scenario and application for a robot in a home environment, allowing for the assessment of application scenarios based on quantitative metrics.
- Propose behaviour selection strategies, based on either simple heuristics or bioinspired mechanisms, and implement them in the robot simulation, comparing the outcomes of each approach for different classifiers and modalities employed for the activity recognition module.

1.3 Organisation of the Work

This thesis is organised as a collection of five papers, organised to compose the next four chapters. Three papers have been already peer-reviewed and published, and the reproduction of all material respected the copyright rules of the publishers, as depicted at the beginning of each chapter. The two other papers are preprints uploaded to arXiv. Upon publication of those papers to peer-reviewed sources, the preprint metadata at arXiv will be updated with the identifier to the published paper, without any infringement of copyright agreements.

Chapter 2 corresponds to the first research question. The results of experiments on publicly available datasets are presented, as published in the proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN 2020). Different types of neural networks were adapted to classify two datasets of videos and inertial sensors. A feature-level fusion approach was presented and compared to a simpler late fusion method.

Chapter 3 corresponds to the second research question. The HWU-USP activities dataset, designed and evaluated as part of this thesis, is presented, along with the techniques for classifying it, derived from the experiments presented in the previous chapter. These results were published

to Sensors, an open access journal by MDPI. This paper was reproduced in the chapter, providing a thorough literature review on activities datasets from different modalities, the data collection procedure, the algorithms applied for classifying the dataset, and different analyses regarding the behaviour of the classifier in limited-length segments, important for real-time applications. The documents required for ethics clearance of the data collection procedure are reproduced in Appendix A.

Chapter 4 presents the most relevant outcomes in the context of computational neuroscience, which were, later, introduced to the application scenario. This participation began with a 10-months scholarship at Heriot-Watt University, under supervision of professor Patrícia Amâncio Vargas, in which the author of this thesis participated on the Neuro4PD project, aimed at studying the neurophysiological correlates of Parkinson's Disease. Two papers were introduced to this chapter. The first of them consisted of an analysis, based on neural networks, of data from healthy and lesioned marmoset monkeys, published in the proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN 2020). The second one is a preprint uploaded to arXiv, in which a computational model of Parkinson's disease was calibrated to fit data from these marmoset monkeys according to a data-driven approach based on differential evolution. Although this research did not relate directly to the objectives of subsection 1.2.1, they were integrated to the robot simulation, since the computational model of brain regions affected by this disease could also be applied to behaviour selection, with the adequate adaptations.

Chapter 5 corresponds to the third research question. The paper reproduced is another preprint at arXiv, describing the different modules that composed an application scenario in HRI. This scenario relied on the outputs of the activity recognition framework, presented in the previous chapter, to build behaviour selection strategies. Approaches based on simple heuristics were presented and compared to a neurorobotics approach, significantly more complex, developed based on results of research on computational neuroscience, presented in the previous chapter.

Chapter 6 presents the conclusions, with a contextualised presentation of the different modules developed with respect to the integration between them and their role in evaluated the hypothesis and research questions the thesis. Also, the activities developed by the candidate during the PhD will be depicted, including other papers co-authored, participation in events and summer schools, and teaching internships. Finally, directions for future research are suggested. Papers published during the development of this thesis are listed in Appendix B. The summary of the other activities performed in parallel to the completion of this thesis is shown in Appendix C.

CHAPTER

2

MODELS FOR HUMAN ACTIVITY RECOGNITION

In this chapter, the multimodal techniques for human activity recognition, designed with deep neural networks and evaluated on public datasets, are presented. The achievement of this part of the thesis, which corresponds to the first research question of section 1.1, are described in the paper "Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach" (RANIERI; VARGAS; ROMERO, 2020), published to the 2020 International Joint Conference on Neural Networks (IJCNN), the IEEE conference on Neural Networks, with H5-index of 45 (Google Scholar) and qualified as A1 in the latest Qualis CC. This paper is reproduced in the following pages.

Contribution statement: Ranieri performed the literature review, experimental design, implementation, and evaluation of the proposed methods, under supervision of the other authors. Vargas and Romero revised the paper and provided enhancement of its final presentation.

Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach

1st Caetano M. Ranieri
ICMC
University of São Paulo (USP)
São Carlos, SP, Brazil
cmranieri@usp.br

2rd Patricia A. Vargas
Edinburgh Centre for Robotics (ECR)
Heriot-Watt University (HWU)
Edinburgh, Scotland, UK
p.a.vargas@hw.ac.uk

3th Roseli A. F. Romero
ICMC
University of São Paulo (USP)
São Carlos, SP, Brazil
rafrance@icmc.usp.br

***Abstract*—Recent breakthroughs on deep learning and computer vision have encouraged the use of multimodal human activity recognition aiming at applications in human-robot-interaction. The wide availability of videos at online platforms has made this modality one of the most promising for this task, whereas some researchers have tried to enhance the video data with wearable sensors attached to human subjects. However, temporal information on both video and inertial sensors are still under investigation. Most of the current work focusing on daily activities do not present comparative studies considering different temporal approaches. In this paper, we are proposing a new model build upon a Two-Stream ConvNet for action recognition, enhanced with Long Short-Term Memory (LSTM) and a Temporal Convolution Networks (TCN) to investigate the temporal information on videos and inertial sensors. A feature-level fusion approach prior to temporal modelling is also proposed and evaluated. Experiments have been conducted on the egocentric multimodal dataset and on the UTD-MHAD. LSTM and TCN showed competitive results, with the TCN performing slightly better for most applications. The feature-level fusion approach also performed well on the UTD-MHAD with some overfitting on the egocentric multimodal dataset. Overall the proposed model presented promising results on both datasets compatible with the state-of-the-art, providing insights on the use of deep learning for human-robot-interaction applications.**

***Index Terms*—Deep learning, CNN, LSTM, TCN, RNN, human activity recognition, human-robot-interaction.**

I. Introduction

Current development on different research fields have risen interest on applications of social robots as interactive tools to assist humans, usually elderly people or people with special needs. In real-world scenarios, roboticists may rely on human activity recognition [1]. This consists in processing sensing data from smartphones and wearable devices to identify semantically understandable interactions amongst the user, the environment and the robot. These technologies are important for the development of automated solutions for human-robot interaction applications that are still mostly based on Wizard of Oz approaches [2].

São Paulo Research Foundation (FAPESP), grants 2017/02377-5, 2018/25902-0 and 2017/01687-0, and Brazilian National Council for Scientific and Technological Development (CNPq), grant 306151/2018-9. This research was carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Nvidia Grants program.

Here we address this challenge by proposing a deep learning model for human activity recognition from videos and inertial sensors. Inertial data may be made available from smartphones or wearable devices such as smartwatches. In situations in which social robots are present, video data may also be obtained from the robot's camera(s). Regardless of the modality, deep learning techniques have shown promising results on activity recognition, although feature-based approaches are still competitive in some cases [3]. Most advances on video classification were built on the Two-Stream ConvNet [4], whereas satisfactory results on inertial data have been provided by the combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) [5]. Some other attempts dealing with multimodal data, focused on fusing inertial data with depth images [6].

In our investigation we have used two datasets for daily activities: the egocentric dataset presented by Song *et al.* [1] and the University of Texas at Dallas Multimodal Human Activity Recognition Dataset (UTD-MHAD) [7]. Our proposed model relied on RGB videos and inertial data, experimenting different possibilities for modelling the temporal dependencies on both modalities. In this regard, first, we are proposing to add Temporal Convolutional Networks (TCN) [8], which consists on an feasible alternative to Recurrent Neural Networks (RNN) on sequence modelling. Second, a feature-level fusion approach is considered as an alternative to the late fusion generally used when dealing with video temporal streams.

II. Human Activity Recognition

Human activity recognition comprises of a wide research field, involving different input modalities and classification of activities on distinct levels of abstraction. In the case of videos, this modality may rely not only on structured data built on controlled environments, but also on unconstrained videos obtained from the Internet [9]. The same has not been true for raw sensors such as inertial measurement units (IMU), which may aggregate, for instance, 3D accelerometer, gyroscope and magnetometer [10]. For those, datasets are typically designed and recorded under controlled environments. In this work, we address a multimodal approach, in which data from inertial sensors has been applied to enhance video-based activity recognition. Even though, single-modality approaches

also influenced our research, providing guidelines on several developments on our proposed methods.

The UCF101 dataset [11] is probably the most relevant benchmark for video classification available. It is composed by 101 categories distributed into human-object interaction, body movements, human-human interaction, musical instruments playing and sports. More recently, large-scale datasets have been deployed and the most relevant one is the Kinetics dataset [12]. Since its volume of data may take several terabytes, it is not always feasible to work directly with datasets of such scale. As we discuss thoroughly on Section III, a Convolutional Neural Network (CNN) trained from scratch on data derived from the UCF-101 dataset has been adopted as a building block for most of our proposed architectures.

A detailed literature review regarding methods for video classification was presented by Herath *et al.* [13]. Most deep learning approaches may be belong into two categories: multiple stream networks or spatio-temporal networks. The most influential multiple stream network is the Two-Stream ConvNet proposed by Simonyan *et al.* [14]. It was composed by a spatial CNN trained to classify RGB frames and a temporal CNN trained on stacks of dense optical flows from sequential frames. This approach have evolved and an important advance was the Temporal Segment Network [15]. Spatio-temporal networks are characterised by combinations between CNN and LSTM, such as the Long-term Recurrent Convolutional Networks (LRCN) [16], or 3D ConvNets (C3D), as presented by Tran *et al.* [17]. Our approach is composed of multiple streams. However, the video temporal streams were built with similar basic principles as the LRCN.

For inertial sensors, a dataset often used in studies centred on wearable devices is the PAMAP2 [18]. The OPPORTUNITY [10] dataset is also relevant, as it provides a large set of sensors not only wearable, but also placed on objects or distributed around an environment. The neural networks architectures used to classify those datasets are almost always based on combinations between CNN and LSTM. A systematical analysis of deep learning techniques for inertial data, experimented in datasets such as the both mentioned, was performed on Hammerla *et al.* [19], in which regular deep neural networks (DNN) were compared to CNNs and three LSTM-based architectures. In Rueda and Fink [20], features extracted from CNNs were on the basis of three architectures: a regular CNN, a variation called DeepConvLSTM, in which LSTM layers would replace fully-connected layers, and the CNN-IMU, composed of parallel convolutional blocks whose outputs were concatenated and fed to fully-connected layers. The InnoHAR architecture [21] consists of a stack of Inception modules followed by two recurrent layers based on Gated Recurrent Units (GRU), and led to improved results on both PAMAP2 and OPPORTUNITY datasets. A detailed overview of the literature regarding smartphone sensors was provided on the recent work of Sousa Lima *et al.* [22], in which different datasets and algorithms, including deep networks, were broadly revised.

Regarding multimodal datasets with videos and inertial sensors, most of them were recorded with depth cameras, as discussed on the survey provided by Chen *et al.* [6]. Datasets such as the UTD-MHAD [7], adopted in the experiments, and the 50 Salads [23] provide not only video and inertial measurements, but also positioning of skeleton joints, which are often used as an important input for the proposed methods [24]. In Chen *et al.* [7], Depth Motion Maps (DMM) were obtained from depth images, statistical descriptors were adopted for the inertial data and the RGB videos were not considered. Classification was performed with Collaborative Representation Classifiers (CRC). Song *et al.* [3], another object of our analysis, brought a different approach, in which scripted actions were performed by 10 participants and recorded with a Google Glass. In a following paper [1], the authors applied the two-stream ConvNet to classify the videos from their dataset and a DeepConvLSTM to classify the sensor data, performing fusion by averaging or max-pooling their outputs.

III. Proposed Model

In this article, we propose to build on the Two-Stream ConvNet [14] and extend it to the case in which another modality composed by IMU sensor data is present. This modality, comprised by multivariate 1D temporal series, has been considered as an additional stream, called *inertial*, as illustrated in Fig. 1. An Inception-V3 network [25], adapted to take pairs of optical flow matrices (U, V) as inputs, has been previously trained on the UCF-101 dataset. Therefore, instead of taking three input color channels of the RGB images, the network would take the two optical flow channels: vertical and horizontal. Further, its last layer was removed, in order to provide a feature vector for each timestep of the video. In other words, the penultimate layer of the Inception-V3 would generate a feature vector of length 2048 of a given timestep, and this network would be applied independently for each timestep considered. A much simpler CNN was implemented to extract features from the inertial stream, which could be used as inputs to a LSTM or a TCN block. Those outputs could be concatenated to the features obtained from other time-dependent streams, particularly the video temporal stream. In the later case, we are assuming that both of them are related to the same amount of time on the sample, so that $c = t_s \times \omega_s$, where t_s is the number of timesteps of stream $s \in \{\text{video}, \text{inertial}\}$, ω_s is its frequency and c is the time amount, in seconds, shared between the streams. Given such assumption, discrepancies on the number of timesteps at the time of the concatenation could be resolved by sampling from the stream with more timesteps.

More precisely, the LSTM and TCN models for temporal modelling were applied to the features extracted by CNNs, and its outputs were fed to a softmax layer for classification. Although LSTM was already applied for video classification on previous literature [26], the suitability of TCN, which has shown to lead to equivalent or even better results in sequence modelling [8], has not been extensively applied to this context.

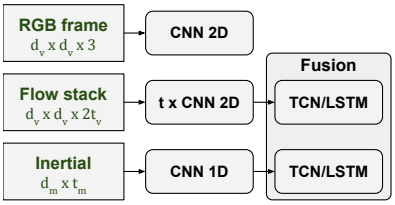


Fig. 1: Proposed framework for multimodal activity recognition, where d refers to the number of features and t, to the number of timesteps considered for a given stream. Fusion may be performed at feature-level, by combining the features from different modalities obtained from the CNNs. Both LSTM and TCN layers were considered for modelling long-term dependencies.

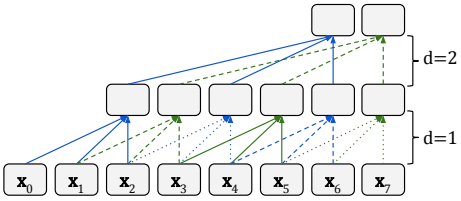
A. Temporal Convolutional Networks

The LSTM architecture is a classical approach for dealing with long-term temporal dependencies in sequences [27]. Recently, it has led to several advances on deep learning, especially regarding language and speech recognition [28]. The success of the LSTM and of its most famous variation, the Gated Recurrent Unit (GRU) [29] turned recurrent neural networks the standard starting point when dealing with deep learning for sequence modelling. However, as Bai *et al.* [8] argued, approaches based solely on convolutional networks could provide results as good as recurrent approaches, and therefore it may be worth to consider them as well. In this context, the temporal convolutional network (TCN) comprises of a neural architecture capable of dealing with long-term dependencies.

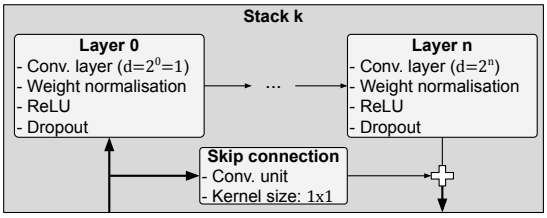
The temporal information would be dealt in such networks by stacks of *dilated causal convolutions*, which are illustrated in Fig. 2a. The *causal* denomination is derived from the connections between the layers. A filter of size k processes a timestep t plus the $k-1$ preceding timesteps, in order to capture the idea of causality. The *dilated* denomination refers to the inclusion of a dilation factor d, responsible for amplifying exponentially the receptive field of the convolutions, as more levels are added to the network. A regular convolution is the particular case in which $d = 1$. Considering a 1D input sequence $\mathbf{x} \in \mathbb{R}^n$ and a filter $f : \{0, \dots, k-1\} \rightarrow \mathbb{R}$, the dilated convolution operator may be defined as in equation 1, where $s - d \cdot i$ refers to the direction of the receptive field to the past.

$$F(s) = (\mathbf{x} *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-d \cdot i} \qquad (1)$$

Each convolutional stack would be followed by weight normalisation, an activation function (e.g., ReLU) and spatial dropout, composing residual blocks as shown in Fig. 2b. The advantage of such blocks is the so-called skip connections, which allow the input data to be fed directly not only to the next block, but also to each of the following blocks. To fix the differences of dimensions, 1×1 convolutions may be applied to adjust the previous inputs before they are combined to the output of a block .



(a)



(b)

Fig. 2: Elements of a TCN. (a) Stack of dilated causal convolutions, with $k = 3$ and dilation factors $d = \{1, 2\}$. The convolutional layers within each stack are comprised by dilated causal convolutions. (b) Generic residual stack, with n dilations, where the dilation factor is increased each layer as a power of 2. Multiple stacks may be concatenated one after another, as in a residual neural network.

B. Video Classification

The architecture for activity recognition on videos was based on the Two-Stream ConvNet, in which the spatial features are extracted by a CNN with RGB frames as input, and temporal dependencies, by a CNN which takes optical flow matrices. Before being fed to the correspondent neural network, RGB frames or optical flow matrices were supposed to be cropped to $d \times d$. For both streams, the InceptionV3 network [25] was adopted as a base model. For the spatial model, we applied a straightforward transfer learning from a model previously trained on the ImageNet dataset [30], in which only the softmax layer was replaced and further trained with the weights of all other layers being fixed.

For the temporal stream, illustrated in Fig. 3, t_v successive pairs of optical flow with shape $d_v \times d_v \times 2$, each corresponding to a single timestep of a sequence, were fed independently to the CNN. This approach is different from Simonyan *et al.* [14], in which a CNN took as input a stack of optical flow matrices related to successive timesteps, i.e., the architecture was composed by a single CNN with input shape $d_v \times d_v \times 2t_v$. Here, a determined CNN, trained from scratch to classify the UCF-101 dataset and deprived from its last softmax layer, would process the pairs of optical flow matrices. The result would consist of a feature vector with shape (a_v, t_v), where a is the number of features generated by the output of the CNN - in the case of the network InceptionV3, $a_v = 2048$. In other words, this feature vector would be a multivariate time series with a_v variables

and t_v timesteps. LSTM networks are commonly seen as a good choice for modelling such one-dimensional signals, so as TCNs, as discussed in subsection III-A. Therefore, LSTM and TCN were both considered as candidate layers for this part of the proposed architecture. Finally, the last output of whichever network was used would be fed to the softmax layer for the classification.

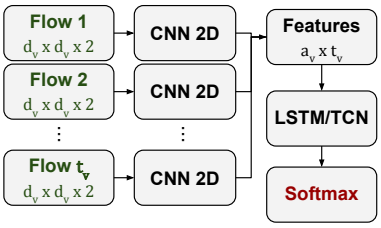


Fig. 3: Network architecture for the temporal stream. The inputs are the pairs of dense optical flows from a frame sequence. Each pair with shape $d_v \times d_v \times 2$, is processed by a shared CNN, and the a_v features obtained from the last layer of the CNN (prior to the softmax layer previously withdrawn) are taken as timesteps for a LSTM or TCN.

C. Inertial Data Classification

In order to classify the inertial data, it has been adopted a one-dimensional version of the same principle as that one for video temporal stream: a CNN to extract features, followed by a LSTM or TCN to model long-term temporal dependencies. This approach has similarities to the work of Rueda and Fink [20]. However, a network architecture was deployed having in mind the particular issues that would arise when performing a fusion with the video temporal stream. Particularly, since the convolutions on the inertial data would be performed on the time domain, and each pooling layer would reduce the resolution at this given domain to the ratio of its kernel, we had to be cautious with the increasing of the depth of this CNN. With this aim, we have considered only two Conv1D layers: the first one with kernel size 1, to increase the number of feature maps, and the second, with size 3, to perform feature extraction. Those layers were followed by a maximum pooling of kernel size 2, which would reduce the number of timesteps t_m to its half, t_n, while still representing the same amount of time (i.e., the time resolution has dropped). The CNN architecture is shown in Fig. 4a.

The a_n features extracted from this CNN were, then, applied as input to a LSTM or TCN block, whose last output was connected to a softmax layer for classification (see Fig. 4b). An important difference between this neural network and that of the video temporal stream is that all the free parameters of both CNN and LSTM/TCN were set to be trainable, i.e., training would be performed end-to-end.

D. Temporal Fusion

In most research on activity recognition based on multiple-stream deep neural networks, fusion was performed at a later stage. For instance, by averaging the outputs of the

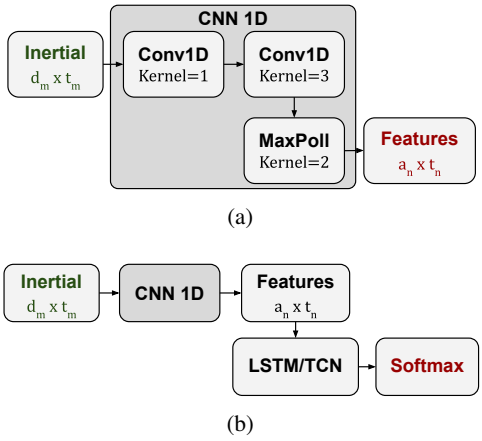


Fig. 4: Neural network applied for the classification of the inertial stream. (a) CNN applied prior to the LSTM or TCN module. Since the convolutions are performed in the time domain, the maximum pooling with kernel size 2 reduces the initial temporal resolution t_m to $t_n = \frac{t_m}{2}$. The number of output features a_n was determined by the number of filters of the second Conv1D layer. (b) For the inertial stream, the inputs are one-dimensional sample sequences. The whole sequence is processed by a CNN in the time domain, which reduces the number of timesteps from d_i to d_j.

last layer. Song *et al.* [1] adapted this approach to fuse the video features to those extracted from the inertial data of their egocentric multimodal dataset. Regarding video-only classification, Feichtenhofer, Pinz and Zisserman [31] analysed different methods for feature-level fusion in two-stream ConvNets. Most of the techniques they proposed rely on the spatial dependencies shared by the video temporal and spatial streams. Therefore, they are not suitable for fusion with the inertial stream. However, we could adapt the concatenation of features presented by them to build our feature-fusion approach, since it does not make assumptions on the spatial dependencies between features.

The proposed method here, shown in Fig 5, builds on two assumptions: the numbers of timesteps t_v on the videos and t_m on the inertial data are synchronised, referring to the same period of the sample on both streams despite each modality having a different temporal resolution; and that $t_v \leq t_n$. Thus, after applying each of the t_v $(d_v \times d_v \times 2)$ optical flow matrices to CNN 2D and stacking the outputs, and applying the $d_m \times t_m$ inertial data sample to CNN 1D, two feature vectors would be obtained, with shapes $a_v \times t_v$ and $a_n \times t_n$. If $t_v \neq t_n$, the inertial feature vector should be adjusted, what would be done by sampling points that were equidistant in the time domain. After such adjustment, both feature vectors would have the same number of timesteps t_v. Therefore, they may be concatenated in this dimension, resulting in a feature vector of shape $(a_v + a_n) \times t_v$. This feature vector would be fed to a LSTM or TCN block, whose output would be connected to a softmax layer. It is worth to remind that CNN 2D has fixed

weights, already optimised in an ad-hoc manner.

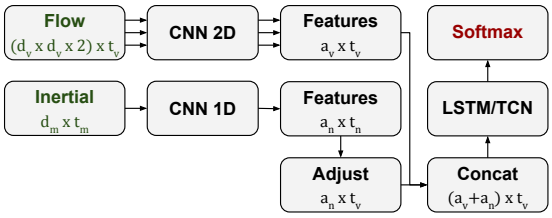


Fig. 5: Framework for feature-level temporal fusion. Features extracted by the video CNN are concatenated to the features extracted by the inertial CNN, composing a feature vector related to a single timestep. The frequencies at each modality are different, thus adjustment by down-sampling is applied to the inertial stream before concatenating, so that the timesteps at all streams are synchronised. After concatenation, the multimodal feature vectors at each timestep are fed to a temporal neural network, be it a LSTM or a TCN module.)

IV. Experimental Setup

All implementations were developed in Python language, using the Keras framework with TensorFlow backend. Before any preprocessing, all videos were proportionally resized so that the smallest side would have size 256. Optical flow was calculated with the TVL1 algorithm [32]. This algorithm has shown, in exploratory experiments, to provide significantly the best classification results among others, although being significantly slower than the Farnebäck algorithm [33], which may be relevant to real-time applications. The networks for the video streams were set to get input frames with shape 224×224, which would be achieved by cropping. Split 1 of UCF101 dataset, suggested by the authors [11], was used to train the CNN for feature extraction in the video temporal stream. The following subsections will present the datasets and the settings for each condition, allowing the results to be reproduced. The code was made available at https://github.com/cmranieri/Deep-Activity-Recognition.

A. Datasets

The experiments were performed on two multimodal datasets: egocentric multimodal [1] and the UTD-MHAD [7]. Those datasets were chosen for their suitability to activities of daily living. Both of them provide the same amount of data from each subject and with respect to each activity. Besides, as they are significantly different in nature, interesting conclusions could be drawn from comparative results.

1) Egocentric Multimodal Dataset: This dataset was generated by a group of 10 participants. They performed a set of 20 activities wearing a Google Glass. Each session length was about 10 seconds. These activities were recorded in different and heterogeneous environments, which provides a lot of visual information, in addition to the movement. Activities were divided into four categories: ambulation, office work, daily activities and exercises. The videos (RGB only) were sampled at 30 Hz, while the sensor data was sampled at 15 Hz. The sensors provided 19 features: the 3D acceleration, magnetic field, linear acceleration, gravity, rotation vector and gyroscope. The data was preprocessed using the L2-norm.

2) UTD-MHAD: This dataset was recorded in a more controlled condition, with 8 participants performing a set of 27 activities, 4 repetitions each. Recordings were performed by a depth and RGB camera (only RGB video was considered in this work) and by two 3D accelerometers. One placed at a band on the user's fist, and the other was placed at the user's waist. Each session lasted about 3 seconds, and the recordings were performed in a controlled room, with the subjects posed facing the camera, at a constant distance and with constant background. The videos were sampled at 15 Hz, and the sensor data, with 6 dimensions corresponding to the two 3D accelerometers, were sampled at 50 Hz.

B. Network Setting

All conditions described in this subsection were experimented on both datasets described in the previous section. The datasets were split following the k-fold cross-validation procedure, with $k = 10$ for the egocentric multimodal dataset and $k = 8$ for the UTD-MHAD, so that data provided by one subject was used for testing, and the remaining data, for training. For the data stream, only one condition was considered, in order to allow for late fusion: an InceptionV3 CNN. As previously stated, transfer learning was applied to a model trained on Imagenet dataset, keeping all weights fixed except for the softmax layer, replaced to match the number of classes of the datasets considered.

The temporal and inertial streams were considered separately and followed the fusion approach of Fig. 5. The CNN applied to extract features of the inertial stream was composed by 256 filters in the first convolutional layer and 512 in the second. As the InceptionV3 ouputs 2048 features, the feature-based fusion provides a vector with video and inertial features in a ratio of 4:1. Both LSTM and TCN were experimented as blocks for temporal modelling, with 128 units and output dropout of 0.3. Regarding the TCN, the kernel size was set to 3, dilations were set do $d = \{1, 2, 4\}$, and the number of residual blocks (i.e., stacks) to 3.

1) Training: The training procedure was adapted from Simonyan *et al.* [14] and Song *et al.* [3]. All models were optimised using the softmax cross-entropy as loss function. The pre-training of CNN for optical flow pairs performed on the split 1 of UCF-101 dataset was ran with the Stochastic Gradient Descent (SGD) optimiser, for $200,000$ steps. In the videos of the goal datasets, data augmentation was performed by random cropping and in the egocentric multimodal dataset, random flipping. We decided not to flip the videos from UTD-MHAD, since some of the activities on that dataset were somewhat symmetric (e.g., wave left and wave right). For the spatial stream, we used SGD with learning rate 10^{-2}, momentum 0.9 and weight decay 10^{-4}, and training was also performed for $30,000$ steps, with batches of size 32. Optimisation on the temporal and inertial streams was

performed in batches of size 16, for $30,000$ training steps, using the RMSProp optimiser [34] with learning rate 10^{-3}.

The number of timesteps was selected so that each snippet would represent 2 seconds of a trial. To reduce the number of video frames, we sampled them so that $t_v = 15$. With the egocentric multimodal dataset, the model was sampled once every 4 frames at the video stream, and the timesteps were set to $t_m = 30$ and $t_n = 15$ for the inertial stream. We sampled once every 2 frames with UTD-MHAD, the timesteps of the inertial stream being set to $t_m = 100$ and $t_n = 50$. Therefore, we had to apply the adjustment depicted in Fig. 5. The same settings were kept when training the inertial stream alone, except the adjustment by sampling in UTD-MHAD.

2) Evaluation: For testing we used the same procedure adopted in the reference papers: a number snippets was considered, with equal time between them, and all of them were submitted to cropping on their four corners and centre. For the egocentric mutimodal dataset, 5 snippets were used to test each video, and the videos from the resulting sequences were also horizontally flipped. For the UTD-MHAD, we considered 2 snippets and no flipping. To make a prediction, output vectors from all snippets of a given sample were averaged.

This procedure was adopted for all models that ran end-to-end, i.e., the models for single-stream and feature-level fusion. For late fusion, one model for each stream was run separately and the output vectors were combined by weighted averaging. The same was done when combining to the spatial stream.

V. Results and Discussion

Fig. 6 shows the number of parameters of each model built for each stream on the egocentric multimodal dataset (UTD-MHAD was fairly alike), including the hybrid model for feature-level fusion. *Late fusion* was not considered a model on itself, since it consists on combining the *spatial* models outputs with one of the *temporal* models. Therefore, at inference time, its number of parameters equals the sum of those present on the models adopted. The *temporal* or *feature fusion* models embed a CNN similar to that of *stream* model. Therefore, their complexity is dependent on the base CNN model adopted.

Since InceptionV3 (adopted on all of our models except for the inertial ones) is expressively more complex than the remaining parts of the architecture, variations on the number of parameters are proportionally small. But yet relevant, since the weights relative to this block are fixed during training. It is noteworthy that TCN model was more complex than LSTM for inertial stream, while the opposite happened for the temporal and feature fusion models. Due to the fact that the temporal block on the inertial stream has shape $512 \times t_v$, against $2048 \times t_v$ on the video stream, thus $2560 \times t_v$ in the feature-fusion models, it may be inferred that the number of input features of the temporal block impacts less the number of parameters in TCN-based than in LSTM-based models. This is expected due to the sparser connectivity of convolutional layers.

The InceptionV3 CNN, which was embedded on the *temporal* and *feature fusion* models to extract features based on single optical flow pairs, was trained separately, prior to the

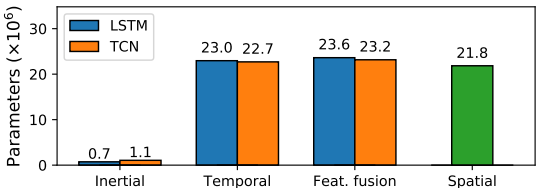


Fig. 6: Number of free parameters of each model analysed, without the softmax layer. The substantially higher number of parameters at the models that involve video processing is given to the InceptionV3 neural network contained on it, which consists of more than 21 million trainable parameters, as made explicit by the number of parameters of the spatial stream.

TABLE I: Mean accuracy of each model for the temporal and inertial streams, using 10 folds for the egocentric dataset and eight for the UTD-MHAD, providing splits such that the test set was composed by all recordings of one subject.

Dataset	Stream	Model	
		LSTM (%)	TCN (%)
Egocentric	Inertial	45.50 ± 7.39	45.50 ± 8.50
	Temporal	69.00 ± 10.68	72.50 ± 11.01
	Feat. fusion	55.50 ± 9.60	53.00 ± 10.77
	Late fusion	74.50 ± 8.20	72.50 ± 9.35
UTD-MHAD	Inertial	63.28 ± 5.71	65.36 ± 9.24
	Temporal	80.02 ± 6.00	81.77 ± 6.49
	Feat. fusion	82.58 ± 5.56	85.47 ± 5.56
	Late fusion	84.90 ± 4.78	83.51 ± 6.25

experiments presented in this paper. It achieved accuracy of **75.15%** on the split 1 of UCF-101, using the same training and evaluation protocol as Simonyan *et al.* [14]. The resulting layers were added as blocks of our architecture, as discussed in section III, and its weights were kept fixed. This was different for the *inertial* stream, whose features were extracted by a simpler network randomly initialised to be optimised together with following layers for temporal modelling. For all models on both datasets, LSTM and TCN blocks were investigated. The mean accuracy of each model for the temporal and inertial streams is shown in Table I.

As some of the results in Table I are close to each other, it may be convenient to compare the performances of each model with respect to some additional aspects. In Fig. 7, we also present the macro F1-score of the models, that is, the average harmonic mean of precision and recall. By penalizing both incompleteness and inconsistency, this measure is a trade-off between type-I and type-II errors per class. The means between the evaluations on each fold were presented in the bars, with standard deviations proportional to the length of the vertical traces on the top of it.

The *spatial* model was obtained by a procedure similar to that of the base CNN block of the *temporal* models. However, it took RGB frames as inputs, instead of pairs of optical flow matrices; and was initialised with ImageNet weights, instead of being trained from scratch. This model was used to build classifications using the three mentioned streams, by fusing it to the models presented in Table I by weighted averaging.

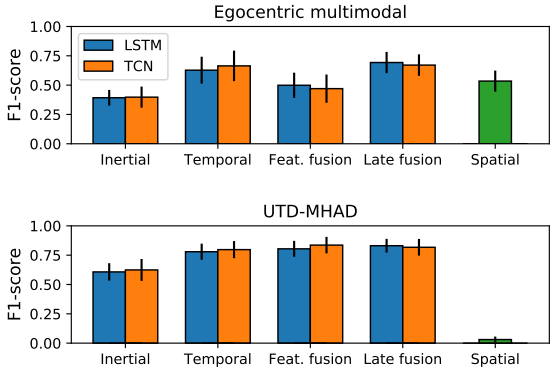


Fig. 7: Macro F1-scores for each model.

TABLE II: Accuracy, for the egocentric multimodal dataset, of the spatial stream models and late fusion by weighted averaging with each of the fused temporal and inertial models. If w_s and w_t are the weights for the spatial and temporal/hybrid streams, the weights are shown using notation $w_s : w_t$.

Temporal model		Weights	Accuracy (%)
Spatial only		1:0	60.50 ± 8.50
LSTM	Video	1:1	72.50 ± 6.42
	Feat. fus.	3:1	69.00 ± 9.69
	Late fus.	1:6	78.50 ± 9.23
TCN	Video	1:2	78.00 ± 10.54
	Feat. fus.	2:1	70.25 ± 9.16
	Late fus.	1:6	80.62 ± 8.81

The fusion weights were selected so that the accuracy was the largest obtained in our experiments. As UTD-MHAD dataset was built on a controlled environment with constant background and without significant differences on objects able to distinguish between activities, the spatial stream was not significantly informative, with accuracy of $(\mathbf{6.74 \pm 3.23})\%$, only slightly above random choice (i.e., 3.70%, given that there were 27 classes). For this reason, fusion between the three streams were made only for the egocentric multimodal dataset. Results are reported in Table II.

A. Discussion

Results from LSTM and TCN-based models were generally quite close to each other, with a slight tendency in favor of TCN models for most single-stream approaches and all models combined with the spatial stream (Tables I and II). The feature-level fusion approach was successful in the UDT-MHAD, surpassing the accuracy of the late fusion when coupled with a TCN block and achieving the best accuracy for this dataset, of 85.47%. Since this dataset is endowed with other modalities, skeleton joints and depth frames, it was expected that the proposed model would perform below the most accurate models on the literature. Still, our proposal may be seen as competitive, since most of our results outperformed those reported on the reference paper [7], which achieved, at most, overall accuracy of 79.10%. It might be noticed that our approach relies only on RGB and inertial data, which are more widely available and may be included in different sorts of systems. With a more complex model, in which LSTM networks also modelled depth information, Li *et al.* [35] achieved an accuracy as high as 95.31%.

On the egocentric multimodal dataset, feature-fusion approaches had suffered from overfitting, with fast optimisation and very high training accuracy. However, average test accuracy is below the temporal stream alone, which has shown lower accuracy during all the training procedure, and actually was harder to optimise than the other models. Considering inertial and temporal streams, the best accuracy was achieved by the late fusion of LSTM-based models (74.50%). This result was curiously different when the models were further combined with the spatial stream, with the late fusion of TCN models achieving the best overall accuracy for this dataset among our models, e.g., 80.62%. Although this was only compatible to the best multiple stream CNN model presented by Song *et al.* [1] which reported 80.50%, it might be noticed that our approach presents some advantages. As we relied on a previously trained CNN to extract features from the optical flow matrices, with a very reduced set of parameters left to be optimised in a LSTM or TCN block, it provides the flexibility to work with different and arbitrarily complex CNNs for this aim. Moreover, since the number of parameters left to be optimised is relatively low, with our approach one can work with larger sequences of data even with a modest hardware.

The F1-scores shown in Fig. 7 were consistent with the accuracy results, thus there were no issues regarding classes with very high precision and low recall or the opposite. Besides, models with higher accuracy have also shown higher F1-score, i.e., both measures were suitable to make comparisons.

The proposed framework may contribute to further applications on human-robot interaction [36] [37], especially on scenarios which demand social interaction between user and robot [38] [39].

VI. Concluding Remarks and Future Work

In this paper, a new model for human activity recognition on videos and data from inertial sensors was proposed. First, different neural networks were analysed as building blocks for the temporal processing, particularly Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCN). Second, fusion between the inertial and video temporal streams were not only performed through late fusion of the output layers, but also at feature-level. All those approaches were analysed separately, for different sets of modalities, and thorough comparisons were done.

Focus was given to modelling the temporal dependencies in sequences of tuples of inertial data, features extracted from optical flow and fusion between those approaches. For the temporal feature extraction, we adopted Long Sort-Term Memory (LSTM) units and Temporal Convolutional Networks (TCN). A feature-fusion approach was also proposed and compared to the more traditional late fusion approach, commonly adopted on multiple stream CNNs. The RGB frames were also contemplated, with output features from a spatial CNN

further combined to the other models through weighted averaging, achieving accuracies up to 80.62% for the egocentric multimodal dataset, and 85.47% for the UTD-MHAD without considering depth data.

Experiments were performed on the egocentric multimodal dataset and UTD-MHAD. Models obtained with LSTM and TCN blocks both led to excellent accuracies, with TCN, which we have brought as a novelty to this application, performing slightly better in many circumstances. The feature-fusion approach led to good results in UTD-MHAD dataset. However, it was unable to generalize well on the egocentric multimodal. Overall, the proposed model presented promising results on both datasets compatible with the state-of-the-art, which provided further insights on the use of deep learning for human-robot-interaction applications.

Future work will contemplate depth images as an additional stream, since this may be introduced to social robots in several circumstances. We have already built a multimodal dataset for activities in domestic environments, with videos and inertial data from smartwatches and smartphones, to be used on deep learning models in human-robot interaction applications. This dataset will be made publicly available once we finish the anonymisation procedures.

REFERENCES

[1] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J.-H. Lim, G. S. Babu, P. P. San, and N.-M. Cheung, "Multimodal multi-stream deep learning for egocentric activity recognition," in *IEEE CVPRW*, 2016.

[2] J. T. Browne, "Wizard of oz prototyping for machine learning experiences," in *2019 Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[3] S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, and J. Liri, "Egocentric activity recognition with multimodal fisher vector," in *IEEE ICASSP*. IEEE, 2016, pp. 2717–2721.

[4] C. Feichtenhofer, A. Pinz, R. P. Wildes, and A. Zisserman, "Deep insights into convolutional networks for video recognition," *International Journal of Computer Vision*, pp. 1–18, 2019.

[5] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 1 2016.

[6] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2 2017.

[7] ——, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE ICIP*, 2015, pp. 168–172.

[8] S. Bai, J. Zico Kolter, and V. Koltun, "Convolutional sequence modeling revisited," in *6th ICLR*. OpenReview.net, 2018.

[9] C. Gan, C. Sun, L. Duan, and B. Gong, "Webly-supervised video recognition by mutually voting for relevant web images and web video frames," in *ECCV*. Springer, Cham, 2016, pp. 849–866.

[10] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen, "The Opportunity challenge: a benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033–2042, 2013.

[11] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," in *CVPR*, 2012.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv:1705.06950*, 2017.

[13] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: a survey," *Image and Vision Computing*, vol. 60, 2017.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *ECCV*. Springer, Cham, 2016, pp. 20–36.

[16] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 39, no. 4, pp. 677–691, 4 2017.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *ICCV*. IEEE, 12 2015, pp. 4489–4497.

[18] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *16th ISWC*. IEEE, 2012, pp. 108–109.

[19] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *IJCAI)*. AAAI Press, 2016, pp. 1533–1540.

[20] F. M. Rueda and G. A. Fink, "Learning attribute representation for human activity recognition," in *ICPR*. IEEE, 2018, pp. 523–528.

[21] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "Innohar: a deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.

[22] W. Sousa Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama, "Human activity recognition using inertial sensors in a smartphone: An overview," *Sensors*, vol. 19, no. 14, p. 3213, 2019.

[23] S. Stein and S. J. Mckenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *UbiComp*. Zurich, Switzerland: ACM, 2013.

[24] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2 2016.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE CVPR*. Las Vegas, NV, USA: IEEE, 2016, pp. 2818–2826.

[26] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *23rd ICM*. ACM, 2015, pp. 461–470.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.

[28] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward torRecurrent LSTM neural networks for language modeling," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 23, no. 3, pp. 517–529, 2015.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS*, 2014.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 12 2015.

[31] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*. IEEE, 6 2016, pp. 1933–1941.

[32] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time TV-L 1 optical flow," in *Patt. Recog. Symp.*, 2007, pp. 214–223.

[33] G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *Scandinavian Conference on Image Analysis (SCIA)*. Halmstad, Sweden: Springer, Berlin, Heidelberg, 2003, pp. 363–370.

[34] M. C. Mukkamala and M. Hein, "Variants of RMSProp and Adagrad with logarithmic regret bounds," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. Sydney, NSW, Australia: ACM, 2017, pp. 2545–2553.

[35] K. Li, X. Zhao, J. Bian, and M. Tan, "Sequential learning for multimodal 3D human activity recognition with long short-term memory," in *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*. Takamatsu, Japan: IEEE, 8 2017, pp. 1556–1561.

[36] P. Vargas, Y. Fernaeus, M. Lim, S. Enz, W. Ho, M. Jacobson, and R. Aylett, "Advocating an ethical memory model for artificial companions from a human-centred perspective," *AI Society*, vol. 26, pp. 329–337, 2011.

[37] B. V. Ferreira, E. Carvalho, M. R. Ferreira, P. A. Vargas, J. Ueyama, and G. Pessin, "Exploiting the use of convolutional neural networks for localization in indoor environments," *Applied Artificial Intelligence*, vol. 31, no. 3, pp. 279–287, 2017.

[38] S. Enz, M. Diruf, C. Spielhagen, C. Zoll, and P. A. Vargas, "The social role of robots in the future—explorative measurement of hopes and fears," *Int J of Soc Robotics*, vol. 3, no. 263, 2011.

[39] C. Rizzi, C. G. Johnson, F. Fabris, and P. A. Vargas, "A situation-aware fear learning (safel) model for robots," *Neurocomputing*, vol. 221, pp. 32 – 47, 2017.

CHAPTER

3

HWU-USP DATASET AND FRAMEWORK

In this chapter, the HWU-USP activities dataset is presented, along with different methods for its classification according to multimodal scenarios. This corresponds to the second research question of section 1.1. The outcomes of the previous chapter were applied as the foundations of such developments, published in the paper ”Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors” (RANIERI *et al.*, 2021a), published to Sensors, an open access journal by MDPI with impact factor 3.275 and qualified as A1 in the latest Qualis for Computer Science journals. The dataset was made available at the Dryad Digital Repository (RANIERI *et al.*, 2021b). The documents for ethics clearance of the data collection, it is, the information sheet presented to the participants, the example of the informed consent form, and the confirmation of the ethical approval by the Heriot-Watt University committee, are reproduced in Appendix A. The paper is reproduced in the following pages.

Ranieri, C.M.; MacLeod, S.; Dragone, M.; Vargas, P.A.; Romero , R.F. ”Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors”. Sensors 2021, 21, 768.

Contribution statement: Ranieri and Romero performed literature review and contextualisation; Ranieri and Vargas designed the dataset and managed to obtain approval from the ethics committee; Ranieri, MacLeod and Dragone prepared the environment and performed the data collection sessions; Ranieri designed the machine learning framework, performed the experiments and analysed the results; Ranieri and MacLeod wrote the paper; Vargas, Dragone and Romero performed the revisions.

Article

Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors

Caetano Mazzoni Ranieri [1], Scott MacLeod [2], Mauro Dragone [2], Patricia Amancio Vargas [2] and Roseli Aparecida Francelin Romero [1,*]

[1] Institute of Mathematical and Computer Sciences, University of Sao Paulo, Sao Carlos, SP 13566-590, Brazil; cmranieri@usp.br
[2] Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, EH14 4AS, UK; sam19@hw.ac.uk (S.M.); M.Dragone@hw.ac.uk (M.D.); p.a.vargas@hw.ac.uk (P.A.V.)
* Correspondence: rafrance@icmc.usp.br

Abstract: Worldwide demographic projections point to a progressively older population. This fact has fostered research on Ambient Assisted Living, which includes developments on smart homes and social robots. To endow such environments with truly autonomous behaviours, algorithms must extract semantically meaningful information from whichever sensor data is available. Human activity recognition is one of the most active fields of research within this context. Proposed approaches vary according to the input modality and the environments considered. Different from others, this paper addresses the problem of recognising heterogeneous activities of daily living centred in home environments considering simultaneously data from videos, wearable IMUs and ambient sensors. For this, two contributions are presented. The first is the creation of the Heriot-Watt University/University of Sao Paulo (HWU-USP) activities dataset, which was recorded at the Robotic Assisted Living Testbed at Heriot-Watt University. This dataset differs from other multimodal datasets due to the fact that it consists of daily living activities with either periodical patterns or long-term dependencies, which are captured in a very rich and heterogeneous sensing environment. In particular, this dataset combines data from a humanoid robot's RGBD (RGB + depth) camera, with inertial sensors from wearable devices, and ambient sensors from a smart home. The second contribution is the proposal of a Deep Learning (DL) framework, which provides multimodal activity recognition based on videos, inertial sensors and ambient sensors from the smart home, on their own or fused to each other. The classification DL framework has also validated on our dataset and on the University of Texas at Dallas Multimodal Human Activities Dataset (UTD-MHAD), a widely used benchmark for activity recognition based on videos and inertial sensors, providing a comparative analysis between the results on the two datasets considered. Results demonstrate that the introduction of data from ambient sensors expressively improved the accuracy results.

Keywords: human activity recognition; multimodal datasets; deep learning; video classification; inertial sensors; human–robot interaction

Citation: Ranieri, C.M.; MacLeod, S.; Dragone, M.; Vargas, P.A.; Romero, R.A.F. Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors. *Sensors* **2021**, *21*, 768. https://doi.org/10.3390/s21030768

Academic Editor: Susanna Spinsante
Received: 15 December 2020
Accepted: 21 January 2021
Published: 24 January 2021

1. Introduction

According to projections by the Department of Economic and Social Affairs of the United Nations, the worldwide proportion of citizens aged between 15 and 64, with respect to those aged over 65 years old, is expected to drop from about 7:1 in 2020 to approximately 4:1 in 2050 [1]. This may lead to a deficit in workforce numbers in the elderly care sector, which has motivated the research on Ambient Assisted Living (AAL) [2]. The idea is to support human carers, with the introduction of assistive technologies. These solutions may help to address issues such as improving limitations of movements, monitoring chronic diseases, minimising social isolation or controlling medicine administration by providing integrated services that may be connected to the Internet of Things (IoT) [3].

Technologies for AAL may be provided in the form of smart homes [4], equipped with sensors, for monitoring different conditions of the environment and its inhabitants [5], and actuators, to effectively help them in their daily activities [6]. To enhance those environments and improve their acceptance towards the end users, the design can include service or social robots [7], which may either introduce additional functionalities and monitoring tools, or provide more natural human–robot interaction. One advantage of introducing such robots to an AAL environment is the possibility to collect visual information with less privacy concerns than those related to fixed cameras [8]. Besides, robots may be endowed with behaviours to manage privacy-sensitive situations [9].

Human activity recognition, which consists of classifying human-centred data from different sensors [10], is a key requirement for AAL applications, as it is essential for allowing proactive behaviours or even basic cooperation between human and the environment. The review provided in Chaaraoui et al. [11] presented a discussion on taxonomies for Human Behaviour Analysis (HBA). According to the authors, an *activity* is a sequence of semantically meaningful actions involving interactions between humans and their environment. The most widely adopted approach to HBA involves the classification of the activities from sensor data capturing sequences of basic human motions, i.e., *action primitives*.

To date, most research on this field has focused on single modality approaches, which may consist of either RGB [12] or RGB-D videos [13], wearables such as inertial sensors (Inertial Measurement Units—IMUs) [14], or ambient sensors [15]. The scenarios in which each of these modalities have been employed for activity recognition vary according to the availability of data, which may be constrained by technical or ethical limitations. RGB videos can be found on different online sources, which allows the gathering of different large-scale, very heterogeneous datasets [16]. Depth videos and IMU data are usually collected in more controlled environments, such as AAL research laboratories [17]. For all of those modalities, deep learning (DL) approaches have shown to provide state-of-the-art classification results [18–20]. In the case of ambient sensors, most datasets provides long-term records of binary data, and the associated research effort usually focus on segmenting and classifying human activities [21].

The availability of data from multimodal sources within a smart robotic environment [22] may help designing more robust methods for activity recognition. For instance, although recent advances on DL approaches have made video-based activity recognition a very powerful approach [23], this modality of data may be unavailable due to privacy restrictions, or it may be compromised by technical issues such as occlusions. Besides, one modality of data can perform better than another in certain conditions. Ambient sensors may be quite informative on some well-defined scenarios in a smart home [24], while wearable sensors can be more suitable for actions that rely on limb motions [25]. Therefore, most recently, multimodal approaches for activity recognition have been investigated [20] as more robust alternatives when compared to single-modality approaches.

To the best of our knowledge, there is no work in the literature that addressed the problem of recognising heterogeneous activities of daily living centred in home environments by building modules that consider, simultaneously, data from videos, wearable IMUs, and ambient sensors. One of the reasons is the lack of a representative dataset suitable for this task, which would be a prerequisite to train and test any data-driven model. Nonetheless, this configuration can be expected in smart AAL environments combining smart home and robotic technology.

Driven by this motivation, our first contribution in this work is the design, collection and curation of the Heriot-Watt University/University of Sao Paulo (HWU-USP) activities dataset, which will be made public. This database was built based on an international collaboration between researchers from the Heriot-Watt University (HWU) in the UK, and the University of Sao Paulo (USP) in Brazil (the dataset is available at https://drive.google.com/drive/folders/1Aq1kOcAxLhZl84R9qAdW_o0uL8s5b30E?usp=sharing). The dataset was designed to capture a set of activities of daily living that took place in the Robotic Assisted Living Testbed (RALT) at the Heriot-Watt University, in Edinburgh, Scotland.

It includes not only activities that involve long-term dependencies, such as preparing a sandwich, but also static activities, such as reading a newspaper. Videos were recorded from the RGBD camera from a robot, positioned at a fixed location in the test kitchen. Two wearable IMUs were placed at the dominant arm and at the waist of each participant, exemplifying the inertial sensors usually found in smartwatches and a smartphones. The ambient sensors from the smart home also have been integrated in the environment.

Besides presenting the dataset in detail, our second contribution is the development of a framework based on Deep Learning (DL) networks for classifying multimodal data not only from videos and inertial units, as performed on related work, but also on ambient sensors. To the best of our knowledge, this is the first approach to consider those three modalities altogether, which could not be done with the other datasets present on the literature. The DL models for the different modalities were trained and evaluated with the HWU-USP dataset. Our investigation included approaches for sensor fusion, a non-trivial problem which drives research in different contexts [26], and has been explored in the field of activity recognition [27]. On our case, fusion was performed mostly at decision-level, though one feature-level approach was proposed for the inertial and ambient sensors. A comparative analysis of the results, quantifying the improvements achieved by each approach, was performed.

The classification framework was based on existing literature for each modality. Regarding the video modality, we have considered the two streams proposed by Simonyan and Zisserman [28]: the spatial and temporal streams. As expected, due to the motion-driven aspect of the datasets analysed, with few background information or objects that could be discriminative regarding to the activity being performed, the appearance-based approaches (i.e., the spatial stream) led to poor results, and hence were not considered on the multimodal scenarios. Instead, our architecture focused on motion-based approaches (i.e., the temporal stream), which led to the best accuracies observed for the single-modality approaches. This consisted of combining CNN modules for feature extraction on dense optical flow maps [29–31], previously computed on the video frames, and a LSTM layer for temporal modelling [32].

With respect to the IMU, we introduced the raw, time-domain data to a DL architecture, another common practice in related work [33,34]. The fusion between IMU data and ambient sensors was performed internally as part of one of the DL architectures presented, after both modalities were temporally aligned in a preprocessing stage, an approach that we are proposing as part of this work. To perform fusion between the video-based models and the models that processed IMU and ambient sensors' data, the output vectors were combined with the outputs of the other modalities, also an approach commonly adopted in related research [35,36].

All predictions are performed on two-seconds-long segments. Following a widely adopted approach in the literature in video-based activity recognition [18,28,35], we have evaluated our models on 25 segments equally spaced between them. We did the same for the other modalities as well, since this approach allows the classifiers to consider partial observations of the activities, as expected for real-world scenarios. Results are presented in terms of the accuracy obtained in each of the conditions analysed, corresponding to different input modalities or fused models. The introduction of ambient sensors has shown to provide significant improvements to the overall accuracy. The results presented here provide a baseline for future work in human activity recognition using multi-modal sensor data in smart robotic environments.

Besides the new HWU-USP dataset, we have also experimented our video and IMU models with another popular public available dataset, the UTD-MHAD [37], providing comparisons with the HWU-USP dataset regarding to the behaviour of the classifiers. Moreover, the classification methods achieved competitive results for the UTD-MHAD. The confidence in predicting the correct label on each segment was also analysed. As was expected, this was quite different when comparing the HWU-USP dataset, consisting of

both complex and simple activities, to a more homogeneous dataset, such as the UTD-MHAD.

The remainder of this article is organised as it follows. Section 2 illustrates and compares the most relevant datasets from the literature, and highlights their key differences from the one presented in this paper. Section 3 provides an overview of sensor-based human activity recognition, focused on techniques able to exploit different input modalities. Section 4 presents a detailed description of the data in the proposed dataset and the protocol used for its collection. Furthermore, it describes the DL methods considered for the classification of data and also the protocols used for their training and evaluation. The results are then presented and analysed in Section 5, and a discussion is presented in Section 6. Finally, in Section 7, conclusions and possible directions for further research are outlined.

2. Datasets of Human Activities

The HWU-USP dataset was built to provide a benchmark for studies on activity recognition in indoor environments. For this reason, combinations of different modalities, namely videos, wearable IMUs, and environmental sensors were considered. In this section, previously developed datasets that includes sensor data from these modalities, regardless of the context, will be presented, in order to contextualise the construction of the HWU-USP dataset. The nature of available datasets and associated approaches for data collection vary greatly for different sensor modalities considered in human activity recognition research. For example, for RGB video datasets, there is a vast availability of data on the Internet, from movies or other non-dedicated sources, which can be labelled and made available, resulting in fairly large datasets. This is more difficult for depth videos, IMUs or environmental sensors, hence this type of datasets are more often collected in controlled settings, usually in research laboratories simulating domestic environments. In the next subsections, datasets for each modality or set of modalities will be presented separately.

2.1. RGB Videos

As already mentioned, most commonly used benchmarks of regular RGB videos can avail of amateur videos, movies or sports broadcasts. Most of these datasets are pre-segmented, which means that each video is entirely associated to one category (e.g., "biking" or "playing piano"), with a few exceptions. The categories in which the activities of these datasets are usually labelled are generally at a comparatively high level of abstraction and granularity, including activities such as *playing basketball*, instead of low-level, primitive activities such as *walking* or *running*. A summary of representative RGB video datasets is provided in Table 1.

Table 1. Video datasets made available and widely used in related works.

Dataset	Number of Instances	Categories	Source	Pre-Segmented
UCF101 [16]	13,320	101	YouTube	Yes
HMDB51 [38]	6766	51	Movies, YouTube, etc.	Yes
CCV [39]	9317	20	YouTube	Yes
Hollywood2 [40]	1707	12	Movies	Yes
Sports-1M [41]	+1 M	487	YouTube	Yes
Kinetics 700 [42]	+600 K	700	Youtube	Yes
THUMOS [43]	+23,700	101	YouTube	No
ActivityNet [44]	13,837	203	YouTube	No

The two most relevant benchmarks, on which the most renowned video-based HAR techniques have been evaluated, are the UCF101 [16], from the *University of Central Florida*, and the *Human Motion Database* (HMDB51) [38]. The *Columbia Consumer Video Database* (CCV) [39] is also commonly referenced, as it presents similar properties, but longer videos. The Hollywood2 [40] and the Sports-1M [41] datasets present an additional challenge, as the videos contain editions and camera transitions. Although, as seen in Table 1, the Sports-1M dataset is quite large, a newer dataset—the Kinetics dataset [42]—has been preferred for testing DL architectures, requiring a large amount of data. Regarding datasets that were not pre-segmented, some of the most relevant ones are the THUMOS [43] dataset, provided with the same set of categories as the UCF101, and the ActivityNet [44], annotated according to a semantic hierarchy of activities designed by the U.S. Department of Labour to perform the American Time Use Survey (ATUS).

All of the above-mentioned datasets consist of heterogeneous and realistic sets of videos, usually thanks to user-created content. This variety of data is not possible, at least at present, for data from other modalities, such as RGB and depth videos, wearable and environmental sensors. Consequently, multimodal datasets are usually collected in controlled environments, mostly with static backgrounds, few variations in camera angles and artefacts shared among the data samples. These limitations are inherent to any dataset consisting of modalities that does not count on large amounts of user-created content, which is the case for almost all multimodal datasets, including ours.

2.2. Depth Videos

With the popularisation of RGBD (RGB + depth) cameras, such as the Microsoft Kinect [45], it became possible to provide not only RGB and depth videos, but also previously extracted skeleton joints from humans being observed. The categories within these datasets are usually from levels of abstraction compatible with those that could be acquired by RGBD devices, although less diverse, with several activities sharing the same background, objects for manipulation and light conditions. In Table 2, a collection based on the datasets adopted by Amir Shahroudy et al. [46] is shown. These datasets presented were collected using a Microsoft Kinect device, except for the NTU RGB+D, which was collected using a Microsoft Kinect v2. Both devices may collect data on either 15 Hz or 30 Hz.

Table 2. Selection of datasets for depth videos, adapted from the list by Amir Shahroudy et al. [46].

Dataset	Classes	Subjects	Repetitions	Instances
ORGBD [47]	7	24	2	336
MSR-DailyActivity3D [48]	16	10	2	320
3D Action Pairs [49]	12	10	3	360
RGBD HuDaAct [50]	13	30	-	1189
NTU RGB+D 120 [51]	120	106	-	114,480

The datasets listed at Table 2 share a lot of common points. The Online RGBD Action dataset (ORGBD) [47] contains videos from different environments, allowing *cross-environment* evaluation of HAR techniques. The MSR-DailyActivity3D [48] is characterised by a higher intra-class variation. The 3D Action Pairs [49] was designed to include pairs of opposite activities, such as *pull a chair* and *push a chair*. An initiative for providing a larger dataset resulted on the RGBD HuDaAct [50]. Finally, the NTU RGB+D was extended and formed the NTU RGB+D 120 dataset [51], with more than 100K videos distributed on 120 categories.

2.3. Wearable and Ambient Sensors

In this subsection, we are addressing sensors that may be worn by the subjects (i.e., wearable sensors) or placed at predefined locations of the environment (i.e., ambient sensors). We focused our review in inertial measurement units (IMU), since most multimodal datasets address this modality. However, we also referenced setups including sensors embedded in the environment, usually at fixed locations, because these can help to get very discriminative information. This is the case of our own dataset, which includes data from different sensors from a smart home, as discussed in Section 4.1. The data provided by these devices usually consist of measurements from accelerometers, gyroscopes, and, sometimes, magnetometers, all of them three-dimensional. All datasets examined in Table 3 were collected under controlled conditions, with the sensors placed on the surfaces of objects or, most commonly, as wearable devices.

Table 3. Datasets based on environmental or wearable sensors. Except for the OPPORTUNITY dataset, the IMUs were all contained on wearable devices.

Dataset	Sensors	Rates	Attributes	Subjects	Classes
OPPORTUNITY [52]	Wearable accelerometers: 12 Wearable IMUs: 7 Wearable tags: 4 Objects' accelerometers: 12 Objects' gyroscopes: 12 Environmental accelerometers: 8 Switches: 13	64 Hz 30 Hz 87 Hz 64 Hz 64 Hz 98 Hz 100 Hz	242	4	17
PAMAP2 [53]	Colibri wireless IMUs: 3 Heart monitor: 1	100 Hz 9 Hz	52	9	18
REALDISP [54]	Xsens IMUs: 9	50 Hz	120	17	33
SBHAR [17]	Samsung Galaxy S2 IMU	50 Hz	561	30	12
Skoda [55]	IMUs: 20	98 Hz	141	1	10
DG [56]	IMUs: 3	64 Hz	9	10	2

The OPPORTUNITY [52] dataset has been widely used as benchmark in the literature for activity recognition tasks involving wearable or environmental sensors, as it consists not only of several inertial sensors placed in objects of daily living and worn by the subjects, but also tags and switches positioned in different parts of the environment. Another widely adopted dataset is the *Physical Activity Monitoring for Aging People* (PAMAP) and its extension, the PAMAP2 [53], designed for identifying patterns in subjects performing physical exercises. The *Realistic Sensor Displacement Benchmark Dataset* (REALDISP) [54] also addresses physical activities. The positioning and availability of sensors are not usually practical and intended for large-scale adoption, except when dealing with standardised conditions, such as smartphones, as addressed on the *Smartphone-Based Human Activity Recognition dataset* (SBHAR) [17]. Datasets for other scenarios have also been developed, such the *Skoda Mini Checkpoint dataset* [55], composed of work activities in a car factory, and the *Daphnet Gait* (DG) [56], composed of motion patterns of patients affected by Parkinson's Disease.

Bakar et al. [57] presented an extensive survey on sensing approaches for activity recognition in smart homes. Besides cameras, microphones and wearables, these environments allow the introduction on fixed sensors such as temperature, pressure or motion sensors. Binary sensors, such as switches at doors and wardrobes, are also usual, and these categories were also included on our approach. The CASAS project [58] proposed different testbeds that could be used for data collection and experiments in smart homes, based mostly on environmental sensors. Differently from the datasets mentioned in Table 3, these datasets usually result from long-term data collections. As detailed by Lesani et al. [59],

the Twor2009, Tulum2009 and Tulum2010 datasets, from the CASAS project, were collected in periods ranging from 3 to 6 months, in which information from motion, binary, door and item sensors were recorded.

An intrinsic advantage of the above-mentioned modalities is that they can provide additional data that are invariant to the positioning of externally placed observing devices, contrary to when cameras or robots are used. Thus, they may provide valuable information for an activity recognition framework. Besides, inertial and ambient sensors also have the advantage of being less intrusive than video cameras. For this reason, multimodal datasets, including video, IMUs and other modalities, have been proposed.

2.4. Multimodal: Video and IMU

Multimodal datasets with videos and other sensors, especially IMUs, have been proposed in different contexts. Most of these datasets report data from combinations of different sensors and depth videos, which may be accompanied by the RGB videos. A survey on the subject was provided by Chen et al. [27], considering only datasets that provided depth videos and IMU data. In Table 4, we present a collection of the most relevant datasets for any kind of video collected along with data from other sensors.

Table 4. Multimodal datasets, provided with videos, IMU sensors, and possibly others.

Dataset	Sensors	Rate	Subjects	Classes	Instances
CMU-MMAC [60]	Cameras: 5 Microphones: 5 Wired IMUs: 5 Wireless IMUs: 4 Motion capture: 1 eWatch (accelerometer)	30 Hz or 60 Hz - 120 Hz 60 Hz 120 Hz -	18	5	90
Berkeley-MHAD [61]	Motion capture: 8 Stereo cameras: 2 Quad cameras: 2 Microsoft Kinect: 2 Shimmer IMUs: 6 Microphones: 4	480 Hz 22 Hz 22 Hz 30 Hz 30 Hz 48k Hz	12	11	660
UTD-MHAD [37]	Microsoft Kinect: 1 IMU: 2	30 Hz 50 Hz	8	27	861
C-MHAD [62]	Webcam: 1 Shimmer3 IMU: 2	15Hz 50Hz	12	12	240
50 Salads [63]	Microsoft Kinect: 1 Accelerometers: 11	30 Hz 50 Hz	25	51	966
JIGSAWS [64]	da Vinci (kinematic data): 1 Stereo camera: 1	30 Hz 30 Hz	8	15	103
ChAirGest [65]	Microsoft Kinect: 1 Xsens IMUs: 4	30 Hz 50 Hz	10	10	1200
UR Fall Detection [66]	Microsoft Kinect: 1 x-IMU: 1	30 Hz 256 Hz	5	5	70
TST Fall Detection V2 [67]	Microsoft Kinect: 1 Shimmer IMUs: 2	30 Hz 50 Hz	11	8	264

The *Carnegie Mellon University Multimodal Activity* (CMU-MMAC) Database [60] records data from RGB cameras, microphones and wearable sensors worn by a set of subjects performing food in a kitchen environment. The *Berkeley Multimodal Human Action Database* (Berkeley-MHAD) [61] and the *University of Texas at Dallas Multimodal Human*

Action Database (UTD-MHAD) [37] have a similar structure based on short actions recorded with RGBD cameras, IMUs and a variety of other sensors. The recently deployed Continuous Multimodal Human Action Dataset [62] was collected in an environment similar to that of the UTD-MHAD, although without the RGBD camera, containing activities related to smart TV gestures (5 activities) and transitions (7 activities) in continuous, non-segmented recording sessions. The *50 Salads* dataset [63] captures people preparing several salad recipes being recorded by RGBD cameras, and IMUs placed in the utensils used for the food preparation. The ChAirGest [65] dataset focuses on gesture recognition with the aim to be applied in human–computer interfaces. The *University of Rzeszow Fall Detection Dataset* [66] and the *Telecommunications System Team* (TST) *Fall Detection Dataset* [67] were built with data on regular daily activities and falls, which can be used to train models for fall detection, an important field of research with applications as part of AAL solutions for the elderly.

Although the above-mentioned datasets cover a range of applications for multimodal activity recognition, none of them focused generically on activities of daily living in AAL environments. Moreover, none of them are provided simultaneously with data from videos, inertial units and ambient sensors. Our approach aims to alleviate this gap by providing a dataset captured in a heterogeneous, sensory rich environment comprised of a smart home system, a wearable sensor kit, and a domestic robot equipped with an RGBD camera.

3. Human Activity Recognition

Different algorithms can be suitable for the task of human activity recognition, depending on the nature of the data being addressed [68]. For RGB videos, although strategies based on classic feature extraction techniques still provide competitive results [69–71], Deep Learning (DL) architectures have led to increasingly accurate, state-of-the-art results, representing a very active field of research, as discussed by Zhang et al. [72]. Among the most influential studies on this subject is Simonyan and Zisserman [28], which presented the Tow-Stream ConvNets, characterised by a *spatial* and a *temporal* stream. The spatial stream consists of a Convolutional Neural Network (CNN) that classifies individual RGB frames from a video, while the temporal stream is a similar CNN which, instead of an individual image, processes a sequence of dense optical flow maps (horizontal and vertical), computed on a preprocessing step using a suitable algorithm [29–31], from a predefined number of frames. The scores obtained by both streams are then fused, in order to obtain a prediction. Most of the works found in literature built on the basic structure of the Two-Stream ConvNets, including the Temporal Segment Networks [18]. Recent literature has proposed different multiple-stream approaches that could include other input modalities [12]. Our work was based on the multiple stream paradigm, in which the temporal stream was extended to work with a combination of CNN and Long Short-Term Memory (LSTM), as proposed by Donahue et al. [32]. It is worth to notice that spatio-temporal approaches, usually based on 3D CNNs, have been a popular alternative to multiple-stream approaches such as ours [73–75]. In this paper, the approaches implemented for video classification consisted of combining multiple stream principles using optical flow maps, with feature extraction with a CNN and temporal modelling with LSTM.

With respect to depth videos, state-of-the-art results have been obtained from different approaches. Motion from depth images, including optical flow features computed over depth human silhouettes, along with features exracted from human joints, are usually employed to compose Hidden Markov Models (HMM) [76–81], or other representations such as Self-Organising Maps (SOM) [82]. The most successful approaches are based on features extracted from geometrical relationships on skeleton joints [83]. In the context of DL, some researchers investigated the introduction of preprocessing steps such as the computation of depth motions maps [84], or the computation of action maps from scene flow representations [19]. We did not include the depth videos as a modality for computing the temporal stream because there is not a direct correspondence between the preprocessing steps of the most successful approaches on this context and the algorithms that we have

analysed for the other modalities. The three-dimensional version of the optical flow, the scene flow, could be computed based on RGB-D images [85], but led to poor results on exploratory experiments and, hence, were unconsidered. Nonetheless, we included the raw depth images as an additional condition for analysing the spatial stream.

Considering sensors other than video cameras, the survey by Wang et al. [86] defined four modalities: body-worn (i.e., wearable sensors such as smartphones or watches), object (i.e., sensors attached to objects, such as RFID or IMUs attached to utensils), ambient (i.e., sensors attached to to environment, such as door sensors or Bluetooth beacons), and hybrid (i.e., combinations of modalities, typical for smart environments). Here, we are interested on body-worn (specifically IMUs) and ambient sensors, which composed a hybrid setting for our experiments.

Regarding activity recognition based on IMUs, research has addressed scenarios that resemble devices that are expected to be actually worn by the users, such as smartphones and smartwatches [87]. Feature extraction methods include combinations between sequential minimal optimization (SMO) and Random Forest [25], statistical features feeding genetic algorithms [88], and Markov models [89]. DL architectures, such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), autoencoders, Restricted Boltzmann Machines (RBM), and Recurrent Neural Networks (RNN) have also been successfully applied to this modality [33]. In this paper, we designed a module for inertial sensors that resembled the DeepConvLSTM by Rueda and Fink [90], in which a convolutional module would perform feature extraction and feed it to an LSTM layer.

Considering ambient sensors, approaches can be divided into two major categories; data driven and knowledge driven. Domain Knowledge based systems use ontology's and semantic reasoning to aid in recognition. Chen et al. [91] and Liciotti et al. [92] used a knowledge driven approach, including a Partially Observable Markov Decision Process (POMDP) and exploited the task information, while the location is combined with the sensor events in the smart home. Data-driven is mainly focused on use of feature extraction, temporal clustering, and activity recognition. Medina-Quero et al. [93] proposed a method using fuzzy time windows (FTW) to segment the data set, followed by Long Short-Term Memory (LSTM) for activity recognition. Gochoo et al. [94] extracted fixed-length sliding windows into a sparse two-dimensional time matrix to use Convolutional Neural Networks (CNN) for activity recognition. Guo et al. [15] provided a data-driven framework for activity recognition from multiple residents using time clustering.

Although different possibilities for fusion of multimodal data using DL methods have been proposed, especially regarding to different inputs from multiple video streams [95], the most popular approach for dealing with heterogeneous data is to process each modality separately and fuse the obtained scores at a later stage [35,36], which we refer as *late fusion*. This was the approach adopted on all experiments performed in this paper. Considering neural networks, variation of this method that has been adopted is to fuse the outputs of the modules respective to each modality using a fully-connected layer [75]. Other approaches have also been proposed, such as the Correlational Recurrent Neural Network (CorrRNN) presented by Yang et al. [96].

4. Methods

The experiments were performed in order to evaluate the improvements that could be achieved by combining motion information from videos and inertial sensors with static, contextual information from ambient sensors at a smart home. The task was to classify high-level activities, possibly composed by complex sequences of actions, using time-localised data, which is certainly a requirement for a real-time decision-making system. In the literature review, summarised in Section 2, we did not find a dataset suitable for such analyses. Hence, we designed a data collection procedure and collected the HWU-USP dataset, presented in the next subsection. This dataset captures a set of daily activities performed in the simulated kitchen at the Robotic Assisted Living Testbed (RALT), part of the Edinburgh Centre for Robotics in Edinburgh [97]. It was recorded with ambient sensors

such as switches installed on wardrobes and drawers, inertial sensors attached to the wrist of the dominant arm and to the waist of the participants, and videos recorded from the RGBD camera of a domestic robot placed in a fixed observing location.

Besides the construction of the HWU-USP database, we have also performed experiments with the UTD-MHAD [37], already mentioned on Section 2.4, one of the most widely used benchmarks for multimodal activity recognition from videos and IMU. This known dataset differs from ours on its granularity, with actions composed of short movements such as *clap*, all of them with approximately the same length of very few seconds. It also provides more homogeneous data, with the subjects cautiously positioned facing the camera, always in full face (on the HWU-USP dataset, images in profile and in full face are presented on different frames of the videos). Although this is suitable for work on gesture-based interfaces, it is realistic for daily activities such as the ones that we are interested in. Moreover, it is focused on motion information and does not provide data from ambient sensors, which limits our analyses. Another difference is that, whereas the HWU-USP dataset is provided with two inertial units placed on the subjects' waist and dominant wrist, the UTD-MHAD provides inertial data from only one unit, worn on the subject's right wrist. Nevertheless, it may provide an alternative benchmark for our evaluations, besides allowing comparisons with our dataset on the performance of successive predictions over time using the same classifiers. Those analyses will be better described on Section 4.3.

As for the classifiers, we built on DL architectures for data from video and inertial sensors, presented on our previous work [36]. The most relevant contributions of this paper are the models trained not only on data from those modalities, but also considering ambient sensors from the smart home. This data was pre-processed to compose tuples of structured, categorical data which could be introduced as an additional stream to be introduced on the top of the neural network originally implemented for classifying IMU data. The design of the resulting neural network will be depicted on Section 4.2.

4.1. The HWU-USP Activities Dataset

The multimodal datasets presented in Section 2 provide data from different kinds of videos and inertial sensors, but they did not include data from ambient sensors. The main contribution of our dataset is introducing the data from the smart home devices synchronously with videos and inertial sensors. Moreover, we provided videos of either activities made of repetitive patterns, such as *reading a newspaper*, and more complex activities with long-term dependencies, such as *preparing a sandwich*. This makes our dataset more realistic regarding the set activities, with respect to what could be expected on an actual AAL scenario, when compared to the others.

As already mentioned, the data collection was performed at the RALT laboratory [97]. The RALT is a 60 m^2 (square meters), fully sensorised space designed to resemble a typical single level home comprising an open-plan living, dining and kitchen area and a bathroom and bedroom, and hosting a number of smart home, assistive technologies and domestic robots, such as the TIAGo robot, manufactured by Pal Robotics [98]. Besides collecting data from the smart home, people being recorded were asked to wear a wearable kit comprising of a smart watch and a sensor device to be installed on the belt, both equipped with IMUs. Furthermore, a Tiago robot was placed at a fixed location, to record data from its RGBD camera.

The data collection procedure received ethics approval from the Heriot-Watt University ethics committee on the 17th of November, 2019. A set of 16 volunteers participated on this study over the period of 2 weeks, performing a single repetition of each of the nine activities included in our protocol. This was to ensure to capture a degree of variability in the performance of each activity being recorded, including different timings for primitive actions and overall activities performed by different people. The participants, healthy volunteers with neither functional nor cognitive impairments, signed a consent form and the data collected did not include their identity (participants' faces were also

made anonymous by blurring the recorded image). Each participant was brought into the lab and the activities were explained and the participants were set up with the IMU's in the kitchen. The following subsections will provide more information about the sensors and protocols used for data collection.

4.1.1. Sensors and Modalities

The RALT is a "Living-Lab" home-like environment designed to facilitate user-driven design and testing of innovative information and communications technologies (ICT) and robotic solutions for healthy ageing and independent living. In Figure 1, the whole environment is illustrated.

This environment is equipped with ambient sensors to perceive, monitor and understand occupancy's daily activities. The sensors are positioned around the household with uniquely identified identity (ID), together with timestamp to indicate and record of occupancy's activity. In our dataset, we recorded the sensors that were available in the kitchen and that would be meaningful for our purposes. Specifically, we considered four binary switches, two of which were positioned at the doors of two cupboards, respectively containing mugs and dishes, one at the door of the fridge, and one at a drawer used to store cutlery. We also considered the PIR sensor present in the kitchen, and the power measurements by the kettle, used for preparing tea.

Figure 1. Environment in which the data collection was performed, with the TIAGo robot positioned on the corner of the kitchen (on the right side) during the recording sessions.

The TIAGo robot is a mobile service robot designed to work in indoor environments. It has an extendable torso and a manipulator arm to grab tools and objects. Its sensor suite allows it to perform a wide range of perception, manipulation, and navigation tasks and is used for assisted living research in the RALT. For our data collection, we considered only data from its RGBD camera, an Orbbec Astra [99] device installed in its head. According to the manufacturer's specifications, the range of this depth sensor lies within 0.6 and 8 meters. We positioned the robot in such a way that all activities and objects of interest were within this range. The colour VGA 640 $\times$ 480 at 25 fps and depth stream mode VGA 640 $\times$ 480 at 15 fps were used for the HWU-USP dataset. The TIAGo robot was placed in the environment with a clear view of the participants, at a fixed viewpoint across all recording sessions (see Figure 1).

As an wearable device for providing IMU measurements, we adopted the MetaMotionR, by MbientLab [100], a commercial device equipped with inertial, temperature, light and humidity sensors. The participants were asked to wear two MetaMotionR units, one of which placed at a wristband worn at the subject's dominant arm, and the another placed at a clip worn at the subjects waist. These devices and placements are shown in Figure 2. We recorded data from the accelerometers and gyroscopes, synchronised using the robot's internal clock. A sample from the dataset, considering the different modalities present in the dataset, is shown in Figure 3.

Figure 2. Inertial sensors attached to (**a**) a waist clip; and (**b**) a wrist band.

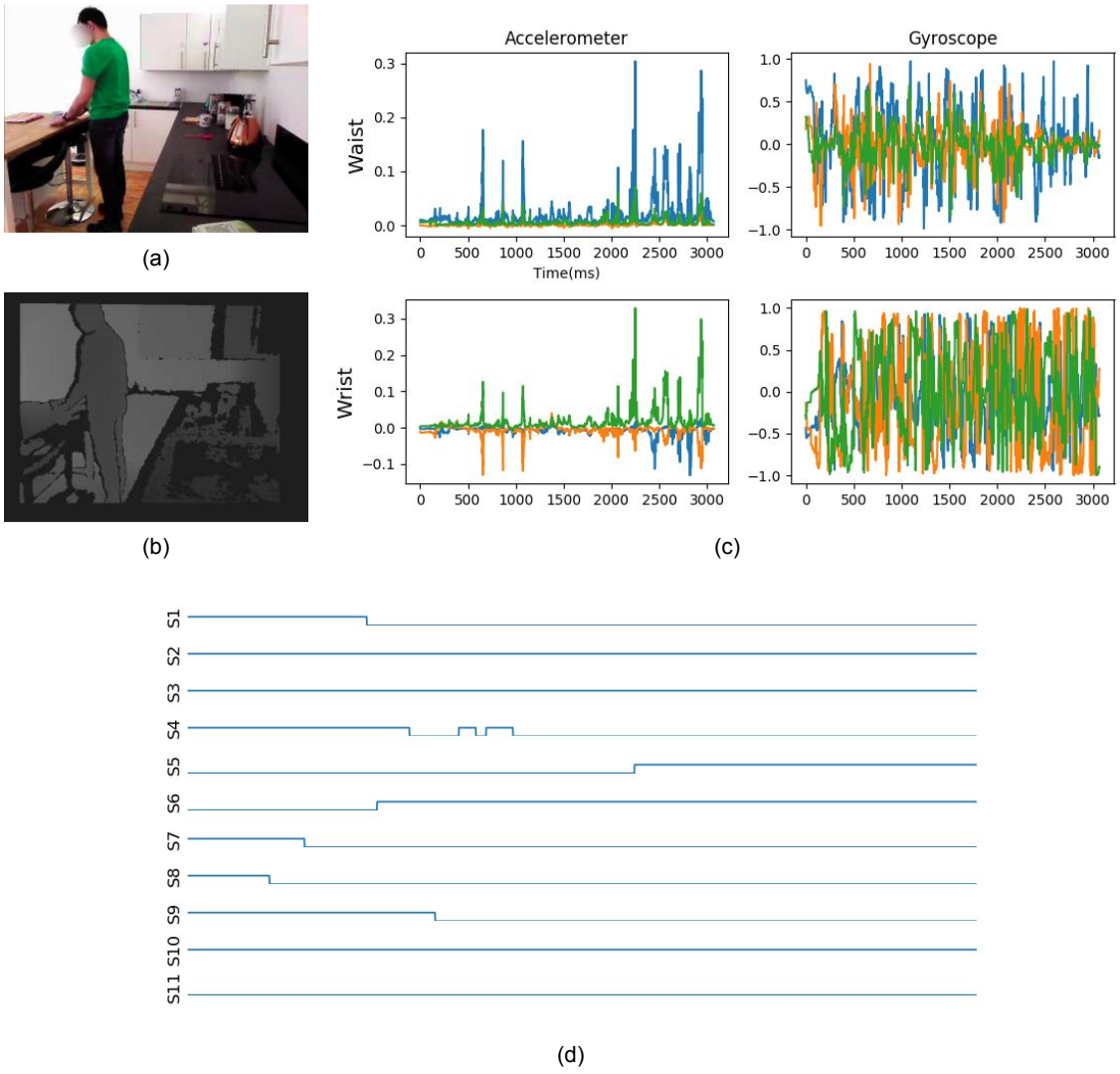


Figure 3. Sample of the dataset collected, consisted of (**a**) an RGB and (**b**) a depth image, both related to one timestep; (**c**) raw data from the inertial sensors, related to a whole sequence; (**d**) raw data from the ambient sensors (binary), where S_k correspond to one of the k sensors available.

4.1.2. Activities List

The activity list was based on the types of activities usually performed in kitchen environments. These were activities of daily living (ADL), tasks that require a level of functional capability and are completed in everyday independent living, such as cooking and cleaning. The activities also required that the participants manipulated a variety of

objects and furniture in the kitchen, especially the cupboards, the simulated fridge and the drawer for cutlery, all of them equipped with binary switches. The participants were also asked to complete the list of activities in their own time. Intervals between recording each activity were implemented, such that the participants could look over the activity list and solve any doubts. The tasks were explained to the participants prior to completing the task, in which they were given a specific order and scripts to complete each of the tasks, such as the location of the items they were instructed to use, and relevant locations where they needed to carry out different actions. Since we were not recording sound, we gave instructions during the completion of the activities as well, so that the participants were not required to necessarily memorise all details respective to each activity. The data collection lasted approximately 20 min per participant as they completed the following set of activities, as illustrated in Figure 4. Notice that those activities have variable lengths, ranging from about 30 s to almost 2 min.

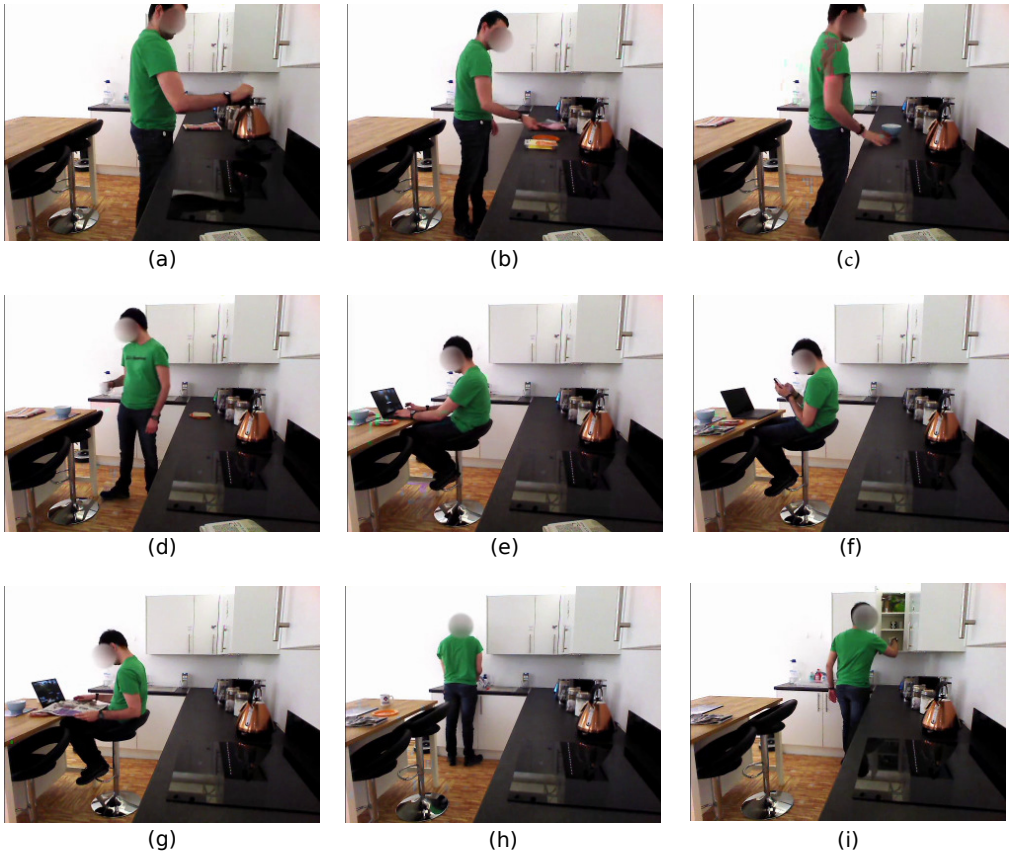


Figure 4. Sample frames of the activities considered for the dataset. (**a**) making a cup of tea; (**b**) preparing a sandwich; (**c**) preparing a bowl with cereals; (**d**) setting up the table; (**e**) using a laptop; (**f**) manipulating the cell phone; (**g**) reading a newspaper; (**h**) washing the dishes; (**i**) cleaning the kitchen.

- *Making a cup of tea*: taking the kettle to the sink filling the kettle, turning it on, collecting a mug and teabags from separate cupboards before combining and filling with water.
- *Making a sandwich*: collecting of a plate, bread, ham and cheese from the respective cupboards and fridge, and assembling with all the ingredients on the worktop.
- *Making a bowl of cereals*: collecting of the spoon at the cutlery drawer, the bowl and the cereal from separate cupboards, and the milk and honey from the simulated fridge, placing everything on the worktop and assembling.
- *Setting the table*: moving the prepared sandwich, tea and cereal from the worktop to the place mat on the kitchen table.

- *Using a laptop*: using a laptop while sat at the kitchen table, complicated with the cluttered environment from the previous activities.
- *Using a phone*: similar to "using a laptop", but with a phone device instead, with both the laptop and the meal at the table.
- *Reading a newspaper*: similar to "using a phone", but with the participant reading the newspaper.
- *Cleaning the dishes*: taking the bowl of cereals to the waste bin, dispose it from its content using the spoon, then pretend to wash it in the sink using a sponge. Due to the position of the sink respective to the robot's positioning, the participant partially obscures this activity.
- *Tidying the kitchen*: returning the items to the cupboards and moving throughout the kitchen environment.

Each activity was performed from the same starting point to avoid classification due to starting configuration alone. The participants would walk into the kitchen environment and begin the activity. Once the activity was completed then the recording of the data was stopped. The starting positions of the objects in the smart kitchen environment was kept constant through the course of the data collection, to ensure consistency of the dataset.

The statistics regarding the lengths of the recordings, for each activity considered, are shown in Figure 5. The resulting dataset was composed of a total of 144 instances (i.e., 16 subjects performing a set of nine activities), which summed about 116 min. As shown in the figure, the average length of of the activities was around 48 s, which is considerably larger than the recordings of most other datasets (see Section 2). An important observation is that the proposed activities were designed at a high level of abstraction, so that most of them were composed by complex sequences of shorter-term actions. For example, to prepare a sandwich, the participant had to place a plate on the board, take the ingredients from the fridge, take the bread from the cupboard, assemble the sandwich, and so on. Based on the timestamps of the videos, fine-grained annotations may be provided as needed, so that each of those actions could be treated as a separate label. This could provide a different, more challenging scenario to be addressed on future research. In this paper, however, we are interested on the presentation of the data collection procedure, the dataset, and the multimodal framework for classification, which operates on high-level activities.

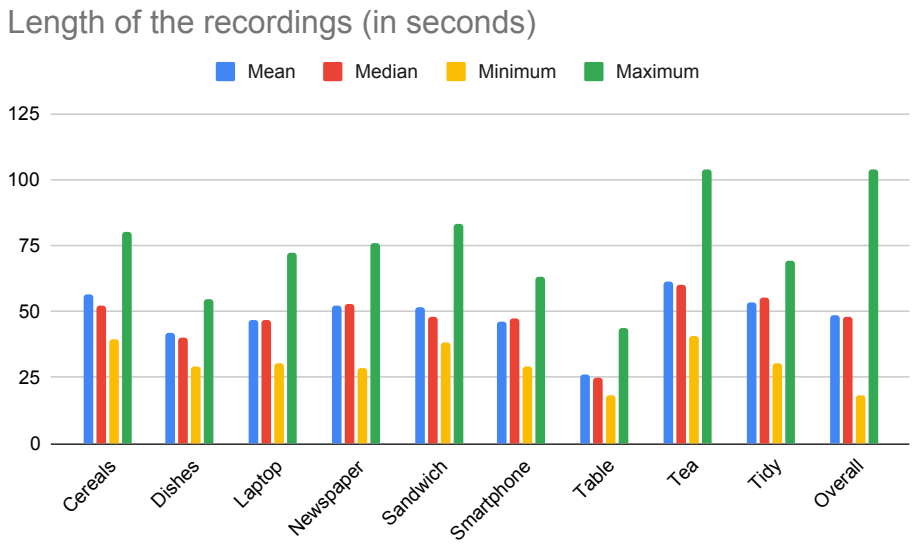


Figure 5. Statistics regarding the lengths of the recordings for each of the activities in the dataset.

4.2. Classification Framework

For feature extraction and classification of activities recorded with videos and inertial data, we have proposed different DL architectures and compared the resulting models and their accuracies on a previous work [36]. Based on the results obtained in this previous paper, we chose the CNN and LSTM models as basis for our experiments. As it already mentioned, our main contribution was the introduction of contextual data from the ambient sensors of the smart home as an additional stream, specifically by including an additional input stream at the neural networks aimed at the inertial data. The different scenarios considered will be presented on the next subsections.

4.2.1. Segment-Wise Classification and AAL Applications

Before presenting the methods proposed for multimodal activity recognition, it may be worth to discuss the type of applications that could benefit from either the HWU-USP activities dataset or the classification framework to be presented in Section 4.2. An AAL application that could be addressed is shown in Figure 6, which summarises scenarios proposed on related work [101]. The sensors made available on the data collection presented in Section 4.1.1, illustrated in the figure, provide inputs for the *activity recogniser* module, which is the focus of this paper. In an actual AAL environment, data would be gathered continuously from the available sensors, and predictions would be provided at each instant t. These predictions, referred in the figure as $pred(t)$, consist of the outputs of the framework presented in Section 4.2, which will be evaluated and discussed in Sections 5 and 6.

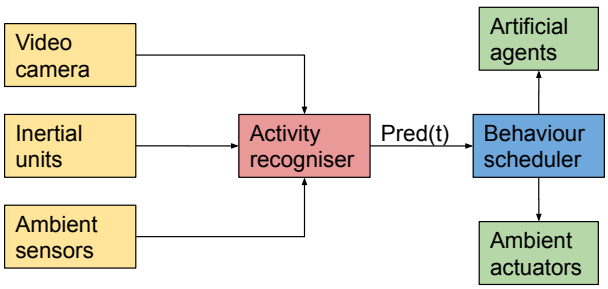


Figure 6. Example of an AAL scenario expected to be addressed by the proposed framework.

The next module that would be part of such an application would be the *behaviour scheduler*, a possible direction for future research. This module would be responsible for orchestrating the different ambient actuators and artificial agents (e.g., social robots or mobile applications), providing useful services or proactive behaviours for the inhabitants of the environment. For a real-time application, these behaviours are expected to be continuously adapted according to the predictions of the activity recogniser at each instant t.

State-of-the-art methods for multimodal activity recognition have been achieved remarkable results by processing previously segmented activities on its whole length [75]. Although this approach makes sense in the case of fine-grained activities, it would be of little use in contexts such as the scenario of Figure 6. There are two reasons for it. First, it requires that the activities have been previously segmented, which is not realistic for real-world applications. Second, it would require the activity to be finished before providing a reliable prediction, which could take more than a minute in the case of the activities of the HWU-USP dataset (see Figure 5). In this case, it is possible that the proactive behaviour of the AAL environment is no longer required, or does not make sense to be performed after the human activity is finished. For example, a robot may need to bring the user's glasses while he is reading the news—it would make little sense to do so after the user has stopped this activity.

Therefore, we designed our framework so that the DL architectures process two-seconds-long segments, and the predictions over a longer sequence could be enhanced at decision-level, by averaging the output vectors at each segment. This is an approach

commonly used for video-based activity recognition [28,35], and we extended it to the other modalities to provide a framework that is able to work with partial data.

4.2.2. Data Preprocessing

Regarding the video modalities, following the proposal by Simonyan and Zisserman [28], we considered multiple streams. Videos were resized to 320×240 before any other preprocessing step. For data augmentation, we implemented random cropping for training, and cropping of all corners and the centre for testing, resulting in frames of size 224×224.

The spatial stream could be composed by individual RGB frames obtained from the videos, as on the original framework. We also adopted a similar approach for taking the depth frames as inputs. To do so, the depth frames had been converted to 3-channel, 8-bit RGB inputs with the same intensity on all channels, composing grayscale samples which could be employed in transfer learning scenarios.

Two approaches were considered for the temporal stream. The first was to feed the learning architectures with pairs of dense optical flow maps, as in the original two-stream ConvNets [28]. Those maps were generated with OpenCV implementation of the TVL1 algorithm [31] on each pair of successive frames on the RGB videos previously converted to grayscale images. The outputs of those algorithms consist of the horizontal and vertical estimations of the displacements of each point from one frame to another, assuming their intensities are preserved on both images. In the case of dense optical flow, all pixels on the image might be considered.

In relation to the inertial and ambient sensors, the recordings were made asynchronously. The alignment was performed independently for each of the 144 recording sessions of the dataset, so far referenced as *instances*. Regarding the inertial sensors, consider that, for a given instance, there is a set of P_{sk} rows of data from a sensor s_k, $k \in \{1,2\}$, with s_1 being the inertial unit of fixed to the user's waist, s_2 the inertial unit fixed to the user's wrist (more sensors could be added to this framework, as needed). Let $\mathbf{x}_{sk}(p)$ be a vector correspondent to the p-th row of data registered by sensor s_k, correspondent to a timestamp $t_{sk}(p)$ obtained from a global clock during the data collection procedure. The alignment procedure intends to obtain an aligned file composed by Q rows, equally sampled at a desired sampling rate r, starting from the highest timestamp registered by any of the sensors. The vector $\mathbf{y}(q)$ is the q-th row of data aligned from both sensors (i.e., the output data). The timestamp correspondent to this row of data is $t_y(q)$, computed as in Equation (1). For each index q, the method consisted of composing a concatenated aligned row $\mathbf{y}(q)$, composed of data from both sensors, by appending the tuple of data $\mathbf{x}_{sk}(i)$, $i \in [1, P]$, from each sensor s_k, so that $t_{sk}(i)$ is the lowest value among the P rows in the instance that satisfies $t_{sk}(p) > t_y(q)$.

$$\begin{cases} t_y(0) = \max\{t_{s1}(0), t_{s2}(0)\} \\ t_y(q) = t_y(0) + q \cdot (1/r) \quad , \quad q = 1, 2, \ldots, Q \end{cases} \tag{1}$$

For preprocessing the smart home data, the same alignment procedure was used to provide one tuple for each timestamp, allowing a one-to-one correspondence with each tuple of the inertial data. Apart from implementation details, this is equivalent to including a sensor s_3 to the above-mentioned alignment procedure, correspondent to the set of ambient sensors from the smart home. An inertial input to the DL architecture would consist of a sub-sequence of a recording session of length N_t^{raw}. The data from the ambient sensors were introduced to an additional preprocessing step before feeding the DL models: the attributes of the N_t^{raw} correspondent samples on the ambient sensors were averaged, composing a feature vector. Finally, the data from inertial sensors and from the smart home ambient sensors were L2-normalised, in order to compensate the scales of each variable.

All experiments performed were based on classifiers being applied to two-seconds long data segments, regardless of the modality. One example of input, with all the different modalities represented on the HWU-USP dataset, is shown in Figure 7. For the UTD-

MHAD, although data from ambient sensors is absent, the remaining modalities were arranged on the same structure.

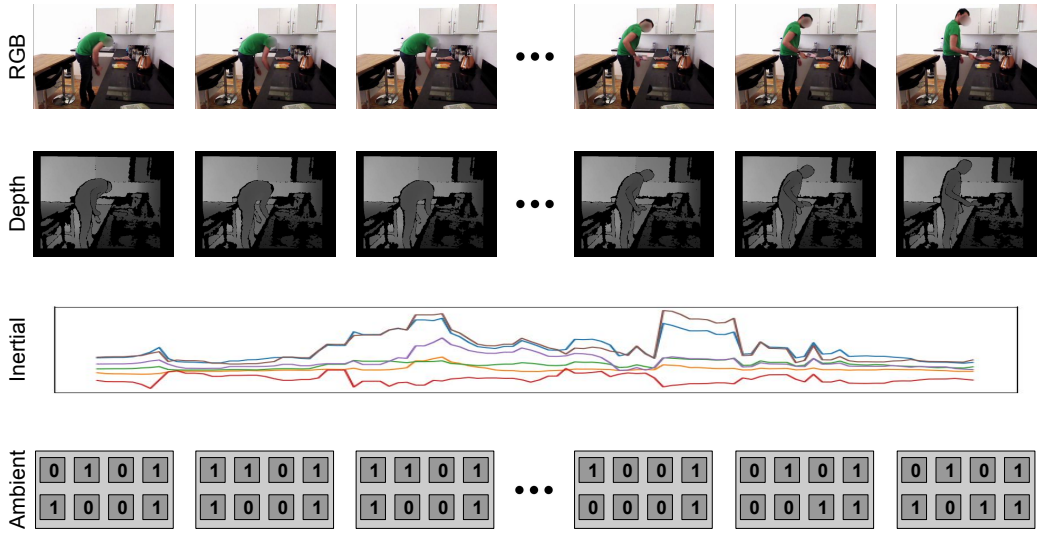


Figure 7. Example of two-seconds long segment fed to the architectures that process each modality.

4.2.3. Deep Learning Architectures

Our multimodal strategy for video and inertial data was based on late fusion of the output vectors of each single-modality model. This approach is also known as decision-level fusion [75]. Indeed, we implemented independent neural networks for videos and IMUs, and, at a later stage, performed weighted averaging of the scores at the outputs of the softmax layers of each network. Data from ambient sensors of the smart home were introduced as an additional input vector on the same neural network used for classifying the IMU data, so that the resulting output vector could also be combined to the video output to provide a classification framework with all modalities considered.

In Table 5, the DL architectures employed for each modality are summarised. We began our analyses by considering two baseline architectures resembling the spatial stream of Simonyan and Zisserman [28]. A consolidated CNN model, the InceptionV3 [102], was employed to train two models for each dataset: one for processing RGB frames, and another for depth frames. We named these modalities *RGB* and *Depth*, respectively. The models were pre-trained on the ImageNet dataset [103], and had all their layers fine-tuned for training the activity recognition datasets under analysis.

Table 5. Summary of the DL architectures employed for each architecture.

Model	Input Description	Structure
Spatial	RGB frame	InceptionV3 [102]
Depth	Grayscale frame	InceptionV3
Optical flow (single frame)	Optical flow pair	InceptionV3 with two input channels
Optical flow (sequence)	Sequence of optical flow pairs	See Figure 8a
IMU	Sequence of raw data	See Figure 8b
Ambient (shallow)	Average vector from sequence of tuples	Fully-connected NN with two hidden layers
IMU+ambient	Two inputs (IMU and ambient)	See Figure 8c

For the video-based temporal stream, which processes optical flow maps, we implemented the neural network of Figure 8a. This consisted of a CNN, which was trained and

evaluated previously for performing the same classification task. Their inputs were a set of flow maps respective to a single pair of frames. This input consisted of two-channels, which comprises the optical flow previously computed. The InceptionV3 architecture was adopted for this aim. We refer to the classification models composed by this CNN alone, without temporal modelling of a sequence of frames, as *optical flow (single frame)*.

To model sequences of optical flow pairs (i.e., the *optical flow (sequence)* condition), the 2048 features, extracted right before the softmax layer of the InceptionV3 architecture, were considered. The CNN was applied N_t^{video} times, each to the optical flow input respective to one timestep of a sequence of timesteps, generating a set of N_t^{video} feature vectors. Those were fed to an LSTM module, whose outputs were the input of a softmax layer for classification. The LSTM module was composed of 128 units and dropout of 50%, with sigmoid activation. L2-normalisation was employed for regularisation.

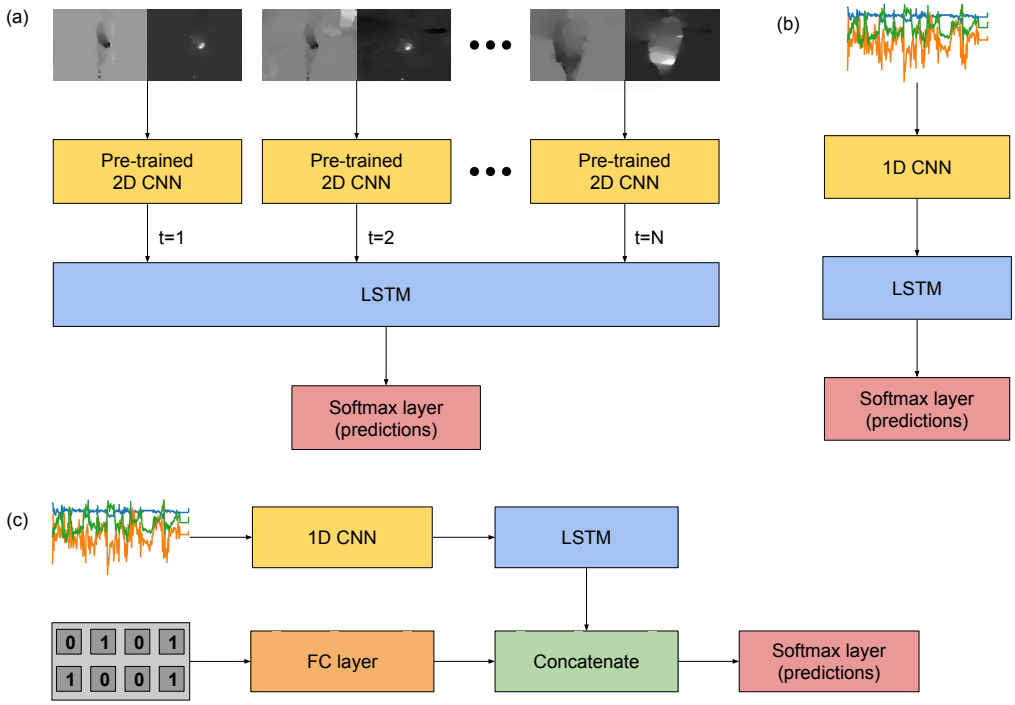


Figure 8. DL Architectures considered for each input modality. (**a**) Video inputs preprocessed by optical flow (i.e., two-channels input maps) algorithms. (**b**) IMU inputs, with a custom one-dimensional CNN as a feature extraction step before feeding a LSTM layer. (**c**) Multimodal scenario with fusion between inertial and ambient sensors within the modules of the neural network.

The same structure was designed for the IMU data, as shown in Figure 8b, with the difference that, instead of an InceptionV3, we used a 1D CNN for feature extraction, which performed convolution and pooling operations on the time domain. Let the length of these sequences be N_t^{raw}. We implemented this CNN with three convolutional layers with kernel size 11 and ReLU activations, interspersed with max-pooling layers of kernel size 2. The convolutional layers were composed by 128, 256 and 378 units, respectively. Batch normalisation was introduced before the first and the last convolutional layers. This convolutional block, referred as 1D CNN in Figure 8b, was followed by an LSTM layer with 128 units, ReLU activation, and dropout of 50%.

Regarding the machine learning aspect, the most noticeable novelty in this work was the introduction of data from ambient sensors of the smart home on the learning framework, which could be done with the new HWU-USP dataset, but not with the UTD-MHAD. An additional input vector, composed by structured data from binary sensors and voltage measurements from the kettle, was added to the same network designed to learn features and classify the IMU data. One condition was included to process this input vector with a shallow neural network: a fully-connected neural network with two hidden layers

composed of 512 and 256 units, respectively, with ReLU activation and dropout of 50% after each layer, and a softmax redout layer.

A feature-level fusion architecture between the IMU and ambient data was also proposed. The approach, as shown in Figure 8c, was to process the input of the ambient sensors in parallel to the convolutional and recurrent layers of the IMU architecture (Figure 8b), this time using a single fully-connected layer with ReLU activation function and dropout of 50%. The outputs of this layer would be concatenated to the features learnt by the convolutional module of the inertial data, and then classified with a softmax layer.

Skeleton joints, which may be extracted by RGBD cameras, were not considered on our framework. When designing our dataset, we were interested in providing a framework based on DL techniques, which have shown to provide good results for video classification on highly unstructured scenarios, closer to real-world applications. However, the proposed architectures could not be employed directly to data from skeleton joints without a feature extraction stage. To properly consider data from skeleton joints for our dataset, we would have to process it with unrelated techniques, which we understand to be out of the scope of this work.

4.3. Experimental Setup

For the implementations of the models presented in the previous subsection, we adopted the TensorFlow library, particularly the Keras module, which provides support for GPU training and evaluation. The models were trained on different hardware devices: the cluster Euler, at the Centre for Mathematical Sciences Applied to Industry (CeMEAI) at ICMC-USP, with GPU nodes provided with a Nvidia Tesla P100; a research computer at the Robots Learning Laboratory (LAR), at ICMC-USP, provided with a Nvidia Titan V GPU; and an ASUS TUF Gaming laptop, provided with a Nvidia Geforce RTX2060 GPU.

All architectures were fed with sequences which correspond to two-seconds-long segments of the recordings. For the case of the video modalities, the inputs were sequences of length $N_t^{\text{video}} = 15$ with period $T = 2$ (i.e., the frames were downsampled on the temporal dimension to half of its original frequency) for the UTD-MHAD dataset, and $N_t^{\text{video}} = 25$ with $T = 1$ for the HWU-USP dataset. For the inertial and ambient modalities, the length of the sequences were $N_t^{\text{raw}} = 100$ for both datasets, as the two of them were converted to a $r = 50$ Hz sampling rate. The optimisation algorithm was Stochastic Gradient Descent (SGD) with learning rate 10^{-2}, momentum 0.9 and decay 10^{-4}. For the video models, training was performed for 40,000 steps, and, for the others, for 20,000 steps.

The evaluation protocol consisted of cross-subject training and testing, with a leave-one-out-approach. That means recordings from one subject were used for testing, while all others were used for training. Consequently, for each input modality, we have trained eight models, and reported the mean and standard deviations of their performance in the test sets. The predictions were obtained using the same principle as recommended in Simonyan and Zisserman [28]. They consisted of evaluating 25 segments on each session recorded, equally spaced between them, and the resulting scores of all outputs were averaged before to produce a prediction. This was done on all modalities. Although this could negatively affect our overall accuracy, this setting is more consistent to real-world applications in which an agent must take actions based on limited, time-localised information. Analyses of the confidence of the predictions through time could be performed, which allowed to better understand the behaviour of the classifiers on the activities of different levels of complexity of the HWU-USP-MHAD dataset, and to compare these results to those obtained with the simpler and shorter activities of the UTD-MHAD.

Fusion of the video streams and the other modalities was performed ad-hoc, after the predictions were already obtained and recorded. The procedure consisted of averaging the outputs of the modalities that were being combined, with different weights for different modalities. The accuracies on different multimodal scenarios were computed on the output vectors respective to this average. For HWU-USP dataset, the weights were set to 1 and 6 for the IMU and video modalities, respectively. For UTD-MHAD, they were set to 1 and

2. On both cases, the weights were chosen in order to maximise the accuracy obtained on each fusion approach.

5. Results

In this section, results of the experiments with respect to each modality are presented, along with multimodal approaches, in Table 6. The *RGB* and *Depth* modalities were computed by feeding an InceptionV3 newtwork with regular 3-channels frames extracted from the videos. In the case of the RGB frames, these consisted of the colour channels of the image, as usual in CNNs. In relation to the depth frames, the 16-bit inputs were converted to 8-bit maps, which were repeated on the three channels, composing grayscale images, already mentioned in Section 4.2.2.

The modality *optical flow (single frame)* refers to a modification of InceptionV3 network to receive a 2-channels input, which was fed with one pair of dense optical flow (see Section 4.2.2), hence considering only one pair of timesteps on the sequence. On the other hand, the *optical flow (sequence)* models refer to LSTM modules processing the features extracted by CNNs for two-seconds long segments of the recordings (see Section 4.2.3 and Figure 8a). The *ambient (shallow)* modality refers to a shallow fully-connected neural network applied directly to the subsequence (see Section 4.2.3), whereas the *IMU* modality refers to the one-dimensional CNN-LSTM models applied to the data from inertial sensors, also computed on two-seconds long subsequences (Figure 8b), the *IMU + ambient* comprises one multimodal setting with both modalities combined within the neural network (Figure 8c). Finally, the *Optical flow + IMU* and *Optical flow + IMU + ambient* multimodal conditions refer to the late fusion approach presented in Section 4.3, consisted of combining the output vectors of each modality before making a final prediction.

It is important to emphasise that all results were computed with predictions from the average output vector from 25 segments on the test data of each modality, and that the train and test partitioning followed a cross-subject approach with eight folds (see Section 4.3), hence the table shows the mean and standard deviation over these eight folds. In Table 6, we presented results on both the HWU-USP and UTD-MHAD datasets, despite the important differences existing between them (see Section 4).

Table 6. Accuracy Measures (%) for each input modalities, for UTD-MHAD and HWU-USP datasets. Models for a single input modality and multimodal models are listed. The accuracy shown is the mean value of 8 cross-subject folds (i.e., leave-one-out cross-subject evaluation protocol), with inputs from a single subject being left for testing, a costly, yet rigorous evaluation protocol.

	UTD-MHAD	**HWU-USP**
RGB (single frame)	6.39 ± 2.16	19.57 ± 6.76
Depth (single frame)	5.46 ± 2.29	36.36 ± 7.28
Optical flow (single frame)	82.47 ± 5.42	86.72 ± 6.74
Optical flow (sequence)	84.79 ± 5.25	93.75 ± 3.33
IMU	82.23 ± 6.55	65.56 ± 13.16
Ambient (shallow)	-	51.39 ± 4.61
IMU + ambient	-	74.30 ± 11.09
Optical flow + IMU	92.33 ± 5.40	96.53 ± 3.87
Optical flow + IMU + ambient	-	98.61 ± 2.41

The confusion matrices for part of the above-mentioned models, for both datasets, were computed in order to allow a more in-depth discussion on the behaviour of each model. For the HWU-USP dataet, these matrices are shown in Figure 9, and, for the UTD-MHAD, in Figure 10.

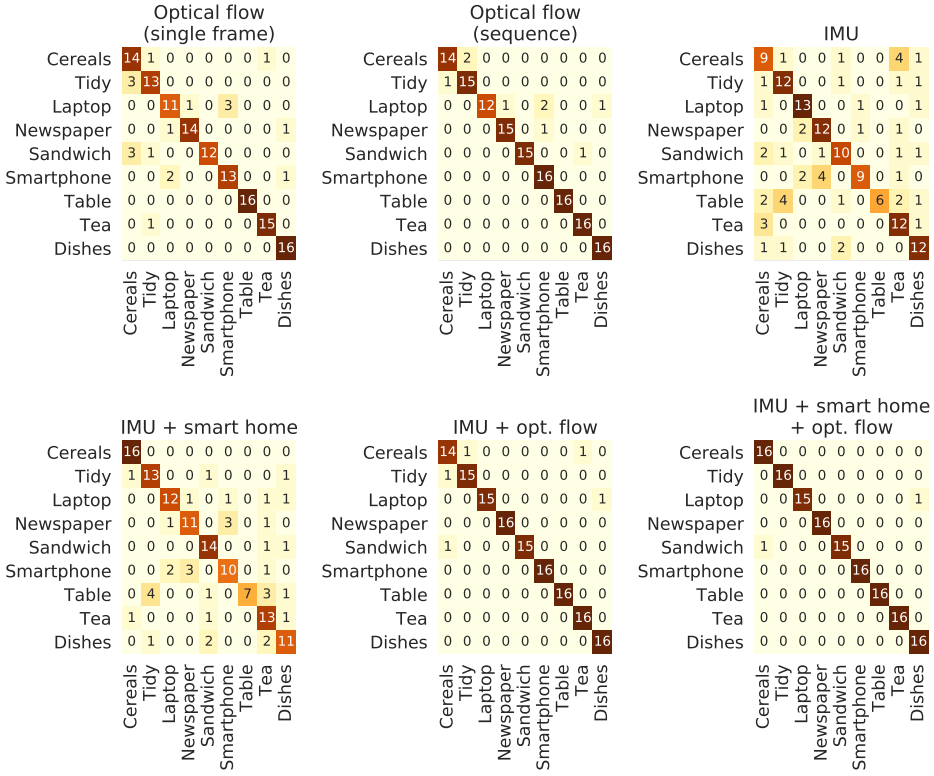


Figure 9. Confusion matrices of different input modalities and architectures for classifying the HWU-USP dataset. The values consist of the summed number of predictions over all folds.

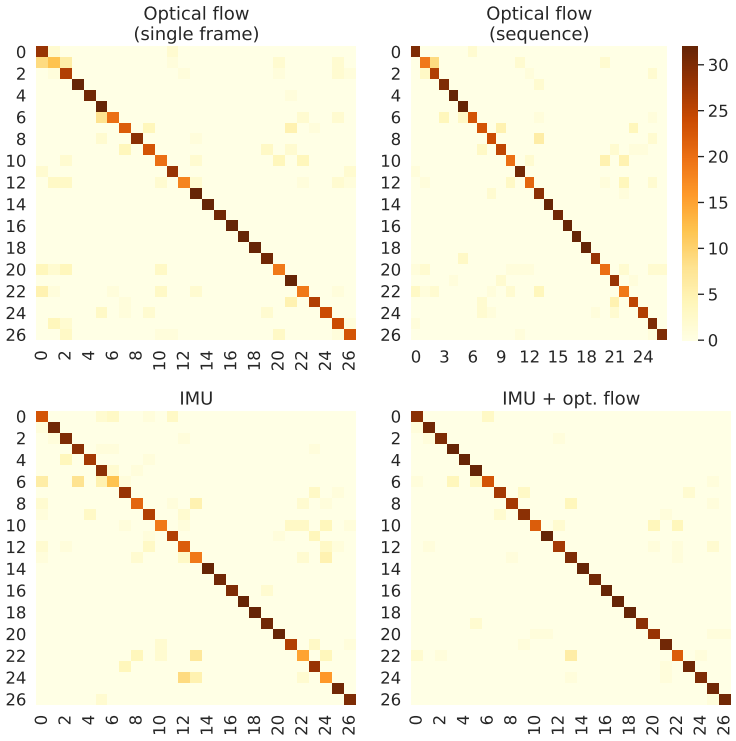


Figure 10. Confusion matrices of different input modalities and architectures for classifying the UTD-MHAD dataset. The values consist of the summed number of predictions over all folds.

We have provided another analysis to evaluate how each model performs across the different segments used to compute the final prediction. These results may lead to important discussions when considering which approach will be adopted for a real-time application, in which partial results computed on a limited range of time might be used in decision-making systems. Figures 11 and 12 present, respectively, for the HWU-USP and UTD-MHAD datasets, the maximum score on the output vectors correspondent to the actual class, across each of the 25 segments used for prediction. For example, taken an input that belongs to the *laptop* class, if the output vector of the first sequence fed to a given classifier gives a 25% confidence for predicting the correctly, and the first sequence of another instance from the same class gives a 32% confidence, the value considered for the figure will be 32%.

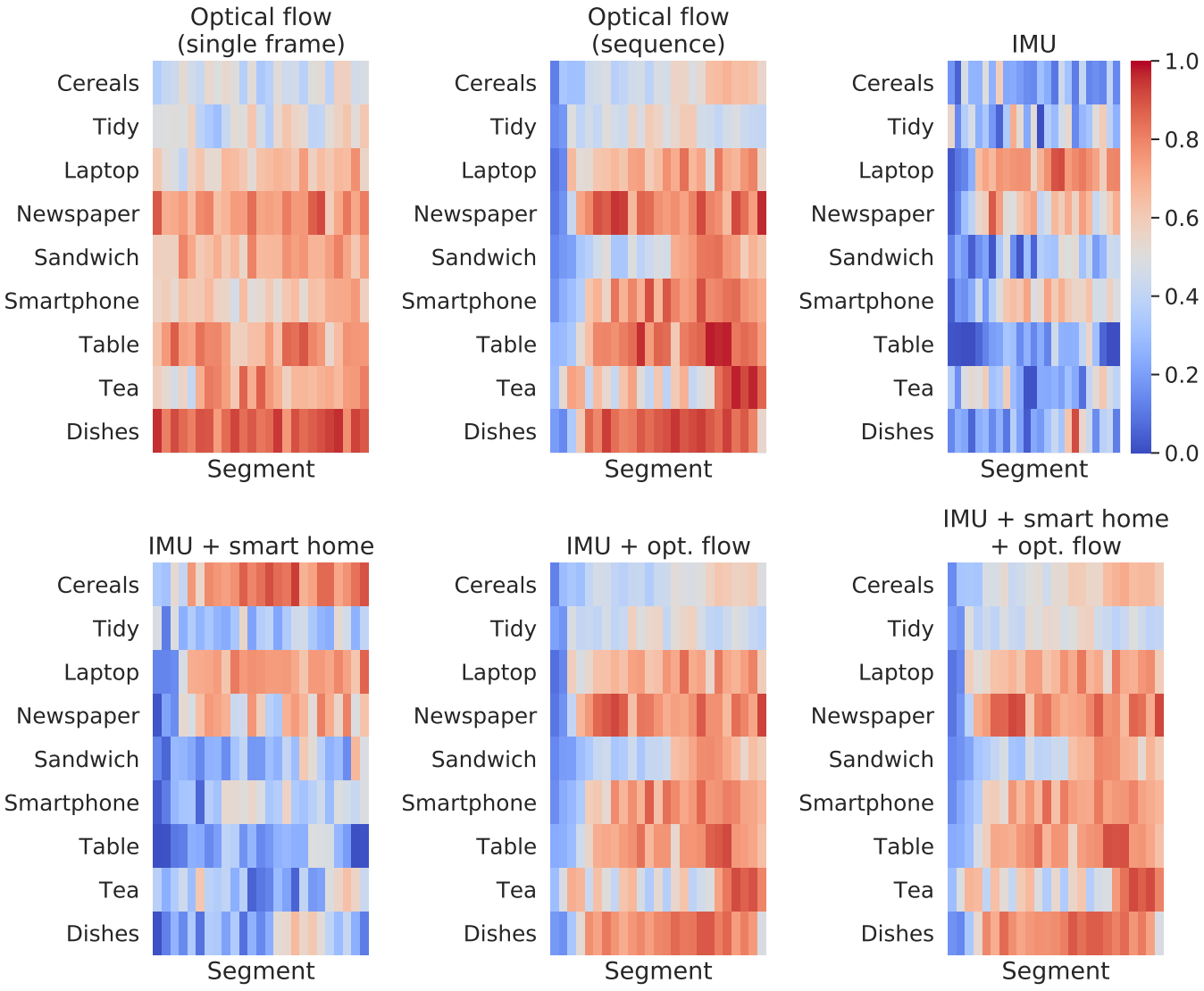


Figure 11. Confidence of predicting the correct label for the HWU-USP dataset, at each of the 25 segments evaluated, equally spaced between them.

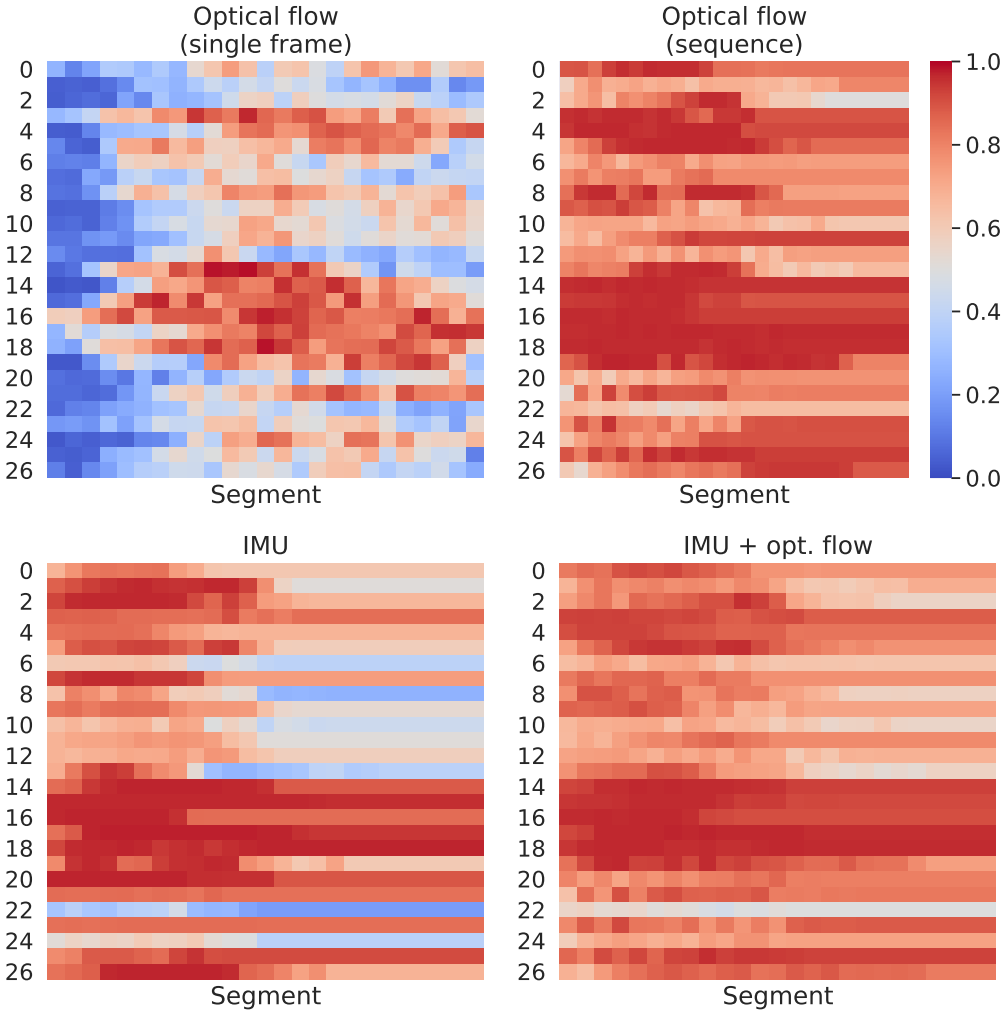


Figure 12. Confidence of predicting the correct label fot the UTD-MHAD, at each of the 25 segments evaluated, equally spaced between them.

Comparison with the State-of-the-Art

HWU-USP database is being presented for the first time in this paper, hence the above-mentioned results are the first to be ever published. For this reason, there is still no literature to compare it with. On the other hand, UTD-MHAD dataset is a widely used benchmark which we can use to evaluate our multimodal approach with only videos and IMU, since this dataset does not provide data from ambient sensors. In Table 7, a collection of results from the literature was put along with the best result that we achieved. To select those studies, we followed the criteria that videos and IMU data were both employed, preferably without skeleton data, so that the comparison with our approach would be as fair as possible. Besides, we only considered studies in which it was explicitly stated that the evaluation protocol was cross-subject. However, it is hold to note, that only Wei et al. [75] adopted a leave-one-out approach, similar to ours, while the others followed the protocol by Chen et al. [37], which used a hold-out approach with half of the data being used for testing.

Table 7. Comparison between our model and others in the literature that deal with similar modalities, for the UTD-MHAD dataset. All of those evaluations adopted a cross-subject protocol.

Method	Modalities	Accuracy (%)
Chen et al. [37]	Depth + IMU	79.10
El Din El Madany et al. [104]	Depth + IMU + skeleton	93.26
Wei et al. [75]	RGB-only	76.00
Wei et al. [75]	Inertial-only	90.30
Wei et al. [75]	RGB + inertial	95.60
Imran and Raman [105]	RGB + inertial	92.32
Ours	RGB + inertial	92.33

6. Discussion

The RGB frames (i.e., the *spatial stream* by Simonyan and Zisserman [28]) led to accuracy measures slightly above random choice for both datasets (i.e., 6.39% for 27 classes on the UTD-MHAD, and 19.57% for 9 classes on the HWU-USP dataset). The depth frames did not led to better results for UTD-MHAD (i.e., 5.46%), but led to an important improvement for HWU-USP dataset (i.e., 36.36%). Still, both approaches led to poor results, if compared to the other models. These results differ from those obtained for video datasets in the literature of multiple stream classification methods [18,28], in which the spatial stream alone led to competitive performances.

Even though, the low accuracy obtained in our experiments was expected due to the nature of the datasets analysed. Applied directly to RGB, or even depth images, a CNN is able to distinguish between the objects, backgrounds and other appearance-based aspects within a scene. Thus, it may be effective when comparing videos from heterogeneous datasets with large inter-class variability regarding those aspects. This may lead to comparatively high accuracy even if motion information was disregarded. The datasets considered in this study present constant background and a limited variability regarding other appearance aspects. Different from the UTD-MHAD, the HWU-USP dataset was recorded from a perspective in which the subjects changed their position constantly with respect to the depth dimension, which may explain the improvements that happened only for this dataset when compared the depth to the spatial models. Nevertheless, for both datasets and any other that shares these characteristics possibly inherent to home environments, a reliable classification method might be based on motion information.

Motion information contained in dense optical flow maps (i.e., the *temporal stream*) led to expressive improvements, even on the *single frame* scenarios. These models were, by far, the ones that led to the highest accuracy on the HWU-USP dataset, which points to the relevance of motion information from computer vision on the scenario analysed. The results were also the bests for the UTD-MHAD, however the IMU condition was still competitive.

When comparing the *single frame* to the *sequence* optical flow architectures, HWU-USP dataset was characterised by a greater increase in accuracy (i.e., 86.72% to 93.75%) than the UTD-MHAD (i.e., 82.47% to 84.79%). This was probably because the HWU-USP dataset is composed of longer recordings with longer-time dependencies. This illustrates how the LSTM-based module is effective in modelling the long-term dependencies that were introduced.

Compared to the video modalities, especially the *optical flow (sequence)*, IMU-based models performed better on the UTD-MHAD, in which the 82.23% accuracy was even competitive when compared to the optical flow models, than in the HWU-USP dataset, which appeared to be favourable for computer-vision approaches. This was probably because the actions on the UTD-MHAD dataset are shorter and more well-defined, so that the most discriminative features were present on most snippets of the inertial data. For the HWU-USP dataset, some of the activities are complex, composed of sequences of actions

that may, isolated, be part of different activities. The visual information contained on videos may be more informative than the IMU data with respect to these more challenging dependencies.

Analyses involving the binary data from the ambient sensors of the smart home could be made only on the HWU-USP dataset, and led to promising results. On its own, feeding a shallow fully-connected network with minimum preprocessing, this modality led to an accuracy of 51.39%, which is expressively below the 65.56% obtained by the one-dimensional CNN-LSTM applied to the IMU data alone. However, when combined, the model hit the accuracy of 74.30%, the best performance obtained without the use of optical flow data.

When combined, the models in which the optical flow models were fused to the other modalities led to the best accuracies. For the UTD-MHAD, this approach led to 92.33%. For the HWU-USP dataset, two conditions were considered: combining the optical flow and the IMU models, as with the UTD-MHAD, and combining the optical flow model to the IMU + ambient model (see neural architecture on Figure 8c). For the first condition, the accuracy was 96.53%, an increment of almost three percent points when compared to the optical flow model on its own. For the second condition, which was possible only because we had made available data from the smart home sensors on the HWU-USP dataset, the accuracy was 98.61%, which may be seen as a remarkable result.

It may be worth discussing some aspects regarding the confusion matrices shown in Figures 9 and 10. Considering the models for the HWU-USP dataset, the most solid observation is that the *cereals* and *tidy* activities are the most sources of wrong predictions on the models with higher accuracy, for either the computer-vision or IMU models. The introduction of the ambient sensors caused an important impact on the recognition of these classes, bringing the error down to zero, which may explain its relevance of the accuracy results of Table 6. Multimodal models provided basically a reduction on the mistakes made in some classes, when compared to the predictions made by each single-modality model. The UTD-MHAD models performed more uniformly across the different modalities, which may explain why the results did not vary too much for the single-modality approaches. For the multimodal scenario, the classes with less precision on each modality seem to have been compensated, causing the observed increment of accuracy.

The confidence scores through time, shown in Figures 11 and 12, seem to have been expressively improved on the UTD-MHAD when comparing the single frame to the sequence approaches. However, these accuracy improvements were more prominent on the HWU-USP dataset, even though the differences of the confidence scores through time seemed to be smaller on these figures. This was because the activities from the UTD-MHAD dataset were all short and made of simple gestures, hence a small snippet on the middle of a video recording could be more informative of the actual activity, providing a correct prediction with high confidence. The same is usually not true for the HWU-USP dataset.

For the UTD-MHAD, the confidence over segments (Figure 12) also presented important differences between the conditions. The *optical flow (sequence)* model provided high confidence scores on the segments closer to the middle of the recordings, in which it differs from the *optical flow (sequence)*, with confidence scores approximately uniform on the whole sequences. For several classes, the IMU model provided high confidence scores only for the first half segments of each modality. The multimodal *IMU + optical flow* provided an improved version of the *optical flow (sequence)* model, except for activity 22, which appears to have its confidence degraded by the IMU scores.

On the other hand, for the HWU-USP activities dataset (Figure 11), a diverse behaviour of the classifiers on each modality was observed. Regarding the optical flow conditions, the *sequence* approach is less uniform across segments than the *single frame*, but seems to provide higher confidence on certain parts of the activities. The activities performed with participants sitting down and performing repetitive movements (i.e., *laptop*, *smartphone* and *newspaper*) led to higher confidence scores for the IMU modality. The combination between IMU and ambient sensors increased drastically the confidence of the *cereals* class across

all segments. The *tea* activity led to high confidence scores on its ending, especially when considering the *optical flow (sequence)* model. An increment on this same region may also be seen when combining the ambient sensors to the IMU, which may be due to the power measurements, which change only during the final moments of the *tea* activity, when the participant turns on the kettle. Few differences may be seen when comparing the *IMU + optical flow* and the *IMU + smart home + optical flow* conditions, however the final segments of the *cereals* activity seems to reflect the most noticeable increment.

By analysing the confusion matrices in Figure 9 and the heat maps in Figure 11, it becomes clear that the predictions on the *cereals* class were the most benefited, which explained the expressive improvements aggregated to the *IMU + ambient* and the *optical flow + IMU + ambient* models with respect to the conditions without this modality. The increment of the confidence on the last segments of the *tea* activity (Figure 11) is also worth a mention, since it was probably due to the power measurements of the kettle, which was turned on at the end of all recording sessions of this activity. In any case, the improvements provided by the multimodal models give additional confirmation on the usefulness of combining videos and IMU modalities whenever they are both available, corroborating to other results from related work [35,75,106].

Comparisons with the state-of-the-art datasets, made only with the UTD-MHAD, pointed that, with respect to videos and inertial sensors, our methods led to results that were compatible. Our best approach led to an accuracy of 92.33%. The reference results, 79.10% for the multimodal condition, provided on the presentation paper of the UTD-MHAD dataset, were successively surpassed on the following years. Models that consider skeleton data led to the higher accuracies, such as in El Din El Madany et al. [104], which hit 93.26%. Nonetheless, these results are not comparable to ours, since this approach considered information extracted from skeleton joints, an additional, rich data input.

Different approaches restricted to the IMU and the video data have also been proposed. Imran and Raman [105] performed experiments with different sets of modalities, and hit 92.32% using RGB videos and inertial sensors, a result which is very close to ours. The most successful approach restricted to these modalities, nonetheless, was provided by Wei et al. [75], which hit 95.60%. This was the only result on the literature that surpassed ours without the use of skeleton information. We have included the RGB-only and inertial-only results in order to situate their results with respect to ours. It is important to note that the video-based method presented by them actually performed less accurate than ours (76.0%, against 84.8% of our approach), which means that the overall accuracy on their work benefited especially from the inertial-only model, which hit 90.3%, against 82.2% of ours.

However, the IMU architecture was based on a two-dimensional matrix representation of the input data, which required the whole sequence to provide a representation. This would not allow a segment-wise classification such as the approach proposed by us, in which the neural network processed two-seconds long data segments, and therefore its application would not be possible in scenarios with partial data, such as the real-time decision-making systems that would be expected on AAL environments.

7. Conclusions

In this paper, we presented the HWU-USP activities dataset, collected at the RALT lab in Edinburgh at Heriot-Watt University. More specifically, the dataset was composed of RGB and depth videos from the camera of a TIAGo robot, data from IMU sensors attached to the users' wrist and waist, and a set of ambient sensors (i.e., switches at the doors of wardrobes and drawers, motion sensors and power measurements) from a smart home. The objective was to build and study a multimodal dataset composed of RGB and depth videos, inertial and ambient sensors from a smart home in the context of activities of daily living, all of them sharing a kitchen environment and performed in the context of a regular breakfast. A set of 16 participants performed 9 activities, resulting in a total of 144 instances

that composed 116 min of recordings in total. All data were stored, made anonymous and will be available to the research community.

This dataset allowed the proposal of multimodal approaches involving not only videos and data from inertial sensors, but also ambient sensors. To the best of our knowledge, this is the first public multimodal activities dataset that provides these three modalities altogether and synchronously. We also proposed a deep learning framework to perform experiments on a multimodal approach. It is based on two-dimensional CNN modules for feature extraction on RGB frames, depth images and optical flow pairs, and LSTM layers for temporal modelling, when applicable. Data from inertial sensors were fed to a similar architecture, with a one-dimensional CNN being applied to extract features to be modelled by a LSTM module. For these modalities, we performed the same experiments on both the HWU-USP and the UTD-MHAD datasets. Results varied from one modality to another, especially for the HWU-USP, in which the architectures based on computer vision, specifically after computing dense optical flow, performed significantly better. These differences were smaller for the UTD-MHAD dataset.

The data from the ambient sensors, present only on the new HWU-USP, were introduced as an additional channel of information on the neural network that processed the inertial data, with no feature extraction: the binary variables were fed to a fully-connected layer whose output was concatenated to the IMU features extracted by the CNN-LSTM modules. The presentation of this fusion architecture is another contribution of our work As expected, the introduction of this modality led to expressive improvements in accuracy. The best multimodal model led to a very high accuracy, which points to the relevance of considering different sources of data to perform activity recognition tasks.

Future work will apply the models trained with this dataset to experiments in the smart home, allowing interventions to be made based on the predictions provided. This may be promising for application scenarios involving human–robot interaction (HRI). For example, a robot may use the successive predictions of an activity recognition framework to decide whether it might remember an user to take his medicines when he is having a meal, or bring an used glass from the living room to the kitchen when the user is washing the dishes. This may be important for designing Ambient Assisted Living solutions with automated technologies, such as robot carers, for monitoring the inhabitants of a smart environment. Moreover, fine-grained annotations may be provided for training models that suit most of those application scenarios and in accordance with the necessities that may arise during those experiments.

To react proactively to the users' needs, this type of applications requires a framework able to provide reliable predictions in real-time, before the user finishes his current activity. This requirement was addressed by our approach, which relied on two-seconds-long segments, whose predictions may be combined to provide better results. In this sense, another direction for future research is to analyse how these predictions may be employed in real-world scenarios in order to complete the most adequate proactive behaviours in a timely manner. We provided an analysis of the confidence of the predictions across segments, which may give a hint on the performance of the framework in real-time applications. For other applications, in which those requirements are absent, experiments with longer segment lengths may be designed, which may foster research on novel learning architectures.

Although the classification methods provided excellent results when the video modality was present, there is still room for improvements regarding the other modalities. Such developments are important because, for real AAL environments, the video data may be frequently unavailable. This may be due to privacy issues or technical limitations, for example if the videos can only be registered by the camera of a social robot, which may not always be accompanying all the inhabitants of the environment. Yet, the very high accuracies provided by the video methods may serve to provide labels on a semi-supervised scenario for new data collections, which may rely on more cameras and more visual perspectives.

Author Contributions: C.M.R. and R.A.F.R. performed literature review and contextualisation; C.M.R. and P.A.V. designed the dataset and managed to obtain approval from the ethics committee; C.M.R., S.M. and M.D. prepared the environment and performed the data collection sessions; C.M.R. designed the machine learning framework, performed the experiments and analysed the results; C.M.R. and S.M. wrote the paper; P.A.V., M.D. and R.A.F.R. performed the revisions. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Sao Paulo Research Foundation, grants 2017/02377-5, 2018/25902-0 and 2017/01687-0, and METRICS (H2020-ICT-2019-2-#871252). It was carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Nvidia Grants program.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the Heriot-Watt University (17 November 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in the Dryad Digital Repository, at https://doi.org/10.5061/dryad.v6wwpzgsj.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. World Population Prospects 2019—Population Division—United Nations. 2019. Available online: https://www.un.org/development/desa/publications/world-population-prospects-2019-highlights.html (accessed on 1 December 2020).
2. Calvaresi, D.; Cesarini, D.; Sernani, P.; Marinoni, M.; Dragoni, A.F.; Sturm, A. Exploring the ambient assisted living domain: A systematic review. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *8*, 239–257. [CrossRef]
3. Maskeliunas, R.; Damaševicius, R.; Segal, S. A review of internet of things technologies for ambient assisted living environments. *Future Internet* **2019**, *11*, 259. [CrossRef]
4. Amato, G.; Bacciu, D.; Chessa, S.; Dragone, M.; Gallicchio, C.; Gennaro, C.; Lozano, H.; Micheli, A.; Hare, G.M.P.O.; Renteria, A.; et al. A Benchmark Dataset for Human Activity Recognition and Ambient Assisted Living. In *International Symposium on Ambient Intelligence*; Springer: Cham, Switzerland, 2016; pp. 1–9.
5. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors* **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
6. Domb, M. Smart home systems based on internet of things. In *Internet of Things (IoT) for Automated and Smart Applications*; IntechOpen: London, UK, 2019.
7. Hasenauer, R.; Belviso, C.; Ehrenmueller, I. New efficiency: Introducing social assistive robots in social eldercare organizations. In Proceedings of the 2019 IEEE International Symposium on Innovation and Entrepreneurship, TEMS-ISIE 2019, Hangzhou, China, 23–25 October 2019; Institute of Electrical and Electronics Engineers Inc. (IEEE): New York, NY, USA, 2019.
8. Yao, Y.; Plested, J.; Gedeon, T. Deep feature learning and visualization for EEG recording using autoencoders. In *Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science (LNCS, volume 11307)*; Cheng, L., Leung, A., Ozawa, S., Eds.; Springer: Cham, Switzerland, 2018; pp. 554–566.
9. Fernandes Junior, F.E.; Yang, G.; Do, H.M.; Sheng, W. Detection of Privacy-sensitive Situations for Social Robots in Smart Homes. In Proceedings of the 2016 IEEE International Conference on Automation Science and Engineering (CASE), Fort Worth, TX, USA, 21–24 August 2016; pp. 727–732. [CrossRef]
10. Jobanputra, C.; Bavishi, J.; Doshi, N. Human activity recognition: A survey. *Procedia Comput. Sci.* **2019**, *155*, 698–703. [CrossRef]
11. Chaaraoui, A.A.; Climent-Pérez, P.; Flórez-Revuelta, F. A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living. *Expert Syst. Appl.* **2012**, *39*, 10873–10888. [CrossRef]
12. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87. [CrossRef]
13. Ahmed, A.; Jalal, A.; Kim, K. RGB-D images for object segmentation, localization and recognition in indoor scenes using feature descriptor and Hough voting. In Proceedings of the 2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 14–18 January 2020; pp. 290–295.
14. Sousa Lima, W.; Souto, E.; El-Khatib, K.; Jalali, R.; Gama, J. Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview. *Sensors* **2019**, *19*, 3213. [CrossRef]

15. Guo, J.; Li, Y.; Hou, M.; Han, S.; Ren, J. Recognition of Daily Activities of Two Residents in a Smart Home Based on Time Clustering. *Sensors* **2020**, *20*, 1457. [CrossRef]
16. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv* **2012**, arXiv:1212.0402.
17. Reyes-Ortiz, J.L.; Oneto, L.; Samà, A.; Parra, X.; Anguita, D. Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing* **2016**, *171*, 754–767. [CrossRef]
18. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*; Springer: Amsterdam, The Netherlands, 2016; pp. 20–36. [CrossRef]
19. Wang, P.; Li, W.; Gao, Z.; Zhang, Y.; Tang, C.; Ogunbona, P. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
20. Gumaei, A.; Hassan, M.M.; Alelaiwi, A.; Alsalman, H. A Hybrid Deep Learning Model for Human Activity Recognition Using Multimodal Body Sensing Data. *IEEE Access* **2019**, *7*, 99152–99160. [CrossRef]
21. Du, Y.; Lim, Y.; Tan, Y. A Novel Human Activity Recognition and Prediction in Smart Home Based on Interaction. *Sensors* **2019**, *19*, 4474. [CrossRef] [PubMed]
22. Bacciu, D.; Di Rocco, M.; Dragone, M.; Gallicchio, C.; Micheli, A.; Saffiotti, A. An ambient intelligence approach for learning in smart robotic environments. *Comput. Intell.* **2019**, *35*, 1060–1087. [CrossRef]
23. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
24. Guesgen, H.W. Using Rough Sets to Improve Activity Recognition Based on Sensor Data. *Sensors* **2020**, *20*, 1779. [CrossRef] [PubMed]
25. ud din Tahir, S.B.; Jalal, A.; Batool, M. Wearable Sensors for Activity Analysis using SMO-based Random Forest over Smart home and Sports Datasets. In Proceedings of the 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020; pp. 1–6.
26. Hall, D.L.; Llinas, J. An introduction to multisensor data fusion. *Proc. IEEE* **1997**, *85*, 6–23. [CrossRef]
27. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425. [CrossRef]
28. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
29. Farnebäck, G. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Scandinavian Conference on Image Analysis (SCIA)*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 363–370. [CrossRef]
30. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3024, pp. 25–36. [CrossRef]
31. Zach, C.; Pock, T.; Bischof, H. A duality based approach for real-time TV-L 1 optical flow. In *Joint Pattern Recognition Symposium*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 214–223. [CrossRef]
32. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
33. Ordóñez, F.; Roggen, D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [CrossRef]
34. Garcia, F.A.; Ranieri, C.M.; Romero, R.A.F. Temporal approaches for human activity recognition using inertial sensors. In Proceedings of the 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), Rio Grande, Brazil, 22–26 October 2019; pp. 121–125. [CrossRef]
35. Song, S.; Chandrasekhar, V.; Mandal, B.; Li, L.; Lim, J.H.; Babu, G.S.; San, P.P.; Cheung, N.M. Multimodal multi-stream deep learning for egocentric activity recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 378–385. [CrossRef]
36. Ranieri, C.M.; Vargas, P.A.; Romero, R.A.F. Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020.
37. Chen, C.; Jafari, R.; Kehtarnavaz, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International conference on image processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172. [CrossRef]
38. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2556–2563. [CrossRef]
39. Jiang, Y.G.; Ye, G.; Chang, S.F.; Ellis, D.; Loui, A.C. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval—ICMR '11, New York, NY, USA, 8–11 June 2011; pp. 1–8. [CrossRef]
40. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 22–24 June 2009; pp. 2929–2936. [CrossRef]

41. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.-F. Large-scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1725–1732. [CrossRef]
42. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A Short Note on the Kinetics-700 Human Action Dataset. *arXiv* **2019**, arXiv:1907.06987.
43. Idrees, H.; Zamir, A.R.; Jiang, Y.G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.* **2017**, *155*, 1–23. [CrossRef]
44. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 961–970.
45. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
46. Shahroudy, A.; Ng, T.T.; Gong, Y.; Wang, G. Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1045–1058. [CrossRef] [PubMed]
47. Yu, G.; Liu, Z.; Yuan, J. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2015; Volume 9007, pp. 50–65. [CrossRef]
48. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–24 June 2012; pp. 1290–1297. [CrossRef]
49. Oreifej, O.; Liu, Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013, pp. 716–723. [CrossRef]
50. Ni, B.; Wang, G.; Moulin, P. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1147–1153. [CrossRef]
51. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.Y.; Kot Chichung, A. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]
52. Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S.T.; Tröster, G.; Millán, J.d.R.; Roggen, D. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognit. Lett.* **2013**, *34*, 2033–2042. [CrossRef]
53. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012; pp. 108–109. [CrossRef]
54. Baños, O.; Damas, M.; Pomares, H.; Rojas, I.; Tóth, M.A.; Amft, O. A benchmark dataset to evaluate sensor displacement in activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing—UbiComp '12, New York, NY, USA, 5–8 September 2012; pp. 1026–1035. [CrossRef]
55. Zappi, P.; Lombriser, C.; Stiefmeier, T.; Farella, E.; Roggen, D.; Benini, L.; Tröster, G. Activity Recognition from On-Body Sensors: Accuracy-Power Trade-Off by Dynamic Sensor Selection. In *European Conference on Wireless Sensor Networks*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 17–33. [CrossRef]
56. Bächlin, M.; Roggen, D.; Tröster, G.; Plotnik, M.; Inbar, N.; Meidan, I.; Herman, T.; Brozgol, M.; Shaviv, E.; Giladi, N.; et al. Potentials of enhanced context awareness in wearable assistants for Parkinson's disease patients with the freezing of gait syndrome. In Proceedings of the 2009 International Symposium on Wearable Computers, Linz, Austria, 4–7 September 2009; pp. 123–130. [CrossRef]
57. Bakar, U.A.; Ghayvat, H.; Hasanm, S.F.; Mukhopadhyay, S.C. Activity and anomaly detection in smart home: A survey. In *Next Generation Sensors and Systems*; Springer International Publishing: New York, NY, USA, 2016; Volume 16, pp. 191–220. [CrossRef]
58. Cook, D.J.; Crandall, A.S.; Thomas, B.L.; Krishnan, N.C. CASAS: A Smart Home in a Box. *Computer* **2013**, *46*, 62–69. [CrossRef] [PubMed]
59. Lesani, F.S.; Fotouhi Ghazvini, F.; Amirkhani, H. Smart home resident identification based on behavioral patterns using ambient sensors. *Pers. Ubiquitous Comput.* **2019**, 1–12. [CrossRef]
60. De la Torre Frade, F.; Hodgins, J.K.; Bargteil, A.W.; Artal, X.M.; Macey, J.C.; Castells, A.C.I.; Beltran, J. *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*; Tech. Rep. CMU-RI-TR-08-22; Robotics Institute: Pittsburgh, PA, USA; Carnegie Mellon University: Pittsburgh, PA, USA, 2008.
61. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, USA, 15–17 January 2013; pp. 53–60. [CrossRef]
62. Wei, H.; Chopada, P.; Kehtarnavaz, N. C-MHAD: Continuous Multimodal Human Action Dataset of Simultaneous Video and Inertial Sensing. *Sensors* **2020**, *20*, 2905. [CrossRef]
63. Stein, S.; Mckenna, S.J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), Zurich, Switzerland, 8–12 September 2013.
64. Lin, H.C.; Shafran, I.; Yuh, D.; Hager, G.D. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Taylor Fr.* **2010**, *11*, 220–230. [CrossRef]

65. Ruffieux, S.; Lalanne, D.; Mugellini, E. ChAirGest—A Challenge for Multimodal Mid-Air Gesture Recognition for Close HCI. In Proceedings of the 15th ACM on International conference on multimodal interaction—ICMI '13, New York, NY, USA, 9–13 December 2013; pp. 483–488. [CrossRef]
66. Kepski, M.; Kwolek, B. Fall Detection on Embedded Platform Using Kinect and Wireless Accelerometer. *Comput. Help. People Spec. Needs* **2012**, 407–414. [CrossRef]
67. Gasparrini, S.; Cippitelli, E.; Gambi, E.; Spinsante, S.; Wåhslén, J.; Orhan, I.; Lindh, T. Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In *International Conference on ICT Innovations*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 99–108.
68. Rafferty, J.; Nugent, C.D.; Liu, J.; Chen, L. From Activity Recognition to Intention Recognition for Assisted Living Within Smart Homes. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 368–379. [CrossRef]
69. Luo, G.; Wei, J.; Hu, W.; Maybank, S.J. Tangent Fisher vector on matrix manifolds for action recognition. *IEEE Trans. Image Process.* **2019**, *29*, 3052–3064. [CrossRef] [PubMed]
70. Arif, S.; Wang, J.; Fei, Z.; Hussain, F. Video Representation via Fusion of Static and Motion Features Applied to Human Activity Recognition. *KSII Trans. Internet Inf. Syst.* **2019**, *13*, 3599–3619. doi:10.3837/tiis.2019.07.015
71. Nadeem, A.; Jalal, A.; Kim, K. Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural Network. In Proceedings of the 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020; pp. 1–6.
72. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors* **2019**, *19*, 1005. [CrossRef]
73. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [CrossRef]
74. Anjali, C.; Beena, M.V. Human Activity Recognition using Convolutional 3D Network. *Int. J. Res. Eng. Sci. Manag.* **2019**, *2*, 832–836.
75. Wei, H.; Jafari, R.; Kehtarnavaz, N. Fusion of Video and Inertial Sensing for Deep Learning–Based Human Action Recognition. *Sensors* **2019**, *19*, 3680. [CrossRef]
76. Jalal, A.; Kim, J.T.; Kim, T.S. Development of a life logging system via depth imaging-based human activity recognition for smart homes. In Proceedings of the International Symposium on Sustainable Healthy Buildings, Seoul, Korea, 27–28 February 2012; Volume 19.
77. Jalal, A.; Kamal, S.; Kim, D. Shape and motion features approach for activity tracking and recognition from kinect video camera. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 25–27 March 2015; pp. 445–450.
78. Kamal, S.; Jalal, A.; Kim, D. Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. *J. Electr. Eng. Technol.* **2016**, *11*, 1857–1862. [CrossRef]
79. Jalal, A.; Kamal, S.; Kim, D. Depth Silhouettes Context: A new robust feature for human tracking and activity recognition based on embedded HMMs. In Proceedings of the 2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Goyang City, Korea, 28–30 October 2015; pp. 294–299.
80. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [CrossRef]
81. Kim, K.; Jalal, A.; Mahmood, M. Vision-based Human Activity recognition system using depth silhouettes: A Smart home system for monitoring the residents. *J. Electr. Eng. Technol.* **2019**, *14*, 2567–2573. [CrossRef]
82. Farooq, A.; Jalal, A.; Kamal, S. Dense RGB-D Map-Based Human Tracking and Activity Recognition using Skin Joints Features and Self-Organizing Map. *KSII Trans. Internet Inf. Syst.* **2015**, *9*. [CrossRef]
83. Franco, A.; Magnani, A.; Maio, D. A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognit. Lett.* **2020**, *131*, 293–299. [CrossRef]
84. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P.O. Action Recognition from Depth Maps Using Deep Convolutional Neural Networks. *IEEE Trans. Hum. Mach. Syst.* **2016**, *46*, 498–509. [CrossRef]
85. Jaimez, M.; Souiai, M.; Gonzalez-Jimenez, J.; Cremers, D. A Primal-Dual Framework for Real-Time Dense RGB-D Scene Flow. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 98–104. [CrossRef]
86. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]
87. San-Segundo, R.; Blunck, H.; Moreno-Pimentel, J.; Stisen, A.; Gil-Martín, M. Robust Human Activity Recognition using smartwatches and smartphones. *Eng. Appl. Artif. Intell.* **2018**, *72*, 190–202. [CrossRef]
88. Quaid, M.A.K.; Jalal, A. Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. *Multimed. Tools Appl.* **2020**, *79*, 6061–6083. [CrossRef]
89. Li, H.; Derrode, S.; Pieczynski, W. An adaptive and on-line IMU-based locomotion activity classification method using a triplet Markov model. *Neurocomputing* **2019**, *362*, 94–105. [CrossRef]

90. Rueda, F.M.; Fink, G.A. Learning attribute representation for human activity recognition. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 523–528. [CrossRef]
91. Chen, L.; Nugent, C.D.; Wang, H. A Knowledge-Driven Approach to Activity Recognition in Smart Homes. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 961–974. [CrossRef]
92. Liciotti, D.; Bernardini, M.; Romeo, L.; Frontoni, E. A Sequential Deep Learning Application for Recognising Human Activities in Smart Homes. *Neurocomputing* **2019**, *396*, 501–513. [CrossRef]
93. Medina-Quero, J.; Zhang, S.; Nugent, C.; Espinilla, M. Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. *Expert Syst. Appl.* **2018**, *114*, 441–453. [CrossRef]
94. Gochoo, M.; Tan, T.; Liu, S.; Jean, F.; Alnajjar, F.S.; Huang, S. Unobtrusive Activity Recognition of Elderly People Living Alone Using Anonymous Binary Sensors and DCNN. *IEEE J. Biomed. Health Inf.* **2019**, *23*, 693–702. [CrossRef] [PubMed]
95. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941. [CrossRef]
96. Yang, X.; Ramesh, P.; Chitta, R.; Madhvanath, S.; Bernal, E.A.; Luo, J. Deep Multimodal Representation Learning from Temporal Data. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5066–5074. [CrossRef]
97. Robotic Assisted Living Testbed. Available online: https://ralt.hw.ac.uk/ (accessed on 23 July 2020)
98. PAL Robotics. TIAGo Handbook Version 1.7.1. 2017. Available online: www.pal-robotics.com (accessed on 1 December 2020)
99. Astra Series—Orbbec. Available online: https://orbbec3d.com/product-astra-pro/ (accessed on 1 December 2020)
100. MetaMotionR—MbientLab. Available online: https://mbientlab.com/metamotionr/ (accessed on 1 December 2020)
101. Dragone, M.; Saunders, J.; Dautenhahn, K. On the Integration of Adaptive and Interactive Robotic Smart Spaces. *Paladyn J. Behav. Robot.* **2015**. *6*, 165–179. [CrossRef]
102. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 19–25 June 2016; pp. 2818–2826.
103. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
104. El Madany, N.E.D.; He, Y.; Guan, L. Human action recognition via multiview discriminative analysis of canonical correlations. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 4170–4174. [CrossRef]
105. Imran, J.; Raman, B. Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 189–208. [CrossRef]
106. Lu, Y.; Velipasalar, S. Autonomous Human Activity Classification from Wearable Multi-Modal Sensors. *IEEE Sens. J.* **2019**, *19*, 11403–11412. [CrossRef]

CHAPTER

4

APPROACHES ON COMPUTATIONAL NEUROSCIENCE

Two papers related to developments on computational neuroscience, which influenced the application scenarios presented later on, were reproduced in this chapter. The first paper, named ”Unveiling Parkinson’s Disease Features from a Primate Model with Deep Neural Networks” (RANIERI *et al.*, 2020), was published to the 2020 International Joint Conference on Neural Networks (IJCNN), the IEEE conference on Neural Networks, with H5-index of 45 (Google Scholar) and qualified as A1 in the latest Qualis CC. It consisted of, based on the marmosets dataset, devising and testing a machine learning framework comprised of deep neural networks to identify features in neural dynamics related to healthy and PD states that could contribute to early-stage diagnosis, and to inform novel computational models, as well as generate testable hypotheses on the mechanisms of the disease.

The second paper, a preprint at arXiv, is named ”A data-driven biophysical computational model of Parkinson’s Disease based on marmoset monkeys” (RANIERI *et al.*, 2021). It consisted of deriving a computational model of Parkinson’s Disease by calibrating a previously developed model of the basal ganglia-thalamus-cortex circuit (KUMARAVELU; BROCKER; GRILL, 2016) to fit data from marmoset monkeys according to a data-driven approach based on differential evolution.

These papers correspond to the developments of the Neuro4PD project that counted with most participation of the author of this thesis. This collaboration began to contribute during the scholarship period at Heriot-Watt University, Edinburgh, Scotland, under supervision of professor Patrícia Amâncio Vargas, and continued after this was finished. Based on a recent marmoset monkey database (SANTANA *et al.*, 2014), the Neuro4PD project alleviates gap on the lack of early diagnosis of Parkinson’s Disease (PD), by implementing machine learning (ML) methods to help to unveil its neural correlates. Based on the knowledge to be unveiled about the neural structures of PD, the project aims to reproduce the underlying behaviour in

humanoid robots. Such idea follows the paradigm of neurorobotics, the intersection of robotics and neuroscience with focus on implementing neurobiological structures underlying animal behaviour in robots.

Contribution statement: Ranieri performed literature review and designed the machine learning architectures and experiments. Moioli, Vargas and Romero provided supervision and directions for analyses of the results, implemented and compiled by Ranieri. Moioli, Araújo and Pimentel provided guidance regarding the neuroscience aspects of the analyses. Santana and Araújo provided and curated the data used. Ranieri wrote the paper and performed the analyses. Moioli, Vargas and Romero provided the revisions.

Unveiling Parkinson's Disease Features from a Primate Model with Deep Neural Networks

1st Caetano M. Ranieri
ICMC, *USP*
São Carlos, SP, Brazil
cmranieri@usp.br

2nd Renan C. Moioli
IMD, *UFRN*; and
Santos Dumont Institute (ISD)
Natal, RN, Brazil
renan.moioli@imd.ufrn.br

3th Roseli A. F. Romero
ICMC, *USP*
São Carlos, SP, Brazil
rafrance@icmc.usp.br

4th Mariana F. P. de Araújo
CCS, *UFES*
Vitória, ES, Brazil
Santos Dumont Institute (ISD)
Natal, RN, Brazil
mfparaujo@gmail.com

5th Maxwell Barbosa De Santana
ICTA, *UFOPA*
Santarém, PA, Brazil
barbosadesantana@gmail.com

6th Jhielson M. Pimentel
Edinburgh Centre for Robotics, *HWU*
Edinburgh, Scotland, UK
jm210@hw.ac.uk

7rd Patricia A. Vargas
Edinburgh Centre for Robotics, *HWU*
Edinburgh, Scotland, UK
p.a.vargas@hw.ac.uk

***Abstract*—Parkinson's Disease (PD) is a neurodegenerative disorder with increasing prevalence in the world population and is Characterised by motor and cognitive symptoms. Although cortical EEG readings from PD-affected humans have being commonly used to feed different machine learning frameworks, the directly affected areas are concentrated in a group of subcortical nuclei and related areas, the so-called motor loop. As those areas may only be directly accessed through invasive procedures, such as Local Field Potential (LFP) measurements, most data collection must rely on animal models. To the best of our knowledge, no neural networks-based analysis centred on LFP data from the motor loop was reported so far. In this work, we trained and evaluated a set of deep neural networks on a dataset recorded from marmoset monkeys, with LFP readings from healthy and parkinsonian subjects. We analysed each trained neural network with respect to its inputs and representations from intermediate layers. CNN and ConvLSTM classifiers were applied, reaching accuracies up to 99.80%, as well as a CNN-based autoencoder, which has also shown to learn PD-related representations. The results and analysis provided further insights and foster research on the correlates of Parkinson's Disease.**

***Index Terms*—Parkinson's disease, LFP analysis, deep learning, attribution methods, computational neuroscience.**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder and estimates indicate a prevalence between 1 and 2 per 1,000 individuals. Age is the most relevant factor to influence such incidence [1]. The most common symptoms are motor deficits, such as bradykinesia, rigidity, and resting tremors, although cognitive symptoms, especially dementia, may occur in later stages [2]. PD diagnosis is clinical (there is no feasible biomarker), and current treatments provide symptomatic relief, but do not stop, revert, or slow disease progression [3]. In this context, machine learning techniques are being used to characterise the neurophysiological correlates of PD, which can contribute to unveil disease mechanisms as well as non-trivial features that are present on neural data. Ultimately, this may facilitate early diagnoses and support novel therapies.

With few exceptions, human datasets are comprised of non-invasive electroencephalography (EEG) recordings, which capture the neural dynamics from cortex superficial layers. However, the neural circuits directly associated with PD may only be sampled by invasive electrodes, limiting the availability of human studies. Nevertheless, there are consolidated animal models of PD, in which disease symptoms can be elicited by administering neurotoxins [4]. From implanted electrodes, Local Field Potential (LFP) signals are obtained and these have a close relationship with EEG signals [5]. Furthermore, the basic anatomy and structure of the neural circuitry relevant to PD are conserved across most vertebrate species [6], thus supporting the use of such animal models.

In this paper, we designed a set of deep neural networks able to learn explainable features from raw time-domain LFP data with minimum preprocessing. The trained models were evaluated for their ability to classify structured data segments as belonging to healthy or PD animal subjects. Then, we highlighted which properties of the segments contributed most to the networks' classifications. To accomplish that, we used a marmoset monkey database [7] of simultaneous LFP recordings from PD-related brain regions, namely the basal ganglia-thalamus-cortex (BG-T-C) system, known as the *motor loop*). The network architectures include a fully-connected (FC) network, used as a baseline method, a Convolutional Neural Network (CNN), and an hybrid CNN with Long Short-Term Memory (ConvLSTM). We also employ an autoencoder-based unsupervised framework to analyse not only the consistency

This work was funded by the Neuro4PD project-Royal Society and Newton Fund (NAF\R2\180773), and São Paulo Research Foundation (FAPESP), grants 2017/02377-5 and 2018/25902-0. Moioli, Araujo, and Santana acknowledge the support from the brazillian institutions: INCT INCEMAQ of the CNPq/MCTI, FAPERN, CAPES, FINEP, and MEC. This research was carried out using the computational resources from the CeMEAI funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Robotics Lab within the ECR, and by the Nvidia Grants program.

of the lower-dimension representation (i.e., *embedding*) with respect to the two conditions - healthy or PD - but also the suitability of the learnt features in regular classification.

Our deep learning models trained on LFP readings reached accuracy above 99% and the learnt features resemble those that were previously associated to PD. As the acquisition of LFP data depends on inserting electrodes directly inside the subject's brain, the methods proposed are not directly suitable for diagnosis, but rather to provide an additional framework for better understanding of the underlying mechanisms of the disease. To the best of our knowledge, this paper is the first attempt to apply deep neural networks to better understand PD features extracted from simultaneous multi-region LFP.

II. Related Work

Recent research has shown that deep neural networks may be promising machine learning algorithms for studying biological systems. In [8], classification of Alzheimer's disease was performed based on magnetic resonance brain images fed into a CNN. In [9], the intention to perform certain movements was detected from EEG signals, by generating time-frequency maps through wavelet transforms and feeding them to a CNN.

Regarding PD, in [10], EEG data was collected from 15 patients in early stages of PD, ranging from Hoehn and Yahr (H&Y) [11] stages 1 to 2, and 15 healthy subjects with a similar age profile. They applied the autoregressive Burg and the Wavelet Packet Entropy (WPE) methods to characterise the resulting signals in terms of frequency bands and to identify cortical patterns that may be indicative of PD at its early stages, and found significant differences between the affected patients and the control subjects.

Yuvaraj, Acharya, and Hagiwara [12] provided a machine learning (ML) framework to diagnose the disease in an EEG dataset produced by 20 affected patients with H&Y stage ranging from 1 to 3, though most of them were in stages 2 or 3, and a control group of 20 other subjects with no history of mental illness. They extracted the Higher-Order Spectra (HOS), a well-established technique for feature extraction from biomedical data, and introduced a feature ranking method before applying several classical classifiers, obtaining a state-of-the-art diagnosis with Support Vector Machines (SVM).

An important development in the Brain-Computer Interfaces (BCI) domain was the EEGNet [13], based on the application of a compact CNN in diverse motor tasks. Besides providing a classification framework, the authors explored the interpretation of features by analysing filter outputs, convolutional kernel weights, and single-trial relevance. The SyncNet [14] was another CNN-based deep network capable of handling not only EEG, but also LFP signals from public datasets. When processing EEG data, the framework generated visualisations of the spatial patterns recognised by the network filters with heat maps representing learnt amplitude and phase in different bands of the frequency spectrum. Their work did not describe a visualisation approach for features learnt from LFP data.

In [15], different CNN architectures were employed in order to classify public EEG datasets focused on commands for initiating movement. For visualisation, two types of correlation maps were considered: input-feature unit-output, consisted of bandpass-filtering the input signal to each frequency band of interest and checking the outputs of each unit of the network, and input-perturbation network-prediction, based on perturbations on the network inputs. A paradigm derived from research on video classification was proposed in [16], in which the spatially-coherent readings of the EEG electrodes at a given timestep were represented as a regular 2D image, and stacks of such images were interpreted as sequential frames in a video. The data was collected during a working memory experiment, and different architectures were considered for feature extraction and classification, especially neural networks.

Regarding research aimed at PD diagnosis, [17] presented a thirteen-layer CNN that was applied directly to EEG data from 20 PD patients and 20 healthy subjects from similar age groups for classification. The accuracy obtained was lower than that reported in related work with handcrafted features, though direct comparisons are difficult to make due to the lack of standardised datasets. A different technique was presented on [18], which applied Echo State Networks (ESN) to classify data collected from patients with REM-sleep Behaviour Disorder (RBD), a risk factor for PD, and healthy controls, with promising results. Both papers focused on classification, with few considerations regarding the representations learnt.

Research on learning feature representations from brain signals through unsupervised techniques presented two autoencoder architectures to learn short-time features from EEG data from a public dataset [19]. Each trial was represented as a 2D image whose pixel intensities were related to the power of different EEG frequency bands at the spatial location of each particular electrode in the scalp surface, and channel-wise, in which each EEG electrode was treated as a different channel. The embeddings learnt were fed to fully-connected layers to perform classification tasks, leading to state-of-the-art results in the cross-subject experiments. Another autoencoder-based framework was proposed by Wen and Zhang [20], designed to learn representations related to epilepsy with the so-called AE-CDNN model.

The above-mentioned literature focused on learning representations based on EEG signals from humans by applying different sorts of neural networks, with accurate results in comparison to other approaches. PD-related work was also relied on this modality of data, however work on this subject did not provide an in-depth analysis on the interpratability of the features learnt. Also, LFP data has not been a focus of ML efforts in understanding PD, though we have found research addressing this modality for other purposes. This paper attempts to fulfil those gaps by providing a comparative study via a set of deep networks that learned from a PD-related dataset of marmosets' LFP measurements, in both supervised and unsupervised manners.

III. The Motor Loop

The motor loop of the mammals' brain is formed by the *motor cortex* (M1), the *thalamus* (TH), and the *basal ganglia* (BG), the latter composed of a subset of structures: the *striatum*, which itself includes the *putamen* (PUT) and the *caudate nucleus*, the *globus pallidus*, divided into *pars interna* (GPi) and *pars externa* (GPe), the *subthalamic nucleus* (STN), and the *substantia nigra*, divided into *pars compacta* (SNc) and *pars reticulata* (SNr). McGregor and Nelson [21] provided a discussion about the mechanisms of this loop and presented models to describe it. The most useful model to explain the connections affected by PD is the so-called classic model, illustrated in Fig. 1, which highlights the relationships between the projections of neurons from the SNc to the BG structures, mainly striatum, where dopamine is released.

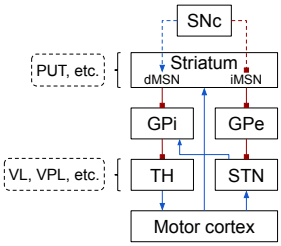


Fig. 1: Excitatory (blue) and inhibitory (red) connections from the circuitry of the motor loop. PD is caused by the loss of neurons of the *substantia nigra pars compacta* (SNc), which weakens the connections represented by the dashed lines. This causes malfunction on both the direct and indirect pathways.

The pathways begin with an excitatory connection from the cortex to the striatum, which projects its output neurons, named *medium spiny neurons* (MSN), to other structures inside the BG. In the direct pathway, the direct MSN (dMSN) inhibits the GPi, which reduces its inhibition to the TH, which then excites the motor cortex. In the indirect pathway, the indirect MSN (iMSN) inhibits the GPe, which reduces its inhibition to the STN, which excites the GPi. Thus resulting on inhibition of the TH and absence of excitatory outputs to the motor cortex. Hence, in summary, the direct pathway excites the cortex (i.e., positive feedback loop), while the indirect pathway inhibits it (i.e., negative feedback loop). PD is characterised by the progressive loss of dopaminergic neurons, especially in the SNc, which causes malfunctions to both pathways.

IV. Methods

This work consists of applying deep neural networks to LFP data collected from marmoset monkeys. We have considered networks for classification, trained to distinguish between healthy and PD-induced individuals, and autoencoders, trained in an unsupervised manner. To explore the representations learnt by each model, we applied attribution methods to segment the input sequences and to look for the most relevant features. All implementations were developed using the TensorFlow/Keras framework. The experiments were performed on a desktop equipped with an Intel Core i7-7700 CPU and a NVidia Titan-V GPU.

A. Datasets

Four adult males and one adult female common marmosets (i.e., *Callithrix jacchus*), weighing 300–550 g, were used in the study performed on [7]. The animals were housed in pairs in a vivarium with a natural light cycle (12/12 hr) and outdoor temperature. All animal procedures followed approved ethics committee protocols (CEUA-AASDAP 08/2011, 11/2011, 02/2015, and 03/2015) strictly in accordance with the NIH Guide for the Care and Use of Laboratory Animals. PD symptoms were elicited in all four male animals with injections of 6-OHDA toxin under deep anesthesia. LFPs were sampled at 1000 Hz and recorded using a 64 multi-channel recording system (Plexon) with fully-awaken animals behaving freely. Electrode coordinates and dopaminergic lesions were verified in all animals.

B. Data Preprocessing

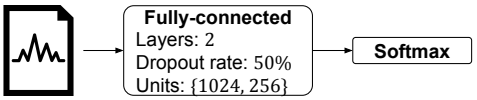
The only healthy individual had recordings from the M1, PUT, GPe, and GPi regions, thus we limited our analysis to those regions. In total, 14 and 16 recording sessions were obtained for the healthy and for the PD conditions, respectively, considering hemispheres independent from each other. Each recording session was segmented in 2-second data segments. As multiple electrodes were recorded for each region, a preprocessing pipeline was required before providing a standardised data structure, as with other approaches in the literature [22]. For each channel, our pipeline began with a low-pass filter (cutoff frequency of 250 Hz), a high-pass filter (cutoff frequency of 0.5 Hz) and a hum notch filter at 60 Hz, 120 Hz, and 180 Hz frequencies. Each signal was then scaled according to a z-score normalisation. The next step was to compute the cross-correlation matrix of each region and discard channels with mean correlation coefficient below the threshold of 0.7. Finally, all channels within a brain region were averaged, which provided a matrix with dimensions 4×2000.

After that, to reduce the amount of noisy or non-meaningful data, we imposed additional criteria to decide whether to keep or discard each resulting instance. An upper threshold of 0.2 was set for the module of the mean of the signal over time at each region, and a lower threshold of 0.1, for the standard deviation. Also, each window was required to show a minimum of 10 peaks.

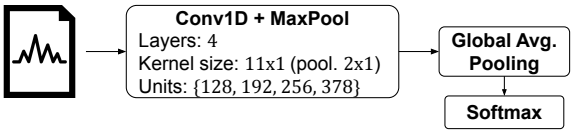
C. Network Architectures

We considered the classification task of distinguishing between healthy and PD-induced individuals and elaborating embedding representations through an autoencoder [23], which could be analysed on its own or coupled with supervised techniques to check its ability to enhance the classification procedure. The different architectures considered are illustrated in Fig. 2. The number of layers and its numbers of neurons were chosen based on literature on EEG classification and

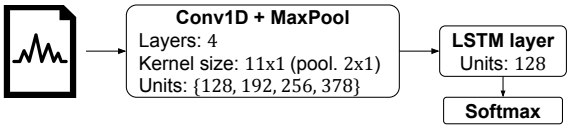
exploratory experiments. The complexity of each model is shown in Fig. 3.



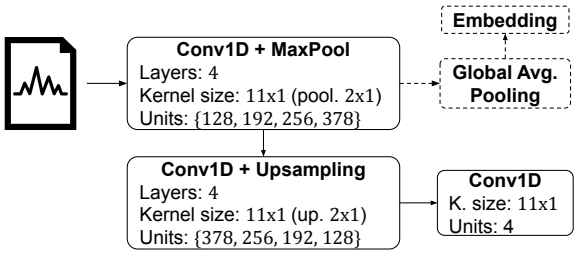
(a) Fully-connected architecture. The two fully-connected layers, both provided with dropout, are followed directly by the softmax layer.



(b) CNN architecture. All convolutional layers within a *Conv1d + MaxPool* block were set to the same kernel size, with one convolutional layer being interspersed with a max-pooling layer. At the top, the features map are processed through global average pooling and fed to a softmax layer.



(c) ConvLSTM architecture. The *Conv1D + MaxPool* block is similar to that of the CNN architecture, however its output is fed to a LSTM layer, whose output is fed to the softmax layer.



(d) Autoencoder architecture. The convolution/upsampling block and the top convolutional layer with the output with the same dimension as the input signal, used for training the autoencoder, is removed and replaced by a global average polling for providing the embeddings.

Fig. 2: Network architectures. The input signals are processed by the intermediate layers, which would be a stack of fully-connected, convolutional, pooling or upsampling, depending on the architecture (icon by Freepic, from www.flaticon.com).

1) Classification Networks: Three different architectures were considered for classification. All of them were endowed with a readout layer made of two neurons, each related to one of the two possible classifications - healthy or PD - and softmax activation function. The baseline, Fig. 2a, was a shallow, fully-connected (FC) neural network, consisted of two intermediate layers with dropout set to 50%. We also considered a 4-layered CNN, Fig. 2b, with the convolutional layers composed of 1-dimensional filters with receptive field

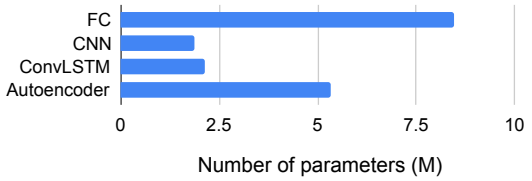


Fig. 3: Complexity of each model, given by the number of parameters, in millions.

of size 11 and interspersed with max-pooling layers with filter size 2, and a ConvLSTM, Fig. 2c, inspired by literature on activity recognition from inertial sensors [24], which consisted of the CNN architecture provided with an additional LSTM layer at the top, right before the softmax layer. The number of units at each layer is depicted in Fig. IV-C1.

2) Autoencoder: The autoencoder, Fig. 2d, reproduced the CNN architecture and endowed it with a reconstruction block. At training time, four convolutional-upsampling pairs were introduced to reverse the encoding produced, followed by a convolutional readout layer to reconstruct the input shape. At test time, the 378×125 encoding following the last max-pooling layer would be processed by global average pooling and turned into a flat feature vector composed of 378 units, which we call *embedding*.

This embedding was employed in two other classification settings. The first consisted of simply feeding the embedding to a fully-connected network, just like the one illustrated in Fig. 2a, and training the fully-connected network regardless of the original CNN that generated the embedding. The second consisted of inserting a softmax layer at the top of the global average pooling layer, resulting in an architecture identical to that of the CNN depicted in Fig. 2b, and fine-tuning all its weights.

D. Attribution Methods

Algorithms to assign a value to the contribution of each input to a given output of a neural network may be called *attribution methods*. A comprehensive summary of different methods was presented on [25]. Formally, given an input $X = [x_1, \ldots, x_N] \in \mathbb{R}^N$ and an output $S(X) = [S_1(X), \ldots, S_C(X)]$, where N is the number of input neurons and C is the number of output neurons, the problem consists of assigning an attribution $R^C = [R_1^C, \ldots, R_N^C] \in \mathbb{R}^N$ of each input feature $x \in X$ with respect to a given output $S_k(X) \in S(X)$. The *Integrated Gradients* method [26], adopted in this work, is based on the gradients obtained through a single backward pass through the network.

Here, we applied the DeepExplain framework [25] to the outputs for computing the attributions of each instance with respect to the input signals and to all intermediate layers. The outcome, in the case of the inputs, may be represented by the example in Fig. 4, which presents the attribution of each timestep of the input channels as colour maps. It is worth mentioning that negative attributions, represented in blue, are

also present, and might be interpreted as evidence *against* the output analysed.

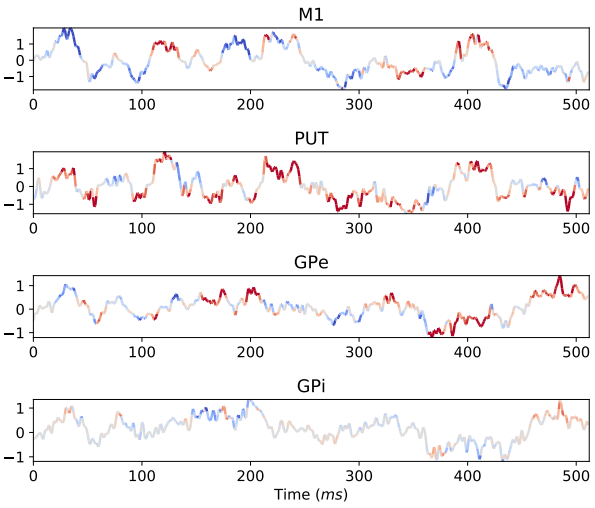


Fig. 4: Example of attributions at the input layer with respect to a given output. Contributions of each timestep are represented in a colour map, with red points corresponding to positive attributions, and the blue points, to negative attributions. In other words, red (blue) points relate to increased (decreased) probability of correct classification.

V. Results and Analysis

After preprocessing, 14 sessions from the healthy condition and 11 sessions from the PD condition were kept. Based on that, the data was split following the rule that segments that belonged to the same recording session would always belong to the same fold. This policy allowed the data to be split into 11 folds, each consisting of one healthy and one PD recording session, preventing the ML algorithms from achieving high accuracy by simply learning session-specific artifacts. The classification networks were trained to optimise the softmax cross-entropy loss, while the autoencoder was trained to optimise the mean squared error (MSE). All neural networks were trained using stochastic gradient descent (SGD), with learning rate 10^{-2} and decay 10^{-4}, for 30 epochs. We have chosen to apply SGD without momentum because this was the most stable training algorithm, generally leading to convergence on both train and test sets. The number of epochs was actually overestimated, since convergence appeared to happen around epoch 10, though it was kept for safety, since the loss remained stable after achieving an optimal set of parameters. The trained models were evaluated with regular evaluation metrics, but they were also analysed with respect to its features, as presented in the next subsections.

A. Performance Evaluation

The classification results are presented in Table I, including the shallow FC network applied to the autoencoder embedding and the pre-trained CNN, which is actually a fine-tuned autoencoder. Accuracy and macro F1-score (i.e., the harmonic mean between macro precision and recall) were close for all models, which suggest an equilibrium between true and false classifications across both classes. The results show that the CNN performed expressively better than the baseline FC network, with a 5.98% accuracy rise from 93.65% to 99.63% and standard deviation an order of magnitude lower. The ConvLSTM presented a perceptible improvement towards the CNN: the error rate, the opposite of accuracy, dropped from 0.37% to 0.20%, with even lower standard deviation. The FC applied to the autoencoder's embedding presented a slight improvement when compared to the baseline FC, despite an increase at the standard deviation, especially regarding the F1-score, which may suggest worse performance at certain circumstances. Pre-training the CNN had little effect on the classification metrics, as the accuracy of the CNN and the pre-trained CNN changed only 0.02%.

TABLE I: Classification metrics for each network architecture, with window size $t = 2,000$ points. Means and standard deviations between all folds.

	Accuracy (%)	F1-score (%)
Fully-connected	93.65 ± 6.03	93.39 ± 6.14
CNN	99.63 ± 0.78	99.61 ± 0.83
ConvLSTM	99.80 ± 0.40	99.79 ± 0.45
AE / FC	95.76 ± 7.93	94.49 ± 10.57
Pre-trained CNN	99.65 ± 0.68	99.63 ± 0.75

The dataset in which we performed the experiments is not public, thus there are no related work to which we can directly compare these results. Also, PD-related LFP data is not readily available for most research on the issue, even considering data from rodents. If compared to EEG datasets, collected under more controlled circumstances, our results would be consistent with the state-of-the-art, in which accuracy of up to 99.62% can be found with HOS features and SVM-RBF classifier [12]. Regarding deep neural networks, the CNN of [17] hit an accuracy of 88.25%, while [18] reported an accuracy around 85% with ESN classifiers.

The autoencoder's embedding went through an additional performance evaluation. Three clustering methods were applied to the feature vector - K-means, agglomerative hierarchical clustering and DBSCAN - and the clusters were evaluated according to entropy-based evaluation metrics [27], which take into account the labels of the instances assigned to each cluster. Those metrics were the Homogeneity of the clusters, according to which each cluster contains only instances of a single class, the completeness, according to which all instances of a given class are assigned to the same cluster, and the V-measure, the harmonic mean between the other two. The results, shown in Table II, give a measurement of whether the features learnt by the autoencoder and grouped by the clustering algorithms, both without considering the annotations, were informative of whether the instance corresponded to a healthy or PD subject.

K-means and agglomerative clustering performed better when the number of clusters was set to $n = 4$. In particular

TABLE II: Entropy-based metrics on clustering methods applied to the autoencoder's embedding. The number associated with the K-means and agglomerative clustering rows refer to the number of clusters n, set as a hyper-parameter of the algorithm.

	Homogeneity (%)	Completeness (%)	V-Measure (%)
K-means_2	53.00 ± 25.84	54.79 ± 24.66	53.63 ± 25.53
K-means_4	86.54 ± 14.19	56.15 ± 9.81	67.91 ± 10.99
Aggl._2	48.84 ± 30.27	51.99 ± 28.07	49.98 ± 29.58
Aggl._4	91.27 ± 14.22	59.85 ± 12.36	72.05 ± 12.87
DBSCAN	69.64 ± 41.65	53.83 ± 27.42	59.42 ± 34.87

considering the proportionally high standard deviations of the other approaches, which indicates that, for some folds, the embedding was not informative with respect to the labels. Even the completeness measurement, which could be expected to be lower when the number of clusters is higher than the number of classes, has actually improved. The homogeneity reached 91.27% with agglomerative clustering, an evidence in favour of the autoencoder's features as discriminative towards detection of PD. The density-based DBSCAN showed intermediate results, though with the highest standard deviations.

B. Feature Analysis

Features learnt by each model were analysed based on the attribution methods (Section IV-D) and spectral analysis. We have considered the input features and the internal representations at the intermediate layers of the convolutional networks.

1) Input Features: The attributions with respect to the input channels (i.e., regions of the motor loop) were used to determine the 1-second segments that show the highest accumulated attributions at each instance (i.e., the highest sum of 1,000 subsequent elements within a given channel of a given input), with the constraint that only segments whose sum of attributions is above a threshold of 1.0 were considered. The power spectral density (PSD) of those segments was computed using the Welsh method [28], and the mean μ_{PSD} of the spectra of each class $\mathcal{C} = \{\text{H}, \text{PD}\}$, where H means "healthy" and PD, "parkinsonian", was considered to calculate the ratio R of Equation 1. The rationale is that a peak at the beta frequency band (13-30 Hz) is a relevant marker of PD brain signals [29].

$$R = \frac{\mu_{\text{PSD}}(\mathcal{C} = \text{PD})}{\mu_{\text{PSD}}(\mathcal{C} = \text{H})} \tag{1}$$

Results for each model are presented in Fig. 5, alongside a baseline spectrum corresponding to random segments of each instance. The beta frequency peak can be clearly seen in random segments, and was enhanced on all models except for the autoencoder without fine-tuning. The autoencoder situation was expected, since the gradients were not updated with respect to the inputs of the network, but only to the encoding produced after the convolutions. The M1 and GPi ratios were close to zero because few segments of PD individuals with relevant attributions were present in the analysis.

As expected, the CNN and pre-trained CNN elicited high attributions to segments with similar spectral densities. The

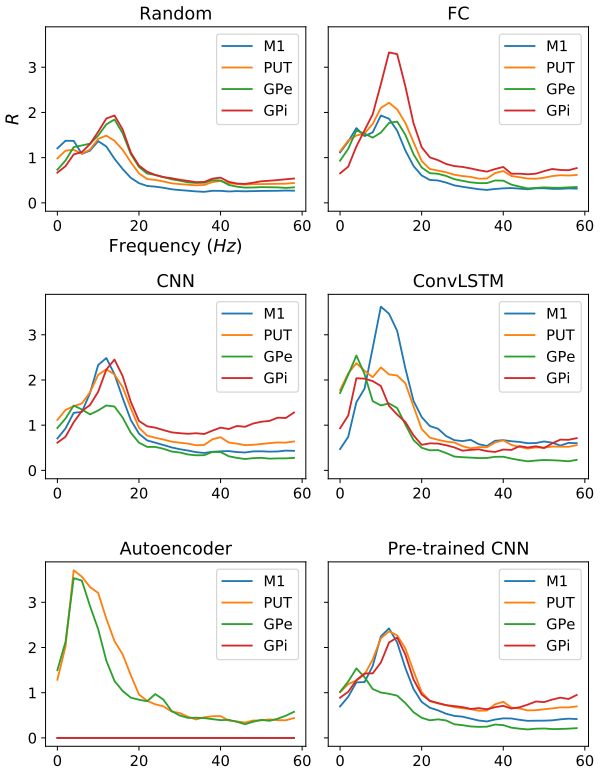


Fig. 5: Ratio between the mean PD and healthy PSD of the 1-second snippets with the highest accumulated attribution per input segment, above a given threshold of 1.0.

FC network was also consistent with the literature, with even a more acute beta peak in GPi. In the ConvLSTM spectrum, this peak was very high in M1, though less prominent in GPe and GPi. These differences in spectra show that each model make predictions based on different input features, however all of them were in consonance with previous PD literature.

To verify the contribution of each region for the models' performance, we evaluated the total number of regions with at least one 1-second segment whose sum of attributions was above the threshold of 1.0, across all folds. In Fig. 6, this evaluation is shown in terms of the proportion of segments above such threshold with respect to the total number of segments within each given region.

This evaluation suggests that the PUT and GPe regions were generally more relevant for recognising the healthy condition for all models, and also for recognising the PD condition for the ConvLSTM and the autoencoder-based network. Therefore, the particularly high frequencies for the GPi spectrum at the baseline FC and for the M1 spectrum at the ConvLSTM, previously shown in Fig. 5, does not imply that those regions have given the highest contributions for the classifications.

2) Intermediate Convolutional Layers: We also evaluated the features at the intermediate convolutional layers. Given the internal representation that followed each max-pooling layer, our analysis computed the spectral power at the delta (1-3 Hz),

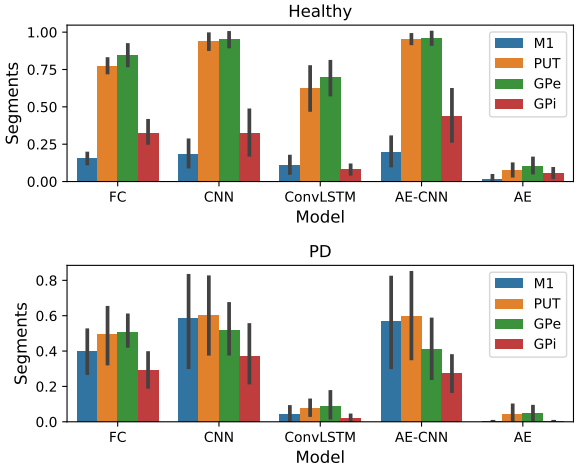


Fig. 6: Proportion of segments above threshold for each classification model (mean between all folds). In the graph, the pretrained CNN was named AE-CNN.

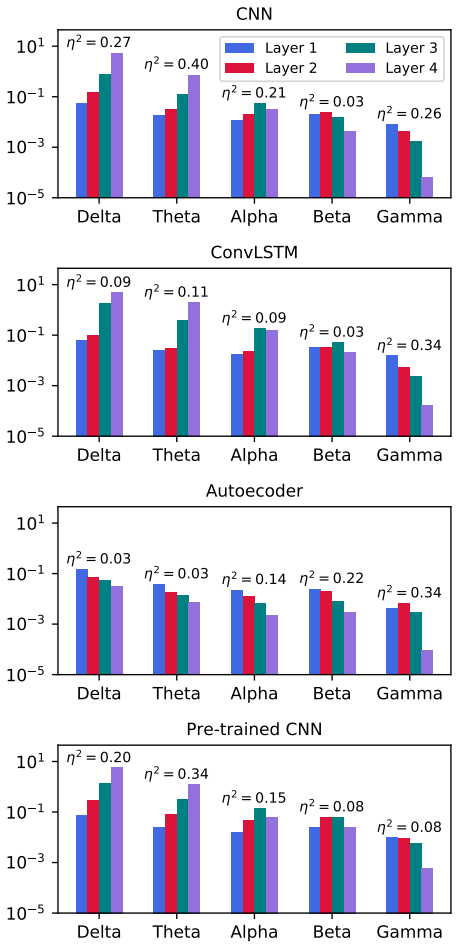


Fig. 7: Average frequency power bands over the spectrum of each max-pooling layer and η^2 effect size over all pairs of layers with summed attributions above a given threshold of 0.5 for the pre-trained CNN or 0.7 for the other models.

theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (30-100 Hz) bands of the LFP. We considered feature maps whose sum of attributions is above a threshold of 0.5 for the pre-trained CNN and 0.7 for the other models. The average spectral power of those representations is shown in Fig. 7.

Differences on the features learnt at each layer were also verified at a given power band. As the number of samples is massive in all of the considered cases, the outcome of a significance test would provide a very low p-value even if the effect of the significance detected was only trivial [30]. In fact, we got $p \approx 0.00$ for all ANOVA tests applied. Hence, in order to understand the effect size of this statistical significance, we measured the η^2 measure [31], also reported in Fig. 7. A small effect size is determined by $\eta^2 \in [0.01, 0.09]$, a medium one, by $\eta^2 \in [0.09, 0.25]$, and a large one, by $\eta^2 > 0.25$.

Except for the autoencoder, the evaluations of all models shared most of its properties. Regarding the sub-alpha waves (i.e., delta and theta), CNN, ConvLSTM, and pre-trained CNN produced feature maps with higher amplitudes the deeper the layer was, with medium to large effect sizes. This pattern started to reverse at the alpha band, with layer 4 producing less power at such frequency interval than layer 3. At the beta band, the pattern was less uniform across models, though relevant (i.e., large effect size for CNN and ConvLSTM). At the gamma frequency, the tendency of the lower bands was reverted, with first layers producing less of those waves. The high effect size was possibly due to the lower resolution at layer 4, which penalises spectral analysis of higher frequencies.

The autoencoder model was considerably different than other models, as one may expect due to its different, unsupervised optimisation strategy. It produced the same pattern at all sub-gamma frequency bands, with a higher prevalence of those frequencies the lower down was the layer. We highlight that only small effect sizes were detected at the delta and theta bands, and medium effects, at the alpha and beta ones. The layers were less specialised regarding the gamma waves. In common with the other models, the autoencoder has also shown a sharp drop of gamma waves at layer 4.

VI. Conclusion and Future Work

In this paper we proposed a deep framework to extract features related to Parkinson's Disease (PD) from Local Field Potential (LFP) brain signals of a marmoset monkey dataset. Different neural networks were applied as machine learning techniques, both as classifiers and autoencoders, and results were reported in terms of accuracy and properties of the representations learnt by each model.

The deep networks presented classification metrics higher than the shallow networks, with accuracy up to 99.80% for the ConvLSTM model. The autoencoder embedding has shown to be informative of the PD-related features, with clustering approaches reaching homogeneity up to 91.27%, and higher

classification metrics when fed to a fully-connected network, in comparison to the raw input (e.g., 95.76% accuracy, against 93.65%). Pre-training the CNN, on the other hand, had little effect compared to training from scratch.

Even though the convolutional networks extract features in the time domain, the input segments with higher attributions presented an enhanced peak at the beta frequency range of the average spectrum of the PD individuals when compared to the healthy ones. Regarding the intermediate representations of the convolutional layers, we have analysed the average power spectra at five frequency bands of feature maps with the highest attributions. Although LFP readings are not a feasible source of data for diagnosing PD, the proposed methods and analysis may contribute for a better understanding of the mechanisms underlying Parkinson's disease.

Future work includes the use of the same deep learning approach to simulated data originated from computational models of PD. This will assist on the validation of artificial models of the motor loop, apart form enhancing our current understanding of the PD neurophisiology. We will also embed such models into a robot, given rise to a neurorobotics model which could simulate the symptoms of this disease and provide a platform to perform preliminary experiments on proposed new therapies. A better understanding of the BG-T-C circuitry might give further insights on related systems regarding decision-making, homeostasis and learning [32]–[35], which are of particular interest to the field of robotics.

REFERENCES

[1] O. B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *Journal of Neural Transmission*, vol. 124, no. 8, pp. 901–905, 8 2017.

[2] J. J. Gaare, G. O. Skeie, C. Tzoulis, J. P. Larsen, and O.-B. Tysnes, "Familial aggregation of Parkinson's disease may affect progression of motor symptoms and dementia," *Movement Disorders*, vol. 32, no. 2, pp. 241–245, 2 2017.

[3] B. S. Connolly and A. E. Lang, "Pharmacological treatment of Parkinson disease: A review," *JAMA - Journal of the American Medical Association*, vol. 311, no. 16, pp. 1670–1683, 2014.

[4] H. Kita and T. Kita, "Cortical stimulation evokes abnormal responses in the dopamine-depleted rat basal ganglia," *Journal of Neuroscience*, vol. 31, no. 28, pp. 10 311–10 322, 2011.

[5] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes." *Nature reviews. Neuroscience*, vol. 13, no. 6, pp. 407–20, 5 2012.

[6] J. B. Koprich, L. V. Kalia, and J. M. Brotchie, "Animal models of α-synucleinopathy for Parkinson disease drug development," *Nature Reviews Neuroscience*, vol. 18, no. 9, pp. 515–529, 8 2017.

[7] M. B. Santana, P. Halje, H. Simplício, U. Richter, M. A. M. Freire, P. Petersson, R. Fuentes, and M. A. Nicolelis, "Spinal cord stimulation alleviates motor deficits in a primate model of Parkinson Disease," *Neuron*, vol. 84, no. 4, pp. 716–722, 11 2014.

[8] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images," *Cognitive Systems Research*, vol. 57, pp. 147–159, 2019.

[9] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep cnn approach to decode motor preparation of upper limbs from time–frequency maps of eeg signals at source level," *NN*, vol. 124, pp. 357–372, 2020.

[10] C. X. Han, J. Wang, G. S. Yi, and Y. Q. Che, "Investigation of EEG abnormalities in the early stage of Parkinson's disease," *Cognitive Neurodynamics*, vol. 7, no. 4, pp. 351–359, 8 2013.

[11] Y. J. Zhao, H. L. Wee, Y.-H. Chan, S. H. Seah, W. L. Au, P. N. Lau, E. C. Pica, S. C. Li, N. Luo, and L. C. Tan, "Progression of Parkinson's disease as evaluated by Hoehn and Yahr stage transition times," *Movement Disorders*, vol. 25, no. 6, pp. 710–716, 4 2010.

[12] R. Yuvaraj, U. Rajendra Acharya, and Y. Hagiwara, "A novel Parkinson's disease diagnosis index using higher-order spectra features in EEG signals," *Neural Computing and Applications*, vol. 30, no. 4, pp. 1225–1235, 8 2018.

[13] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 10 2018.

[14] Y. Li, m. Murias, s. Major, g. Dawson, K. Dzirasa, L. Carin, and D. E. Carlson, "Targeting EEG/LFP synchrony with neural nets," in *NIPS*, Long Beach, CA, EUA, 2017, pp. 4620–4630.

[15] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization." *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[16] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *4th ICLR*, San Juan, Puerto Rico, 2016, pp. 1–15.

[17] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, and U. R. Acharya, "A deep learning approach for Parkinson's disease diagnosis from EEG signals," *Neural Computing and Applications*, pp. 1–7, 2018.

[18] G. Ruffini, D. Ibañez, M. Castellano, S. Dunne, and A. Soria-Frisch, "EEG-driven RNN classification for prognosis of neurodegeneration in at-risk patients," in *ICANN*. Springer, 2016, pp. 306–313.

[19] Y. Yao, J. Plested, and T. Gedeon, "Deep feature learning and visualization for EEG recording using autoencoders," in *ICONIP-LNCS-vol. 11307*, 2018.

[20] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018.

[21] M. M. McGregor and A. B. Nelson, "Circuit Mechanisms of Parkinson's Disease," pp. 1042–1056, 3 2019.

[22] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K. M. Su, and K. A. Robbins, "The PREP pipeline: Standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, pp. 1–19, 6 2015.

[23] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," in *NIPS)*, 2018.

[24] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 1 2016.

[25] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for Deep Neural Networks," in *6th ICLR*. OpenReview.net, 2018, pp. 1–16.

[26] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *34th ICML*, vol. 7. IMLS, 2017, pp. 5109–5118.

[27] A. Rosenberg and J. Hirschberg, "V-Measure: A conditional entropy-based external cluster evaluation measure," in *Proc. of the EMNLP-CoNLL*. Association for Computational Linguistics, 2007, pp. 410–420.

[28] K. D. Rao, M. Swamy, K. D. Rao, and M. Swamy, "Spectral analysis of signals," in *Digital Signal Processing*. Springer, 2018, pp. 721–751.

[29] G. Tinkhauser, A. Pogosyan, H. Tan, D. M. Herz, A. A. Kühn, and P. Brown, "Beta burst dynamics in Parkinson's disease off and on dopaminergic medication," *Brain*, no. 11, 2017.

[30] B. Lantz, "The large sample size fallacy," *Scandinavian Journal of Caring Sciences*, vol. 27, no. 2, pp. 487–492, 6 2013.

[31] T. R. Levine and C. R. Hullett, "Eta squared, partial eta squared, and misreporting of effect size in communication research," *Human Communication Research*, vol. 28, no. 4, pp. 612–625, 2002.

[32] P. Vargas, R. Moioli, F. Von Zuben, and P. Husbands, "Homeostasis and evolution together dealing with novelties and managing disruptions," *International Journal of Intelligent Computing and Cybernetics*, vol. 2-3, pp. 435–454, 2009.

[33] R. C. Moioli, P. A. Vargas, and P. Husbands, "A multiple hormone approach to the homeostatic control of conflicting behaviours in an autonomous mobile robot," in *IEEE CEC*, 2009, pp. 47–54.

[34] M. Keysermann and P. Vargas, "Towards autonomous robots via an incremental clustering and associative learning architecture." *Cognitive Computation*, vol. 7-4, pp. 414–433, 2015.

[35] C. Rizzi, C. G. Johnson, F. Fabris, and P. A. Vargas, "A situation-aware fear learning (safel) model for robots," *Neurocomputing*, vol. 221, pp. 32 – 47, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216310529

Ranieri, C. M.; Pimentel, J. M.; Romano, M. R.; Elias, L. A.; Romero, R. A. F.; Lones, M.; Araujo, M. F. P.; Vargas, P. A.; Moioli, R. C. . "A data-driven biophysical computational model of Parkinson's Disease based on marmoset monkeys". *arXiv preprint arXiv:2107.12536*, 2021.

Contribution statement: Moioli and Vargas designed and supervised the project. Moioli, Araujo and Ranieri preprocessed the raw data from the marmoset monkeys. Romano ported the original code to the NetPyNE framework. Ranieri designed and implemented the machine learning methods to fit the model to the marmoset's data. Lones supervised the implementation of the evolutionary algorithm. Ranieri, Pimentel, Moioli, Romano and Elias performed the analyses of the results. All authors contributed to writing and revising the draft of the paper.

A DATA-DRIVEN BIOPHYSICAL COMPUTATIONAL MODEL OF PARKINSON'S DISEASE BASED ON MARMOSET MONKEYS

Caetano M. Ranieri
Institute of Mathematical and Computer Sciences
University of Sao Paulo
Sao Carlos, SP, Brazil
cmranieri@alumni.usp.br

Jhielson M. Pimentel
Edinburgh Centre for Robotics
Heriot-Watt University
Edinburgh, Scotland, UK
jm210@hw.ac.uk

Marcelo R. Romano
School of Electrical and Computer Engineering
University of Campinas
Campinas, SP, Brazil
marcelorromano@gmail.com

Leonardo A. Elias
School of Electrical and Computer Engineering
University of Campinas
Campinas, SP, Brazil
leoelias@unicamp.br

Roseli A. F. Romero
Institute of Mathematical and Computer Sciences
University of Sao Paulo
Sao Carlos, SP, Brazil
rafrance@icmc.usp.br

Michael A. Lones
Edinburgh Centre for Robotics
Heriot-Watt University
Edinburgh, Scotland, UK
M.Lones@hw.ac.uk

Mariana F. P. Araujo
Health Sciences Centre
Federal University of Espirito Santo
Vitoria, ES, Brazil
mfparaujo@gmail.com

Patricia A. Vargas
Edinburgh Centre for Robotics
Heriot-Watt University
Edinburgh, Scotland, UK
p.a.vargas@hw.ac.uk

Renan C. Moioli
Digital Metropolis Institute
Federal University of Rio Grande do Norte
Natal, RN, Brazil
renan.moioli@imd.ufrn.br

July 28, 2021

ABSTRACT

In this work we propose a new biophysical computational model of brain regions relevant to Parkinson's Disease (PD) based on local field potential data collected from the brain of marmoset monkeys. Parkinson's disease is a neurodegenerative disorder, linked to the death of dopaminergic neurons at the substantia nigra pars compacta, which affects the normal dynamics of the basal ganglia-thalamus-cortex (BG-T-C) neuronal circuit of the brain. Although there are multiple mechanisms underlying the disease, a complete description of those mechanisms and molecular pathogenesis are still missing, and there is still no cure. To address this gap, computational models that resemble neurobiological aspects found in animal models have been proposed. In our model, we performed a data-driven approach in which a set of biologically constrained parameters is optimised using differential evolution. Evolved models successfully resembled single-neuron mean firing rates and spectral signatures of local field potentials from healthy and parkinsonian marmoset brain data. As far as we are concerned, this is

the first computational model of Parkinson's Disease based on simultaneous electrophysiological recordings from seven brain regions of Marmoset monkeys. Results show that the proposed model could facilitate the investigation of the mechanisms of PD and support the development of techniques that can indicate new therapies. It could also be applied to other computational neuroscience problems in which biological data could be used to fit multi-scale models of brain circuits.

***K*eywords** basal ganglia · brain modelling · computational modelling · evolutionary computation · neural engineering · Parkinson's Disease · 6-OHDA lesioned marmoset model

1 Introduction

Parkinson's disease (PD) affects more than 3% of people over 65 years old, with figures set to double in the next 15 years [67]. It is a neurodegenerative disease, whose symptoms include cognitive and motor deficits. In late stages, it can possibly also lead to depression and dementia [89]. There is still no cure, and current therapies are only able to provide symptomatic relief.

PD is characterised by a dopaminergic neuronal loss within the substantia nigra pars compacta (SNc), which leads to a dysfunction of the basal ganglia-thalamus-cortex (BG-T-C) circuit. The BG-T-C circuit is a neuronal network with parallel loops that are involved in motor control, cognition, and processing of rewards and emotions [61, 74]. There are also links between the degeneration of dopamine neurons within those brain regions and changes on electrophysiological behaviour [22].

A commonly used model to explain how PD affects the neural connections within this circuit, also known as the motor loop, is the so-called classic model, illustrated in Figure 1a. It consists of projections from primary motor (M1) and somatosensory cortical areas to BG input structures, specifically the putamen (PUT) and the subthalamic nucleus (STN). In PUT, the cortical projections establish excitatory glutamatergic synapses with medium spiny neurons (MSNs).

The MSNs establish two distinct pathways to the BG output nuclei (globus pallidus pars interna – GPi and substantia nigra pars reticulata – SNr). The MSNs from the direct pathway (dMSN) directly project to the GPi/SNr, while the MSNs from the indirect pathway (iMSN) project to the globus pallidus pars externa (GPe), which in turn send projections to the GPi/SNr directly or indirectly via the STN (for reviews, see Obeso *et al.* [61], Lanciego *et al.* [42], and McGregor and Nelson [49]).

The cortical projection to the STN establish a third pathway, often called the hyperdirect pathway [58]. Activation of the direct pathway facilitates movement by inhibiting the activity of GPi/SNr, thus reducing the inhibition of the ventral anterior nucleus (VA) and the ventral lateral nucleus (VL) and increasing the excitatory thalamic input to the motor cortex. Activation of the indirect and hyperdirect pathways, on the other hand, inhibit movement by increasing the inhibitory activity of the GPi/SNr over the VA/VL, hence decreasing the excitatory thalamic input to the motor cortex.

The activity of the motor loop is modulated by dopaminergic projections from SNc to PUT. The main effect of dopamine (DA) release in PUT is movement facilitation, since DA increases the excitability of the dMSNs and decreases the excitability of the iMSNs.

In PD, the depletion of striatal DA leads to an enhanced activation of the indirect pathway and a decreased activation of the direct pathway, resulting in the characteristic motor symptoms of this neural disorder [93]. In addition to changes in firing rates, the functional imbalance within the motor loop in PD also disrupts the firing patterns within each nucleus and amongst the structures of the BG-T-C circuit, increasing neuronal synchronisation, neuronal bursting, and enhancing the oscillatory activity at the beta frequency band [21].

Brain regions linked to PD present complex interactions, with mutual excitatory and inhibitory feedback loops, which limit a comprehensive understanding of the physiopathology of the disease. Studies aiming at investigating the mechanisms underlying PD often use animal models. In classic animal models of PD, symptoms are elicited by delivering neurotoxins that damage the SNc dopaminergic neurons, such as 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and 6-hydroxidopamine (6-OHDA), or chemicals that transiently inhibit dopamine production, such as alpha-methyl-p-tyrosine (AMPT) [38]. Also, antipsychotics like haloperidol have side effects that may promote dystonia and parkinsonism [35].

Bilateral 6-OHDA lesions in the marmoset medial forebrain bundle induce several PD motor symptoms, including impairments in fine motor skills, limb rigidity, bradykinesia, hypokinesia, and gait impairments. Alpha-methyl-p-tyrosine (AMPT) administration to 6-OHDA lesioned marmosets can transiently increase the severity of these symptoms. However, like MPTP macaques, these animals do not exhibit resting tremor. Santana *et al.* [83] provides an extensive characterisation of these symptoms, that were quantified through manual scoring (adapted version of the unified PD

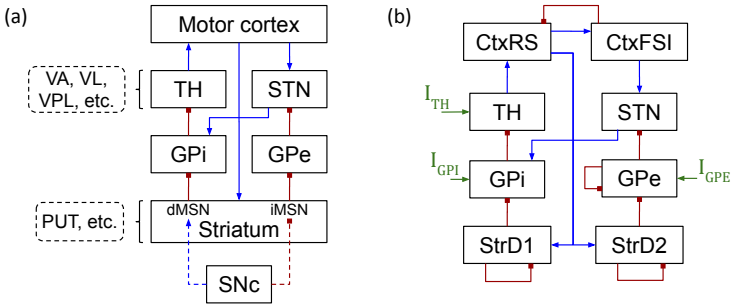


Figure 1: Models of the basal ganglia-thalamus-cortex (BG-T-C) circuit, central to the underlying mechanisms of Parkinson's Disease (PD), including excitatory (blue) and inhibitory (red) connections between the regions involved. (a) Classical model of BG-T-C circuit. The motor loop in the mammalian brain is formed by the *motor cortex* (M1), the *thalamus* (TH) - composed of structures such as the *ventral anterior nucleus* (VA), the *ventral lateral nucleus* (VL), and the *ventral posterolateral nucleus* (VPL) -, and the *basal ganglia* (BG), the latter composed of a subset of structures: the *striatum*, which itself includes the *putamen* (PUT) and the *caudate nucleus*, the *globus pallidus*, divided into *pars interna* (GPi) and *pars externa* (GPe), the *subthalamic nucleus* (STN), and the *substantia nigra*, divided into *pars compacta* (SNc) and *pars reticulata* (SNr). PD is caused by the loss of dopaminergic neurons in the SNc, which weakens the connections represented by dashed lines and leads to malfunctioning of both direct and indirect pathways. (b) BG-T-C network used in this work, based on [40]. The cortex is represented by regular spiking (CtxRS) excitatory neurons and fast spiking (CtxFSI) inhibitory interneurons. The direct and indirect pathways in the striatum were modelled separately, representing the medium spiny neurons (MSNs) modulation by D1 and D2 dopamine receptors, respectively.

rating scale for marmosets), automated assessments of spontaneous motor activity in their home cages (using actimeters), and automated motion tracking while the animals explored two experimental apparatuses.

To date, no animal model of PD fully reproduces human features of the disease. In addition, due to experimental limitations, animal data often include only a limited set of PD-related brain regions, with subjects engaged in different behavioural settings. In this context, computational models, with biologically informed constraints that can be selectively altered, are a promising, complementary approach to advance our knowledge about PD beyond that obtained from anatomical and physiological studies [33, 59]. Some PD-related anomalies observed in animal models, and efforts to reproduce those in computational models, are presented by Rubin *et al.* [79].

Computational models are established tools to facilitate understanding of neural disorders [65, 82, 86] and, in the context of PD, accommodate several levels of description and range from focusing on disease mechanisms to understanding anomalous neuronal synchronisation [33].

For instance, Pavlides *et al.* [63] conducted a detailed study to help unveiling the mechanisms underlying beta-band oscillations in PD and compared computational model predictions with experimental data. Muddapu *et al.* [55] studied loss of dopaminergic cells in the SNc due to neural dynamics between SNc and STN, shedding light on the relevance of ongoing neural activity and neural loss. Gurney *et al.* [24] described mounting evidence relating the BG-T-C network and action selection mechanisms; actually, computational models showed a close relationship between action selection and BG-T-C oscillatory activity [30, 31, 50].

Moren *et al.* [54] proposed a model of the spiking neurons within the BG-T-C circuit, in order to observe the asynchronous firing rates around the 15 Hz beta-range oscillations, as well as on lower frequency bands. Terman *et al.* [96] developed a conductance-based computational network model which shed light on the mechanisms underlying the neural dynamics of STN and GPe, a model which was further developed by Rubin *et al.* [78] to investigate the effects of deep brain stimulation (DBS) to eliminate anomalous synchronisation within the BG-T-C network in PD condition. In fact, one of the key areas in which computational models serve as an invaluable tool for developing novel therapies is that related to predicting the effects of DBS [32, 46].

Finally, based on a collection of previously published studies, Kumaravelu *et al.* [40] developed a computational model of the BG-T-C network tuned for the 6-OHDA *rat model* of PD (Fig. 1b). Compared to other computational models [33], it was the first to specifically consider 6-OHDA and a single species.

Most computational models related to BG-T-C dynamics rely on rodent data [31, 37, 45], with only a handful focusing on primate data [44, 98]. The research by Shouno *et al.* [88], for instance, provided a spiking neuron model of the

recurrent STN-GPe circuit for studying dysfunctions in oscillations within the 8-15 Hz frequency band for PD primate models.

All mammals have a similar set of BG structures that are similarly connected with thalamic and cortical structures. Nevertheless, recent studies suggest subtle differences between species [7,27,44,101], also in the neuropathophysiology of PD [16,38], with primates (including marmosets) being more similar to humans than rodents. For example, there are differences in the distribution of dopaminergic neurons in substantia nigra of rats and primates, and the subthalamic nucleus and internal globus pallidus of rats have less neurons containing parvalbumin than primates [27]. Thus, a primate computational model of PD is of paramount relevance.

In this work, we developed a new computational model of PD based on published data from the BG-T-C brain circuit of marmoset monkeys [15]. We built upon the neuronal computational model of rat models of PD [40], and adjusted its parameters to match the electrophysiology data from 6-OHDA+AMPT marmoset model of PD [83,84].

It is important to highlight that, in our work, we are using the LFP signal data to tune and validate our model, not spikes or other biosignals, thus the whole optimisation framework relates to LFP-based metrics. We are aware that there are several simplifications in our computational model, nevertheless results were shown that LFP power spectral densities at frequencies of interest, firing frequency dynamics, and spike coherence resemble those from healthy and PD marmosets.

The main contributions of this paper are: (i) the first computational model of PD validated on simultaneous, multi-site electrophysiological recordings (e.g., LFP recordings) from a marmoset monkey model of the disease, and (ii) an optimisation framework that can easily include novel biophysical parameters as soon as they become available.

This paper is organised as follows. In Section 2, the building blocks of the computational model are depicted, as well as the free parameters that were optimised, the algorithm to update those parameters, the experimental setup, and the evaluation protocol. In Section 3, the results are presented regarding the optimisation process, the parameters learnt by the machine learning algorithms, and the metrics observed on the simulations of the computational models provided, considering spectral densities from simulated LFP, dynamics of the firing rates from simulated neurons, and coherence analyses. In Section 4, a discussion is presented in order to contextualise our results and compare them with the expectations from the data from animal models, and knowledge from the literature. In Section 5, a conclusion is presented with a brief summary of what was presented.

2 Methods

To provide a computational model of the BG-T-C circuit for PD-related features in primates, we began by re-writing the code by Kumaravelu *et al.* [40], originally implemented in Matlab. We have ported the original code to the Python programming language, with the NetPyNE framework and the libraries from the NEURON simulator [18,28]. Then, we performed a series of adaptations and employed a data-driven approach to calibrate a set of parameters, in order to derive a model that resembles local field potentials from marmoset data [83,84]. More specifically, we employed an optimisation technique called differential evolution (DE), an algorithm based on evolutionary computation [2]. This approach consists of optimising a predefined set of parameters (i.e., genotype) by gradually adapting them through successive steps (i.e., generations), providing variability and selection of the best solutions (i.e., individuals) through mechanisms analogous to biological evolution.

In the model by Kumaravelu *et al.* [40], no noise was introduced in the simulations. Neuronal connectivity and membrane initial conditions can be stochastic, and neuronal models include synaptic transmission delay. The dataset employed in our work was suitable to calibrate such a model, since it was collected from marmoset monkeys that were not engaged in any particular task, that is, they were moving freely, without any time-marked events such as sensory or artificial stimuli.

After having calibrated our marmoset model, different analyses were performed in order to enhance and validate it. The dataset used as ground-truth for adjusting the parameters of the computational model is not publicly available due to legal restriction, but it is available from the corresponding author on reasonable request. The next subsections will provide a detailed description of the methods employed. The code to reproduce the results from this paper, including the machine learning framework and the analyses of the results, is publicly available at `https://github.com/cmranieri/MarmosetModel`.

2.1 Computational Model

The computational model was based on Kumaravelu *et al.* [40]. Their model was build to reproduce the neurophysiological behaviour from rats based on data from healthy and 6-OHDA-lesioned individuals. As an initial step, we did an

alternative implementation for their model within the NetPyNE framework, and we validated this implementation by comparing its outputs with those reported in [40].

Briefly, eight brain structures were modelled and connected based on a simplified version of the classic model (Figure 1b). In particular, the direct and indirect pathway in the striatum were modelled separately representing the MSN modulation by D1 and D2 dopamine receptors, respectively [49]. The cortex is represented by regular spiking (RS) excitatory neurons and fast spiking (FSI) inhibitory interneurons. Neurons from all but cortical regions were modelled using a biophysically based Hodgkin–Huxley [29] single-compartment model, whereas cortical neurons were constructed based on the computationally efficient Izhikevich's model [34]. The reasoning for different neuronal models lies on the fact that PD effects are captured by altering specific conductances in selected structures (see below), thus a conductance-based model is more suitable at these locations. Finally, a bias current was added in the TH, GPe, and GPi, accounting for the inputs not explicitly modelled. Remarkably, even though no oscillatory inputs are present in the model, synaptic delays and network interactions by means of recurrent connections promote sustained firing rate oscillations. For a detailed description of connectivity schemes and other implementation details, the reader is referred to Kumaravelu *et al.* [40].

The computational model described above can shift from the simulations of the healthy to the PD conditions by altering three conductances [40]: decreasing the maximal M-type potassium conductance in direct and indirect MSN neurons (MSN firing disfunction) from 2.6 to 1.5 mS/cm^2; decreasing the maximal corticostriatal synaptic conductance (reduced sensitivity of direct MSN to cortical inputs) from 0.07 to 0.026 mS/cm^2; and increasing the maximal GPe axonal collaterals synaptic conductance from 0.125 to 0.5 mS/cm^2 (increase of GPe neuronal firing). This is implemented in the model with a control flag.

One major addition to the model developed here is the simulation of local field potentials (LFP). These measurements are related to the extracellular activity produced by action potentials of the neurons within a brain region [23]. A discussion on the behaviour of LFP signals within the basal ganglia and its consequences to humans, especially regarding conditions such as PD, was presented by Brown and Williams [9]. The NetPyNE function for LFP calculation is based on the work of Parasuram *et al.* [62]. The LFP amplitude in each simulated electrode is obtained by summing the extracellular potential contributed by each neuronal segment, calculated using the line source approximation and assuming an Ohmic medium with conductivity 0.3 mS/mm. Thus, the electrical activity of neurons from each brain region contributes to the peaks and valleys recorded at each electrode (subject to extracellular medium attenuation).

In our work, first, each simulated brain region is assigned to a spatial 3D coordinate that matches that used in the stereotaxic surgery where electrodes were placed in the real marmoset monkeys [64, 84]. Then, a simulated electrode is placed at the centre of each region. In our model, each neuron is represented as a single cylindrical compartment with a membrane area of 100 μm^2. For each electrode, NetPyNE estimates the simulated LFP by summing the extracellular potential contributed by each neuronal segment (based on the transmembrane current generated from the single cylindrical source neuron), calculated using the "line source approximation" method and assuming an Ohmic extracellular medium with conductivity $\sigma = 0.30\ mS/mm$ [62].

2.2 Dataset and preprocessing procedures

The dataset we used in the present work is based on a previous study by Santana *et al.* [84]. Our dataset includes data from three adult males and one adult female common marmosets (i.e., *Callithrix jacchus*). Data from two males were part of the aforementioned study; data from one male and one female are novel and followed exactly the same experimental procedures. A short summary is presented in the next subsection, followed by the preprocessing steps.

2.2.1 Dataset

The animals, weighing 300–550 g, were housed in a vivarium with natural light cycle (12/12 hr) and outdoor temperature. All animal procedures followed approved ethics committee protocols (CEUA-AASDAP 08/2011, 11/2011 and 03/2015) strictly in accordance with the National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals. PD symptoms were elicited in all three male animals with injections of 6-OHDA toxin in the medial forebrain bundle under deep anaesthesia [83, 84]. Prior to neural recordings, animals that received 6-OHDA were subjected to acute pharmacological inhibition of dopamine synthesis (subcutaneous injections of AMPT 2×3240 mg/kg) to further exacerbate PD motor symptoms, mimicking a more severe stage of the disease. Although 6-OHDA lesions impact on both behavioural and electrophysiological features in all animals [83, 84], there are individual differences at earlier stages of dopaminergic depletion that could hinder our model development considering the relatively low number of subjects.

Both healthy and PD animals were implanted each with two custom-made microelectrode arrays composed of 32 microwires (one array in each hemisphere). The wires were 50 μm in diameter and were organised in bundles aimed to

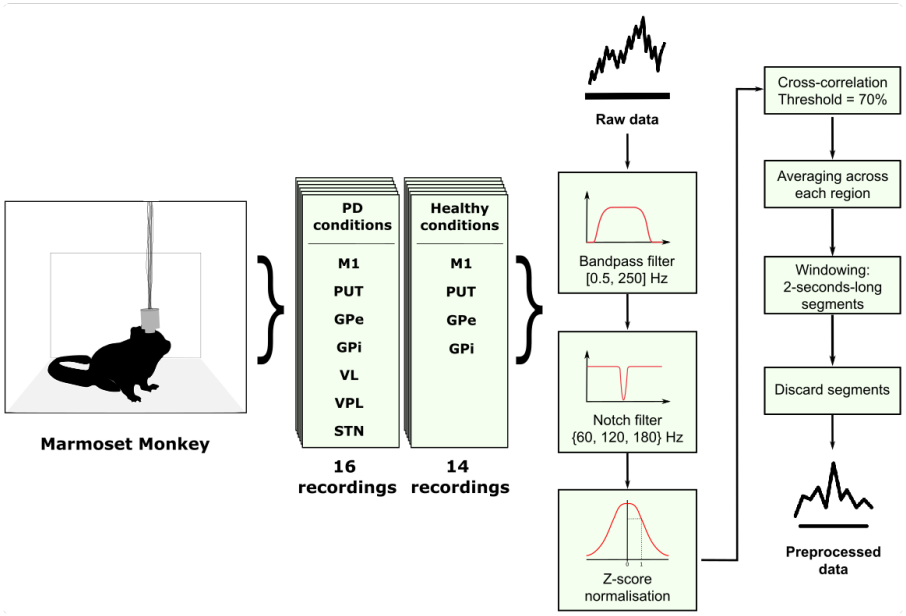


Figure 2: Data acquisition and preprocessing steps implemented in out method. Depending on the monkey condition, different regions of the brain were recorded. The input data was composed of a whole recording session, with variable lengths and numbers of channels (i.e., electrodes) per region. After preprocessing, the data was transformed into 2-seconds-long segments with seven channels, each related to one of the regions analysed.

reach distinct areas of the BG-T-C system. Before the surgery, the animals were sedated with ketamine (10-20 mg/kg i.m.) and atropine (0.05 mg/kg i.m.), followed by deep anesthesia with isoflurane 1-5% in oxygen at 1-1.5 L/min. The arrays were then implanted using a stereotaxic manipulator to position electrodes at the targeted BG-T-C coordinates, which were determined using Stephan *et al.* [92] and Paxinos *et al.* [64] stereotaxic atlas. The microelectrode array and the implant procedures were thoroughly described in Budoff *et al.* [10].

Once the animals recovered from the surgery, recording sessions were performed in fully awaken animals behaving freely. LFPs were sampled at 1,000 Hz and recorded using a 64 multi-channel recording system (Plexon). The position of the recording microelectrodes were verified postmortem through either tyrosine hydroxylase (TH) staining or Nissl staining. Similarly, the extent of dopaminergic lesions were verified through the quantification of striatal fiber density and dopaminergic midbrain cells in TH-stained sections. Further experimental details are described in Santana *et al.* [84].

2.2.2 Preprocessing

For our study, in total, 14 and 16 recording sessions were taken for the healthy and PD conditions, respectively, considering the brain hemispheres independent from each other. For the PD condition, we recorded from M1, PUT, GPe, GPi, ventral lateral (VL) and ventral posterolateral (VPL) thalamic nuclei, and STN, whereas for the healthy animal regions M1, PUT, GPe, and GPi were recorded. The raw data was organised so that, for each recording session, a data structure with $N_{\text{elec}} \times N_T$ was provided, where N_{elec} is the number of electrodes recorded and N_T is the number of samples of the recording session (variable but typically lasting for several minutes).

In Figure 2, are illustratrated the preprocessing steps adopted after data acquisition. For each channel, the pipeline began with a zero-lag low-pass filter (cutoff frequency of 250 Hz) and a high-pass filter (cutoff frequency of 0.50 Hz), to eliminate frequencies that are outside the LFP scope and may relate to electrical or mechanical interference. Then, we minimised power grid interference (hum) with a notch filter centred at 60 Hz and its harmonics (120 Hz and 180 Hz). Each resulting signal was then scaled according to a z-score normalisation, to account for the possible differences in signal amplitude due to different electrode impedance.

In the next step, we computed the cross-correlation matrix Q according to Equation 1, where C_{ij} is the covariance matrix of the filtered and z-scored signals from electrodes i and j, which are located exclusively within a brain region. Channels within each region with mean correlation coefficient below the threshold of 0.70 were discarded. This procedure was employed because electrodes in each recorded region are placed very close to each other (see electrode

and surgical procedures above), thus we expect LFP signals to be highly correlated (if they are not, it may relate to a noisy electrode signal) [12].

$$Q_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} \cdot C_{jj}}} \tag{1}$$

All remaining LFP channels within a brain region were averaged, which provided one data matrix for each recording session with dimensions $N_R \times N_T$, where N_R is the number of brain regions recorded. These average LFP values were computed based solely on channels within each region. Next, we segmented each time-series in 2-second segments, which was the same length of the computational model simulations (Section 2.3 will bring the details). Considering the data sampling rate (1,000 Hz) and frequencies of interest (up to 50 Hz), 2-second segments provide enough data for our analyses. Prohibitively noisy segments were discarded using two criteria: first, segments with abnormal amplitudes, detected using an upper threshold of 0.20 for the absolute value of the mean of the signal over time; second, segments with limited (abnormal) oscillatory patterns, detected using a lower threshold of 0.10 for the amplitude standard deviation and a minimum threshold of 10 amplitude peaks. For each recording session, our preprocessed dataset had a final shape of $N_R \times 2000 \times N_{\text{seg}}$, where N_{seg} is the resulting number of segments.

In the dataset adopted for this work, whether animals were still or moving could have a profound effect on brain oscillatory activity and synchronisation metrics, because all animals were behaving freely and were not engaged in any particular behavioural task during the recording sessions. In fact, especially in motor and pre-motor regions, modulations in neural oscillatory dynamics linked to motor activity are well characterised (see Armstrong *et al.* [1] for a review), and recent studies show that even breathing can modulate neural oscillations [99]. However, we understand that action initiation, movement, or breathing have low influence on averaged LFP amplitude values computed, given that the 2-second window segments were randomly selected without time alignment to any specific movement or action.

2.3 Evolutionary Algorithm

Evolutionary algorithms are optimisation techniques in which a set of parameters, called *genotypes*, are gradually combined and changed according to mechanisms analogous to those of biological evolution, in order to maximise a fitness function dependent of those parameters [2]. Differential evolution (DE) [71] was employed to fit the computational model parameters so that it matches the LFP beta-band power spectrum observed in the marmoset data.

The overall structure of the model was preserved from Kumaravelu *et al.* [40], while a set of conductances, background currents and synaptic modulations, as well as the numbers of neurons in each region of the BG-T-C circuit, were calibrated through the evolutionary algorithm. The connectivity, the delays, the synaptic mechanisms, the remaining conductances, and all other parameters were kept as in the original model (see Section 3 from the Supplementary Material).

More specifically, fourteen parameters compose the set of parameters to be optimised (i.e., the genotype). Parameter I_{TH} ($\mu A/cm^2$) relates to cerebellar input currents to the thalamus, which are linked to sensorimotor inputs [47]. Parameters I_{GPe} ($\mu A/cm^2$) and I_{GPi} ($\mu A/cm^2$) relate to currents at GPe and GPi, respectively, from all sources that were not explicitly modelled. The next two parameters, g_{STN_KCA} (nS/cm^2) and g_{GP_AHP} (nS/cm^2), refer to the maximum slow potassium conductance yielding afterhyperpolarization (AHP) at the STN and the calcium-activated potassium conductance at GPe and GPi, respectively. The sixth parameter, $g_{syn\_CTX\_STR}$ (nS/cm^2), modifies the synaptic conductance from cortex (CTX) to striatum (STR). Finally, parameters seven to 14 map to the number of neurons in each modelled region. All of the aforementioned parameters were chosen because they have a direct influence on the firing rates of neurons within each region, which in turn affect the LFP [62]. Also, comparing marmoset with rodent literature, there is very limited quantitative work on the anatomical and neurophysiological parameters of the BG-T-C neuronal network.

In the DE, each individual from the population was a model $M(G)$ that consisted of an adaptation of the model of Kumaravelu *et al.* [40], in which the parameters of Table 1 were set to the values defined by genotype G. Each model $M(G)$ was simulated for $t_{sim} = 2000$ milliseconds, and the spike trains from each neuron and LFPs from each virtual electrode were recorded. The LFP recordings were applied to calculate the fitness function $f(M)$ as follows.

Given a categorical set R containing N_R brain regions, the mean power spectral density (PSD) of the LFP from the electrode placed in region $r \in R$ is denoted by S_r and defined in Equation 2, where $[\omega_a, \omega_b]$ is the frequency interval of interest and $\hat{P}_r(\omega)$ is the periodogram computed with the Welch's method [73].

$$S_r(\omega_a, \omega_b) = \int_{\omega_a}^{\omega_b} \hat{P}_r(\omega) d\omega \tag{2}$$

Table 1: Free parameters of the computational model, optimised by DE to fit the marmoset data.

ID	Parameter	Range	Description
1	I_{TH}	$[0.6, 1.8]$	Background currents at TH ($\mu A/cm^2$)
2	I_{GPe}	$[1.5, 4.5]$	Background currents at GPe ($\mu A/cm^2$)
3	I_{GPi}	$[1.5, 4.5]$	Background currents at GPi ($\mu A/cm^2$)
4	$g_{\text{STN\_KCa}}$	$[2.5, 7.5]$	Ca^{2+}–dependent AHP K^+ conductance at STN (mS/cm^2)
5	$g_{\text{GP\_AHP}}$	$[5.0, 15.0]$	Ca^{2+}–dependent AHP K^+ conductance at GPe and GPi (mS/cm^2)
6	$g_{\text{syn\_ctx\_str}}$	$[0.8, 1.2]$	Synaptic modulation from cortex to striatum (mS/cm^2)
7	n_{GPe}	$[10, 30]$	Number of GPe neurons
8	n_{GPi}	$[10, 30]$	Number of GPi neurons
9	n_{TH}	$[10, 30]$	Number of TH neurons
10	n_{StrD1}	$[10, 30]$	Number of StrD1 neurons
11	n_{StrD2}	$[10, 30]$	Number of StrD2 neurons
12	$n_{\text{CTX\_RS}}$	$[10, 30]$	Number of CTX_RS neurons
13	$n_{\text{CTX\_FSI}}$	$[10, 30]$	Number of CTX_FSI neurons
14	n_{STN}	$[10, 30]$	Number of STN neurons

According to the literature on the electrophysiology of PD [67,97], a noticeable abnormality is observed typically at the centre of the beta frequency band of LFP recordings from the basal ganglia of PD individuals. This frequency band corresponds approximately to the interval [13,30] Hz, although this range varies within human patients and animal model species. For the formulation of the fitness function, let a coefficient y_r be the summation of the beta-band mean PSD plus the mean PSD of adjacent bands, composing the interval [8,50] Hz, normalised by the mean PSD of all frequencies up to 50 Hz, as stated in Equation 3. This broader interval was defined to account for possible wider spectrum modulations in adjacent bands.

$$y_r = \frac{S_r(8, 50)}{S_r(0.5, 50)} \tag{3}$$

The fitness function $f(M)$ is defined in Equation 4, where $y_{r(target)}$ is the average value of Equation 3 calculated from the preprocessed data of all marmosets of PD condition, and $y_{r(M)}$ is calculated considering the simulated LFP of a computational model M. Notice that the healthy marmoset condition lacks readings from TH and STN regions (i.e., no electrodes were implanted in these regions). In addition, the dataset includes three PD model animals. For this reason, DE optimised parameters for mimicking the PD condition. Fitness values vary from 0, if simulated and marmoset data LFP in all brain regions substantially differ, to N_R, if they match.

$$f(M) = N_R - \sum_{r \in R} \min \left\{ 1, \left| \frac{y_r(M) - y_r(target)}{y_r(target)} \right| \right\} \tag{4}$$

Eight brain regions are simulated, thus $N_R = 8$. PSD target values for the simulated regions StrD1 and StrD2 are drawn from marmoset LFP PSD values for PUT. Simulated TH is tuned based on the average PSD from marmoset VL and VPL, and simulated CtxRS and CTxFSI are tuned based on marmoset M1. Simulated GPe, GPi, and STN LFP PSDs are matched to the respective marmoset LFP PSDs.

The DE initial population was set to 200 individuals, whose initial parameters were drawn from a random uniform distribution in the interval $[0, 1]$. Parameters were normalised to the ranges listed in Table 1 (i.e., the actual values set in the computational model) only at simulation time. In each DE generation, a set of 20 individuals were selected through tournaments of size two. Pairs of those selected individuals were randomly chosen, in order to generate two offspring by applying uniform crossover. This led to a child population of size 20. The mutation rate was set to 10% and followed a normal distribution $\mathcal{N}(\mu = 0.0, \sigma = 0.3)$. The DE implements generational replacement with elitism, with only one elite individual of the parent population being kept, resulting in a population size of 21 individuals. Each model $M(G_k)$, where $k \in \{1, \ldots, N_M\}$, was evolved for $N_{gen} = 60$ generations. We have performed 150 evolutionary runs, so that the highest fitness individual of each run was selected to compose the set $\mathcal{G} = \{G_1, \ldots, G_{N_M}\}$ of evolved genotypes.

2.4 Clustering

Upon completion of parameter optimisation by the DE, we investigated whether high fitness individuals had different genotypes. The rationale is that different parameter sets, even if biologically plausible, could lead to incompatible healthy and PD network dynamics [4]. Considering that the fitness function was computed based on LFP values of the PD condition only, and that the healthy condition was obtained by changing the same parameters listed by Kumaravelu *et al.* [40], there was no guarantee that the genotypes evolved would lead necessarily to models that resemble the healthy and PD conditions of the animal models. For this reason, we performed a clustering analysis [102] to the set $\mathcal{G}$ of evolved genotypes, which we could then evaluate separately based on their spectral densities. This validation step is based on the fact that PD individuals present a peak at the beta band (13-30 Hz) when compared to healthy individuals [97].

Let $\mathcal{C} = \{C_1, \ldots, C_{n_c}\}$ be a set of clusters, with $C_p = \{G_1, \ldots, G_{n_p}\}$, where n_c is the number of clusters, $p \in \{1, \ldots, n_p\}$, and n_p is the total number of genotypes within cluster p. Considering $s_p(G_k)$ to be the sample silhouette [77] of genotype G_k with respect to $C_p \in \mathcal{C}$, consider $s_p(G_k) \geq s_p(G_{k+1})$ for all $k \in [1, P]$, it is, each cluster is ordered from highest to lowest silhouette. In exploratory experiments (not shown), we investigated different clustering paradigms, namely K-means, density-based spatial clustering of applications with noise (DBSCAN), and agglomerative clustering. Based on these experiments, we opted for the K-means algorithm with two centroids (i.e., $p = 2$), because this configuration led to the highest mean silhouette score. Hence, the K-means algorithm was fed with all the individuals with the highest fitness per evolutionary run (i.e., set $\mathcal{G}$), and the euclidean distances for the algorithm were computed on the 14 normalised parameters of the genotype.

2.5 Computational model spike and LFP analysis

The different clusters of genotypes were compared with respect to their parameter values, spike firing rates and LFP power spectra. For each cluster, the 50 highest fitness genotypes were chosen for the following analyses. Spectral analysis was performed by simulating $C_p[1, \ldots, 50]$, for $t_{sim} = 2000$ milliseconds, in both healthy and PD conditions. Thus, for each condition, 50 simulated LFP recordings were analysed per cluster for each condition. Since we simulated the same individuals (i.e., sets of parameters), with the same seeds for generation of random numbers, in each of the conditions (healthy and PD), the samples across these conditions were considered to be dependent. The PSDs were computed and evaluated with respect to the mean of the density spectrum per cluster, and the average power at the beta band.

For PSD analyses on the LFP of either the animal and computational models, to highlight the presence of a peak in the beta band in the PD condition, a ratio R was defined as in Equation 5, where $\hat{P}_r^{PD}(\omega)$ and $\hat{P}_r^{H}(\omega)$ are the mean spectral power across the PD and healthy models, respectively, for frequency ω. A lower threshold value ϵ was defined because, for denominators too close to zero, the ratio may lead to high values that actually has little meaning for interpretation. For the analyses with the animal models, ϵ was defined as the median power across the mean spectrum of the healthy condition. For the computational models, it was set to the percentile 80 of the healthy spectrum.

$$R(\omega) = \begin{cases} \frac{\hat{P}_r^{PD}(\omega)}{\hat{P}_r^{H}(\omega)} & , \quad \hat{P}_r^{H}(\omega) > \epsilon \\ 0 & , \quad \text{otherwise} \end{cases} \tag{5}$$

Regarding spike dynamics, the models within each cluster were simulated for $t_{sim} = 2000$ milliseconds with time step size $dt = 0.10$ milliseconds, always with the same seed for random number generation, and the firing frequency of all neurons was calculated in 50 time bins, each corresponding to 20 milliseconds.

2.6 Computational model validation

Considering that different currents, conductances, and numbers of neurons may influence the firing rate in each simulated brain region, which in turn modulates the LFP power spectra, one may conclude that even if there are different clusters, their neural dynamics are comparable because both clusters are formed by high fitness individuals. However, even if our computational model was optimised to replicate the LFP power spectra from marmoset animal models of PD, it should also mimic the power spectra from healthy marmosets (by changing selected conductances, see Section 2.1). In other words, if the computational model accurately captures the physiological phenomena responsible for the different beta-band centred LFP power spectra from PD marmoset monkeys, it should also replicate the healthy spectra (a scenario in which it was not evolved).

Therefore, we first confirmed that our marmoset animal model of PD presented frequency spectra in accordance with previous works, following Section 2.2. Then, we investigated whether the computational model would also capture this

phenomena. For that, for each genotype cluster found (Section 2.4), we compared the LFP power spectra from the evolved PD computational model with that from the healthy model. This was performed by modifying a predefined set of conductances in the simulation (Section 2.1). To highlight the differences, we first analysed the ratios between the mean PSD of the PD and healthy simulated individuals from each cluster.

During evolution, fitness is given by LFP PSD in the vicinity of the beta band calculated in the whole $t_{sim} = 2000ms$ sequence, hence it is possible that the same spectra relate to different LFP rhythms over shorter time scales. Thus, different neuronal spiking dynamics may lead to similar LFP dynamics over time. Moreover, spikes from single neurons are noisy and vary considerably over time and over repeated simulations. With large recordings, joint neuronal averages over time may hinder comprehension of neural population dynamics. Finally, one of the advantages of computational models such as the one used here is the direct access of each neuron state at any given time, but it is not trivial to interpret the dynamics of large populations of neurons over time. To clarify these issues, we studied low-dimensional neuronal trajectories for both healthy and PD computational model conditions [14].

To compute the neuronal trajectories, we first calculated the firing frequencies for all neurons from each simulated model in a particular cluster and condition (i.e., healthy and PD), based on the mean firing rates (MFR) taken from bins of size 50 ms. Since the number of neurons within each region varies from 10 to 30 (see Table 1), and there are eight regions considered for the computational model, this procedure generates time series with high dimensionality, ranging from 80 to 240, which would be difficult to visualise and analyse. To reduce the dimensionality, we employed principal component analysis (PCA) [105], that is, we analysed neural trajectories by projecting high-dimensional neural population activity in a 3D space using PCA of the spike MFR time series.

However, what if, instead of clearly occupying different regions in the state space, neuronal responses from the same conditions result in similar paths in the reduced dimensional space? To address this hypothesis and to compare PCA trajectories, we used Dynamic Time Warping (DTW) with Euclidean distance [57]. DTW finds the optimum non-linear alignment between two time series, hence it can estimate whether neuronal trajectories share a similar path, regardless of initial conditions. In the analysis performed, we employed the *fastdtw* Python package, which implements the method proposed by Salvador and Chan [81]. Each pair of three-dimensional time series, computed from the MFR signals and dimensionally reduced with PCA, was fed to the algorithm, which provided, as output, a scalar proportional to the dissimilarity between the two time series being compared.

More specifically, we compared the similarity of all possible pairs of neural trajectories considering all individuals within the clusters (healthy and PD dynamics). We compared all pairs of trajectories generated by individuals within the same condition (healthy or PD), which gave a measurement of how different the healthy or PD individuals are compared to each other (i.e., within-group comparison), and we compared pairs of trajectories between healthy and PD conditions (i.e., between-groups comparison).

Finally, one of the hallmarks of PD is the anomalous widespread synchronisation in the BG-T-C network. To validate our model in that aspect, we calculated the magnitude-squared coherence between nuclei and intranucleus. Based on a similar analysis performed in healthy and PD marmosets reported in Santana *et al.* [84], we expect a widespread increase in this metric. The magnitude-squared coherence was calculated from the spike trains of neurons of each nucleus using Welch's method with Hanning windowing without overlap and with spectral resolution of 1 Hz. The average was taken as recommended by Bendat and Piersol [5]: the squared value of the average of the cross spectra divided by the product of the mean values of the auto spectra of each nucleus.

The value of the magnitude-squared coherence between brain regions r_A and r_B, defined as $C(r_A, r_B)$, was computed as in Equation 6, where N_A is the number of neurons in region r_A, and N_B is the number of neurons in region r_B, and $S(r_x^m, r_y^n)$ is the cross spectrum between the spike trains from the m-th neuron from region r_x and the n-th neuron from region r_y.

$$C(r_A, r_B) = \frac{\left[\frac{1}{N_A \cdot N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} S(r_A^i, r_B^j)\right]^2}{\left[\frac{1}{N_A} \sum_{i=1}^{N_A} S(r_A^i, r_A^i)\right] \cdot \left[\frac{1}{N_B} \sum_{i=1}^{N_B} S(r_B^i, r_B^i)\right]} \tag{6}$$

Then, we considered the peak of the coherence in the 7-30 Hz band to highlight PD-related effects [84]. The significance level for coherence was defined as $1 - (1 - \alpha)^{1/(L-1)}$ [76], with $\alpha = 0.95$ and $L = 100$, because the windowing was done with 100 segments and we adopted as 95% the significance level. As the computational models have eight nuclei, an 8×8 matrix was constructed, representing the coherence between each pairs of nuclei. The median of this matrix

was considered as the global coupling metric between nuclei in each simulation, because it is less sensitive to outliers than the mean.

3 Results

Based on two-seconds-long segments, computed according to the data preprocessing steps described in Section 2.2, (see Figure 3a for a sample), the PSDs of LFPs from healthy and PD marmosets were computed (see Figure 3b for the average spectrum). In all regions of the PD subjects, an increased PSD magnitude from 5 Hz to 25 Hz was observed, which is in accordance with the reported electrophysiological signatures of PD [97].

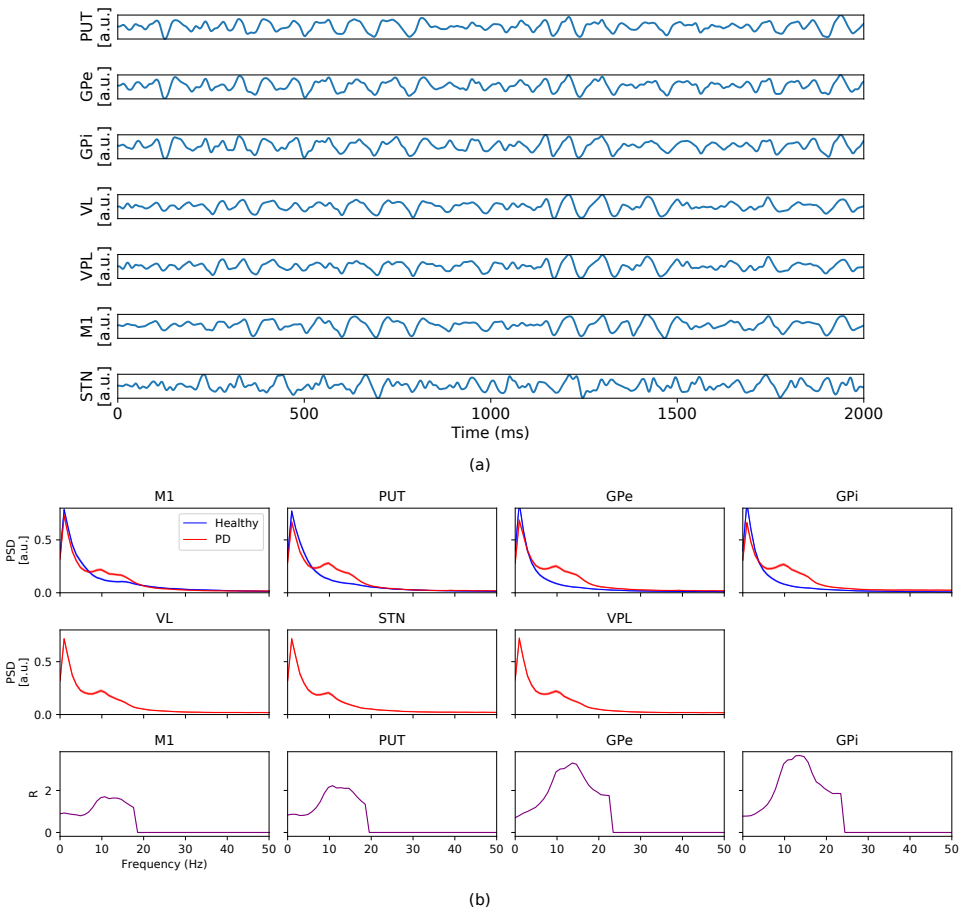


Figure 3: Animal data from marmoset monkeys, collected through electrodes implanted to each region of the BG-T-C circuit in a previous study [84], and made available for our research. (a) Example of a two-second time-window of the preprocessed LFP of a PD-induced (i.e., 6-OHDA lesioned) marmoset. For a clearer visualisation, signals were bandpass filtered to the [8,50] Hz interval, only for this panel. (b) Top two panels show the mean power density spectra (PSD) over all segments for the healthy (blue) and PD (red) marmosets (data for each individual marmoset is included as supplementary material). For thalamic regions (i.e., VL and VPL) and STN, only 6-OHDA lesioned hemispheres are represented, since these regions were not recorded in the healthy marmoset. PSDs were normalised by the maximum PSD value for each time-window. The bottom panel shows the ratio (R) between PD and healthy PSD for each frequency (see Equation 5). To improve visualisation, ϵ is set to the median of the healthy spectrum. a.u.: arbitrary units.

From the estimated LFP power spectra from PD marmosets, the target LFP power spectra values for the computational marmoset model were computed as in Equation 3. The results, presented in Table 2, were fed to the DE fitness function (Equation 4).

Table 2: Target values obtained from the marmoset LFP data from the animal model, PD condition.

y_{StrD1}	y_{StrD2}	y_{TH}	y_{GPi}	y_{GPe}	y_{CtxRS}	y_{CtxFSI}	y_{STN}
0.44	0.44	0.38	0.46	0.42	0.39	0.39	0.37

3.1 Evolutionary algorithm successfully found high fitness genotypes

After running the DE $N_M = 150$ times, the resulting set of high fitness individuals $\mathcal{G}$ (i.e., the highest fitness individual in the population at the end of each of the 60 generations at each evolutionary run) was analysed. The fitness values of all individuals were recorded at all generations of each evolution.

Figure 4 reports the best and mean individual fitness across generations, and the distribution of those values at the end of the evolutionary runs. Concretely, the best individual in a given generation is the set of parameters that led to the highest fitness value according to Equation 4. The mean individual fitness across generations refers to the average fitness of all individuals achieved at each generation.

Regarding the best individual fitness curve, results show that, at every evolutionary run, the initial population contained at least one individual with fitness value close to 6, and that value improved by approximately 1 at the end of evolution (the maximum fitness value possible is 8.0, see Equation 4). Considering the whole population, the initial average fitness was low (approximately 4.5), reaching a plateau close to 5.75 as evolution progressed. The mean fitness across individuals and the best individuals fitness have marginal improvement after generation 40, thus the DE was stopped at $N_{gen} = 60$ generations.

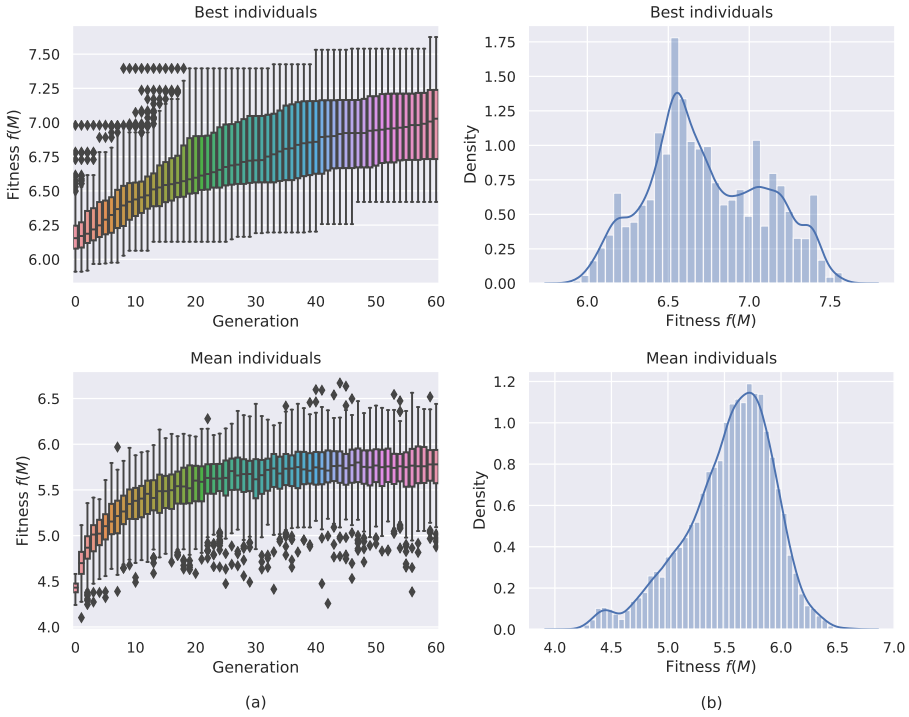


Figure 4: Fitness values $f(M)$ (Equation 4) per generation of the evolutionary algorithm (box-plots regarding the $k = 150$ runs at each generation). The genotypes (i.e., parameter sets for the free parameters elicited in Table 1) were meant to maximise f, which, by definition, would be upper bounded at 8.0. The upper panel refers to the highest fitness individuals at each evolutionary run, and the lower panel, to the mean fitness values of all individuals. (a) Box-plots of the best (upper panel) and mean (lower panel) fitness values at each generation. Outliers were represented by black diamonds. (b) Probability distribution of the best (upper panel) and mean (lower panel) fitness.

For all $G \in \mathcal{G}$, we looked at the distribution of parameter values for clusters C_1 and C_2, represented in Figure 5. Both clusters present similar distributions for most of the parameters, either with small variance (e.g., the numbers of neurons at the cortex populations) or more uniform distributions with high variance (e.g., the number of neurons at the striatum). Other parameters, such as I_{TH} and I_{GPe}, had a clear mean peak and reduced variance in the distribution for C_2, but a large variance for C_1.

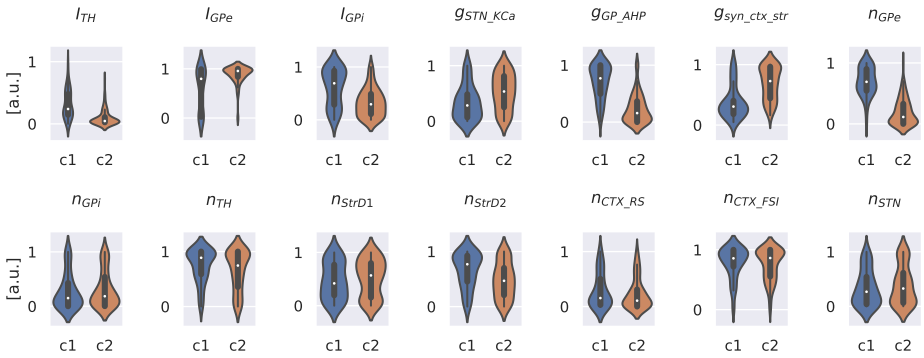


Figure 5: Violin plots showing the distribution of each free parameter (see Table 1) across the best individuals found at each run of the evolutionary algorithm employed for optimising this set of parameters (i.e., genotype). Although scales vary across parameters (see Table 1), all parameters were linearly scaled (i.e., normalised) to the interval $[0, 1]$ at evolution time. For example, for parameters 7-14 (i.e., the numbers of neurons), a value of zero corresponds to the lower bound of the parameter interval, that is, 10 neurons. a.u.: arbitrary units.

3.2 High fitness genotypes form two clusters

A set of 150 high fitness individuals was generated by repeatedly running the evolutionary algorithm with different seeds. It is possible that high fitness individuals do not have a unique parameter distribution, and diverse parameter settings could lead to high fitness values. To investigate this issue, we performed a clustering analysis based on the evolved individuals.

Following the methods from Section 2.4, the K-means algorithm was employed to determine $p = 2$ clusters. Figure 6 provides a radar plot representation of genotypes learnt for each cluster, and the correspondence between the mean value of each parameter and those of the rat computational model by Kumaravelu *et al.* [40].

Figure 6a shows 4 representative genotypes $c_p[1, \ldots, 4]$, chosen based on the highest silhouettes with respect to each cluster. For comparison, the parameters from the rat model [40] are superposed with the mean values between all individuals from both clusters in Figure 6b. This representation highlights substantial differences between clusters. For instance, the I_{GPe} is at its maximum value in C_2, while it shows a much lower value for C_1. On the other hand, the number of neurons at the GPe is higher in C_1 than in C_2.

3.3 Healthy and PD spectral signatures from computational model resembles those from marmoset monkeys

Regarding the spectral analyses of simulated sessions of the computational model, we employed the same procedure for normalisation as we did for the spectra of the animal model (see Figure 3), that is, we normalised each data segment by the maximum value. The sample signals of Figure 7a, shown as an example, were bandpass-filtered to the same range as in Figure 3a to the interval [8-50] Hz. The mean spectral power and the ratio R are shown for the healthy and PD conditions for each cluster in Figure 7b (see Equation 5).

In C_1, results show higher magnitudes of most frequencies up to 50 Hz for PD models, a fact that is less visible for C_2. The mean PSD ratio from genotypes $G \in C_2$ is close to 1 regardless of frequency range and brain region, whereas genotypes $G \in C_1$ show prominent peaks in beta frequencies. A detailed analysis of box-plots (Figure 7c) confirm the significant differences in the beta band for cluster C_1 only. Considering the animal spectra (Figure 3b), in which we observe a significant difference in the beta band of the LFP, results displayed in Figure 7c confirm that spectral signatures from genotypes in C_1 resembles those from marmoset monkeys. Notice that the LFP mean PSDs from the computational model (Figure 7b) has a different shape compared to that from the animal LFP (Figure 3b), but the spectral signature is similar in both healthy and PD conditions and resemble those from marmoset monkeys. This can

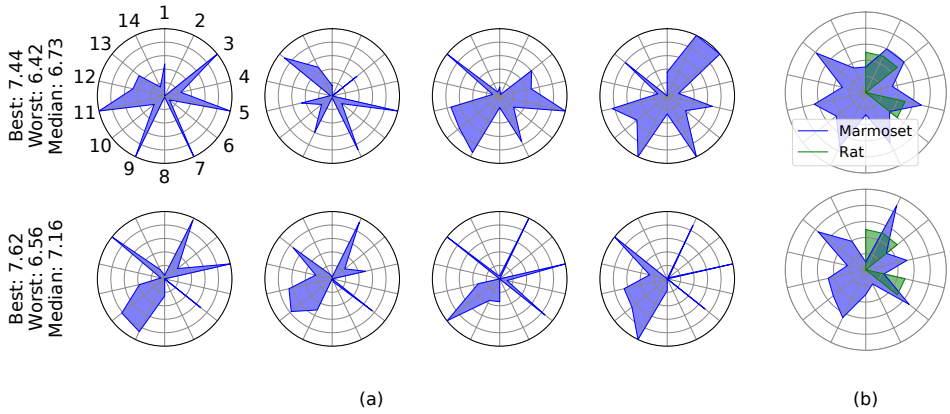


Figure 6: Radar representations of the genotypes (i.e., sets of parameters, see Table 1) from individuals at each cluster obtained by applying the K-means algorithm, applying these parameters as features for the clustering technique. Although scales vary across parameters (see Table 1), all parameters were linearly scaled (i.e., normalised) to the interval $[0, 1]$ at evolution time. For example, for parameters 7-14 (i.e., the numbers of neurons), a value of zero corresponds to the lower bound of the parameter interval, that is, 10 neurons. The first row represents cluster C_1 and the second row cluster C_2. (a) Four individuals with the highest silhouettes with respect to each cluster. Data at the left refers to the fitness f computed as in Equation 4. (b) Comparison between the parameters of the rat model and the mean values from each cluster. As in Figure 5, parameter values were scaled to the ranges shown in Table 1, except parameter 4 ($g_{\text{STN\_KCa}}$) of the rat model, whose original value is 1.0 mS/cm^2.

be explained by the relatively small number of neurons simulated in the computational model [62]. Therefore, for the forthcoming analyses, only $G \in C_1$ will be considered.

3.4 Spike activity from healthy models are significantly different from those of PD models

Regarding spike activity, the marmosets' dataset was not provided with a representative set of spike trains from all regions of the circuit, hence they were not a suitable ground-truth for validating the activity from the computational model. For this reason, the spikes synthesised by the computational model were analysed based on evidence from the literature [69].

First, we assessed the differences in mean firing rates (MFR) between the healthy and PD conditions for the marmoset-based computational models in cluster C_1. Figure 8a shows the simulated MFR in each brain region for $t_{sim} = 2000$ ms, considering the 50 models in C_1 with the highest silhouette with respect to the cluster. Results indicate a counter-intuitive relationship between the MFR and the LFP power spectra observed in Figure 7c. Consider, for instance, the GPe and GPi. Both regions show a higher beta-band LFP magnitude in PD condition, but while GPi MFR in PD condition is higher than that from healthy condition, GPe MFR is the opposite.

From Figure 8b and Figure 8c, we observe that neuronal trajectories are intertwined, with no clear difference in the reduced-dimension state space. This is justified by the relatively mild, though statistically significant, differences in MFR (Figure 8a). As described in Section 2.5, neuronal trajectories were compared with DTW in three scenarios: healthy vs healthy models (HxH), PD vs PD models (PDxPD), and healthy vs PD models (HxPD). As $len(C_1) = 53$, the number of pairs from which the DTW was computed was $len(DTW_{C_1}) = \binom{53}{2} = 1378$ for each scenario. The results from this analysis are shown in Figure 8d, in which the scalar outputs of the DTW algorithm are considered for all possible pairs within groups, for the HxH and PDxPD comparisons, or between groups, for the HxPD comparisons. Since two trajectories generated by the same individual were not compared on any of the analyses, we have computed statistical significance using unpaired tests, differently from the remaining analyses in the paper.

Trajectories from the HxH scenario were statistically more similar than trajectories from the other conditions. Thus, the intertwined trajectories observed in PCA (Figure 8b and Figure 8c) in fact relate to significant differences between healthy and PD neuronal dynamics. Interestingly, PDxPD trajectories differ more than those from HxH, which can be interpreted as a less homogeneous, regarding neuronal dynamics, genotype to phenotype mapping.

Figure 7: Extracellular activity simulated by the computational models resulting from the parameters optimised (i.e., genotypes computed with the evolutionary algorithm), modelled as local field potentials (LFP) at the centre of the regions involved in the BG-T-C circuit. The clusters C_1 and C_2 were computed by applying the K-means technique directly to the genotypes, hence were not influenced by the neurophysiological activity simulated. (a) Example of simulated LFPs for the highest silhouette evolved individual from cluster C_1, PD condition. For a clearer visualisation, signals were bandpass filtered to the [8,50] Hz interval, only for this panel. Compare with Figure 3a. (b) Mean PSD for healthy (blue) and PD (red) conditions from the 50 models with the highest silhouette of each cluster, normalised by maximum PSD value for each time-window, followed by the ratio R between PD and healthy PSD for each frequency (see Equation 5). To improve visualisation, ϵ is set to the percentile 80 of the healthy spectrum. (c) Box-plot regarding the beta band (13-30 Hz) of the LFP from the 50 models with the highest silhouette of clusters 1 (left) and 2 (right). Outliers were represented by black diamonds. Unpaired t-tests were applied to evaluate statistical significance against the null hypothesis that H and PD values are drawn from the same underlying distribution (p-value notation: $p > 0.05 \rightarrow$ ns; $p \in [0.01, 0.05] \rightarrow$ *; $p \in [0.01, 0.001] \rightarrow$ **; $p \in [0.001, 0.0001] \rightarrow$ ***; $p < 0.0001 \rightarrow$ ****). a.u.: arbitrary units.

3.5 Healthy and PD spike coherence from the computational model resembles that from marmoset monkeys

To conclude our model validation, we selected the top five genotypes with highest silhouette from cluster C_1 and calculated the magnitude-squared coherence (MSC) within and between each simulated brain region (Section 3.5) for healthy and PD conditions. Results revealed that computational models ran in the healthy condition provided a lower peak MSC in the 13-30 Hz band when compared to that from the PD condition (Figure 9a), with two important observations: genotype I has higher peak MSC when compared to the other 4 genotypes in the healthy condition, and genotypes II and III have a lower widespread peak MSC when in the PD condition compared to that from other genotypes in the same condition. Statistical analysis confirmed the significant differences in all five genotypes when

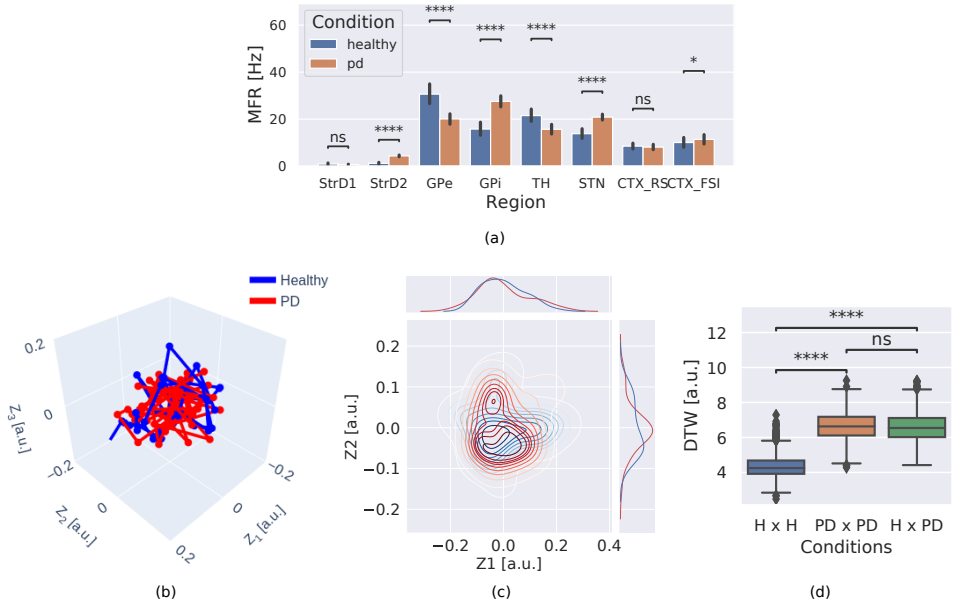


Figure 8: Firing rates and dynamics regarding the spike activity simulated with the computational models derived by the parameter sets from cluster C_1. Simulations were ran for $t_{sim} = 2000ms$. (a) Mean firing rates for each region in cluster (means and standard deviations). (b) Projection of three principal components of the most representative individual (i.e., highest silhouette) of cluster C_1, where Z_1, Z_2 and Z_3 are the principal components with the highest variance. (c) Representation of those components using contour lines. (d) Box-plot of the DTW between the dynamics of one simulation of all genotypes belonging to cluster C_1. All simulations were performed with the same seed for the generation of random numbers. Higher DTW values mean that the pairs of trajectories being compared are less similar to each other. Unpaired t-tests were applied to evaluate statistical significance in (a), against the null hypothesis that H and PD MFR values at each region are drawn from the same underlying distribution, and in (d), against the null hypothesis that a given pair of DTW vectors is drawn from the same distribution as each of the others (p-value notation: $p > 0.05 \rightarrow$ ns; $p \in [0.01, 0.05] \rightarrow$ *; $p \in [0.001, 0.0001] \rightarrow$ ***; $p < 0.0001 \rightarrow$ ****). a.u.: arbitrary units.

comparing the global coupling metric (see Section 2.6 and Equation 6) between healthy and PD conditions (Figure 9b), that is, PD models present a higher widespread coherence in the 13-30 Hz band than that observed in healthy models.

4 Discussion

Marmoset monkeys are prominent in neuroscience research [15, 36, 51, 53]. Although there are anatomical and physiological differences between BG-T-C circuit in rodents and primates, neurophysiological data from rodents are far more available than from primates. Considering that the structure of the BG-T-C circuit presents similar characteristics among all vertebrates [38], we assumed that the rat model presented by Kumaravelu *et al.* [40] was a suitable starting point to build a computational model of those structures in primates. The core hypothesis was that, by keeping the same brain regions and connectivity patterns of the rat model and modifying a set of parameters, the computational model could reproduce neural dynamics of healthy and PD marmoset conditions.

Our dataset comprised simultaneous LFP recordings from regions of the BG-T-C network and power density spectra (PSD) analysis revealed significantly higher 13 to 30 Hz LFP PSD magnitudes for PD marmosets on all regions. This result might be interpreted cautiously, given that one healthy marmoset is being compared to three PD marmosets. Also, results refer to a broad range of frequencies, hence different interval choices may influence the analysis. Nonetheless, one would expect a widespread significant increase in LFP power centred in (but not limited to) the beta band in PD affected brains [84, 97].

Regarding the MFR results from the computational model (Figure 8a), there are significant differences between the healthy and PD conditions. Single-neuron firing rates vary considerably depending on animal species, whether the animal is fully awaken, engaged in behavioural tasks, or anaesthetised [27, 49, 101]). Data from human subjects, even though scarce, are in line with animal results [17]. Moreover, there is a great neuronal diversity within the

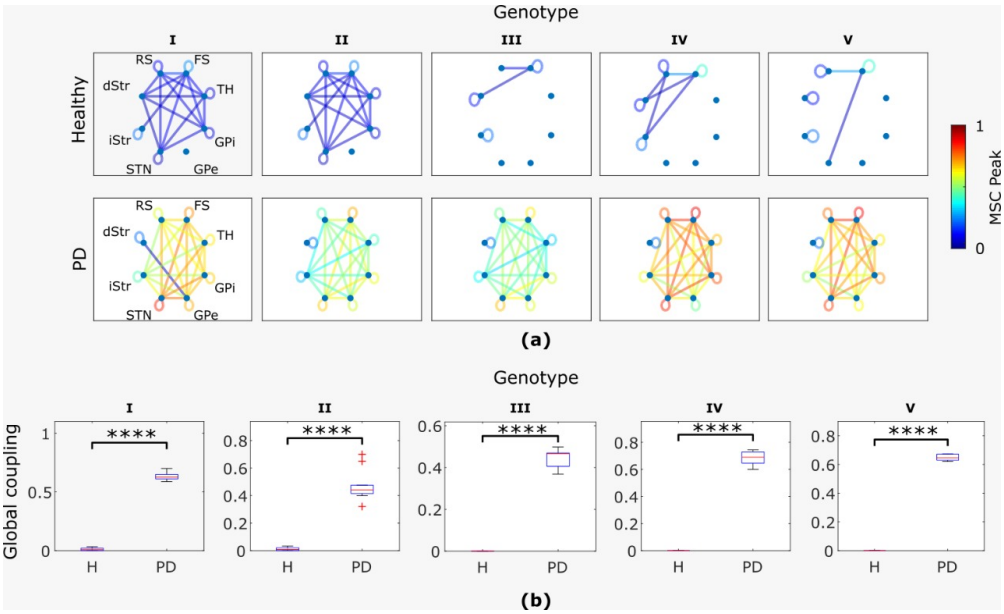


Figure 9: Coherence analyses computed for the spike activity of the five parameter sets, optimised through the evolutionary algorithm, with the highest silhouettes with respect to cluster C_1. These parameter sets could were employed to derive the healthy and PD computational models employed to these analyses. (a) Peak magnitude-squared coherence (MSC) in the 13-30 Hz band within and between each simulated brain region for the top five genotypes with highest silhouette from cluster C_1. Only connections whose peak MSC values are above significance level are shown. (b) Global coupling metric (median value of the MSC matrix) between brain regions for healthy and PD conditions (see Section 2.6 and Equation 6). (p–value notation: $p < 0.0001 \rightarrow$ ****)

BG-T-C network, both in terms of neuronal physiology and connectivity, which have been shown to have a non-trivial relationship with field potentials [6, 8, 30, 87]. Our model partially takes into account this diversity, nevertheless the reported MFR are in agreement with the literature: comparing PD with H conditions, a higher MFR in GPi, STN, and Str, and a lower MFR in GPe, TH, and CTX.

The data-driven modelling strategy adopted in this paper is consolidated in computational neuroscience literature [60], but often leads to multiple models fitting a particular data set [4]. Therefore, model optimisation should be followed by a model selection phase. We clustered high fitness solutions with respect to evolved parameters and obtained two clusters, and found two clear sets of parameters that reproduce the increased beta-band oscillations observed in PD marmosets [84]. However, when perturbing the model to shift from PD to healthy dynamics, only one of the clusters fitted the marmoset data. Notably, we evolved solutions based on LFP data but computational model firing rates resemble those reported in previous works [17, 43, 101]. Nevertheless, as data becomes available, future works should explore different fitness functions based on single-neuron activities or other features of LFP. Lastly, in this context, our simulated neurons are formed by a single cylindrical compartment, thus future works should consider using neurons with more and more complex compartments and connections, possibly including multiple dendritic branches and active ionic channels. This would lead to more realistic simulated LFP signals [62], but at the expense of heavier computing resources.

One of the great challenges in neuroscience is to link the activity of large neural populations to motor and cognitive behaviours. One strategy is to study the intrinsic high-dimensional dynamics of neural populations from its low-dimensional dynamics given by time-varying trajectories [14, 95], thus emphasising circuit over single-neuron function. For example, Humphries *et al.* [33] showed that neural low-dimensional dynamics given by PCA of neuronal activity can explain *Aplysia* rhythmic movement control and propose that only the low-dimensional dynamics are consistent within and between nervous systems. Also, the shape and amplitude of neural trajectories can explain different behavioural outcomes [25]. Combining PCA and DTW, we found that neural trajectories from high-fitness models are more similar in healthy conditions than in PD conditions. This is in line with results from Russo *et al.* [80], who demonstrated, using computer simulations, later confirmed by data from the supplementary motor area in monkeys, that low trajectory divergence is essential in neural circuits involved in action control. PCA is a simple, established method for dimensionality reduction, but other computational tools tailored to neuronal data, such as Gaussian-Process Factor

Analysis (GPFA) [106] and jPCA [13], should be considered in further analyses. Another possible approach is to use more advanced machine learning methods to identify PD-related features from neural data, likewise Ranieri *et al.* [72], who employed a deep learning framework to unveil PD features from marmoset data.

Finally, as part of our model validation, we assessed functional coupling within and between simulated brain regions by means of coherence between spike trains. In contrast to structural coupling, characterised by physical neuronal connections, functional connectivity is an emergent phenomenon commonly linked to synchronisation in neural rhythms in diverse spatiotemporal scales and is the basis of neural communication and cognitive processing [11, 20, 41, 90]. Several neural disorders, including PD, present a disruption in functional connectivity [26, 48, 100]. In particular, Santana *et al.* [84] showed that 6-OHDA marmoset models of PD have a widespread coherence peak in the beta band when compared to healthy individuals. Our computational model is in line with this result, which is relevant not only as further evidence of its biological plausibility, but also because one of the established therapies to alleviate PD motor symptoms is the use of deep brain stimulation (DBS) [67]. Thus, we believe that the work presented here can be used to test hypotheses that employ DBS. For instance, Romano *et al.* [75] performed a comprehensive analysis of frequency-dependent effects of DBS on the same model that we used here, tuned for rodent data [40], and found that neural oscillatory modulations were similar to those observed in electrical brain and spinal cord stimulation of primates [84, 103].

Certain simplifications inherent to our approach may be worth a mention, as they may serve as inspiration for improvements in future research. In our work, LFP generation followed the method described in Parasuram *et al.* [62], and implemented in NetPyNE, which does not consider the influence of sinks. Despite being a simplification, the method has been able to reproduce features of real LFP data, and is computationally feasible. In this approach, LFP peaks and valleys are directly related to transmembrane ionic currents from each neuronal source, which in turn relate to neuronal firing rates, and electrode position. As we have assigned coordinates to the simulated electrodes corresponding to the centre of each simulated region, we can assume that simulated LFP dynamics is due to altered spiking activity in multiple neuronal sources from different brain regions.

In our work, likewise Kumaravelu *et al.* [40] and previous seminal BG-T-C modelling works such as Humphries *et al.* [31], and van Albada and Robinson [101], we did not model any structural synaptic plasticity mechanisms. Our synapses were modelled as bi-exponential and alpha synapses, including transmission delays. Nevertheless, as model dynamics unfold, functional plasticity mechanisms may take place in the sense that the closed-loop, recursive network architecture could lead to single neurons and brain regions whose electrical activity are sensitive to past network states. In fact, the depletion of dopamine, one of the hallmarks of PD, affects structural and functional plasticity. Our model considers the loss of dopaminergic neurons (see Section 2.1, for a complete description), thus we believe that the model is suited for the investigation of functional plasticity phenomena. This analysis is beyond the scope of our work, but the reader can relate the change in oscillatory neural dynamics we described to different functional states. For instance, Humphries *et al.* [31] show that action selection in the BG is closely linked to oscillatory activity. In future works, we plan to use this model in a neurorobotics context, in which sensory inputs and motor responses can be used to highlight functional plasticity mechanisms differences between healthy and PD states.

5 Conclusion

Computational models are invaluable tools for advancing our knowledge on the neural dynamics of our brain, either under healthy conditions or with neurological disorders. In this work, we created the first computational model of PD based on data from Marmoset monkeys both in healthy and parkinsonian conditions. Our data-driven approach used simultaneous, multisite electrophysiological recordings from healthy and 6-OHDA+AMPT marmoset models of PD. Even though the physiopathology underlying PD share similarities across vertebrate species, there are important, species-specific differences in the anatomy and neural dynamics of the BG-T-C circuit. Hence, the design of a primate computational model of PD is of paramount importance.

Electrophysiological datasets from animal models often do not include comprehensive biophysical data such as single-neuron membrane conductances and neuronal cell densities. These parameters are central for building a biophysical computational model. Thus, to address this gap, we implemented a DE to search the multidimensional model parameter space for solutions that could reproduce features of the animal LFP recordings. Our model was based on a well known rat model of PD [40]. The main novelty aspects of our model are: 1) we use a marmoset monkey BG-T-C electrophysiological database; 2) we added LFP simulations to the model, in addition to spike dynamics; and 3) we developed a DE-based optimisation to search for unknown parameters. With this framework, we were able to reproduce several of the previously reported PD electrophysiological biomarkers observed and recorded from the Marmoset monkeys.

Our computational model present beta-band LFP power spectra differences between the healthy and the PD conditions, which Wang *et al.* [104] also found in human patients with dystonia. This is in line with a body of literature that shows that beta-band LFP modulations are not a PD-specific biomarker (see Poewe *et al.* [67] and references therein). Although our model is focused on PD, the electrophysiological features we use are known to be related to other neural disorders and thus should not be considered as exclusive to PD. Also, based on the study of Wang *et al.* [104], we suggest as future work to conduct the phase amplitude coupling in the STN experiment in our computational model.

Most PD computational models do not consider brain-body-environment interactions. Embodied cognitive science studies have provided solid evidence that neural activity is shaped by such interactions [3, 19, 56, 66]. In PD and other neural disorders, body-environment interactions influences motor control [85, 91], but its impact on neural dynamics remains unclear. Moreover, the BG-T-C neuronal network is clearly related to action selection and decision making [31, 52, 94]. Therefore, we believe that our marmoset-based computational model associated with robotics may offer an alternative approach to elucidate the mechanisms underlying brain-body-environment interactions in PD [24, 39, 68, 70]. A possible approach would be to employ this computational model in a sensorimotor loop based on visual inputs from video cameras and motor outputs to actuators such as the robot's motors. In this scenario, computer vision algorithms would transform the images into stimuli for the computational model, so that the resulting currents and action potentials would be used to generate perturbations that would govern the behaviours of the actuators. The resulting framework could become a new tool for studying the underlying mechanisms of PD and the effects of different interventions regarding the simulated circuit.

Acknowledgement

This work is part of the Neuro4PD project, granted by Royal Society and Newton Fund (NAF\R2\180773), and São Paulo Research Foundation (FAPESP), grants 2017/02377-5 and 2018/25902-0. Moioli and Araujo acknowledge the support from the National Institute of Science and Technology, program Brain Machine Interface (INCT INCEMAQ) of the National Council for Scientific and Technological Development(CNPq/MCTI), the Rio Grande do Norte Research Foundation (FAPERN), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Brazilian Innovation Agency (FINEP), and the Ministry of Education (MEC). Romano was the recipient of a master's scholarship from FAPESP, grant 2018/11075-5. Elias is funded by a CNPq Research Productivity Grant (312442/2017-3). This research was carried out using the computational resources from the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Robotics Lab within the Edinburgh Centre for Robotics, and by the Nvidia Grants program.

References

[1] Samuel Armstrong, Martin V. Sale, and Ross Cunnington. Neural oscillations and the initiation of voluntary movement. *Frontiers in Psychology*, 9:2509, 2018.

[2] Daniel Ashlock. *Evolutionary computation for modeling and optimization*. Springer Science & Business Media, 2006.

[3] Paul B Badcock, Karl J Friston, and Maxwell JD Ramstead. The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of life Reviews*, 31:104–121, 2019.

[4] Jyotika Bahuguna, Tom Tetzlaff, Arvind Kumar, Jeanette Hellgren Kotaleski, and Abigail Morrison. Homologous basal ganglia network models in physiological and Parkinsonian conditions. *Frontiers in Computational Neuroscience*, 11(August):1–21, 2017.

[5] Julius S. Bendat and Allan G. Piersol. *Random Data: Analysis and Measurement Procedures*. John Wiley and Sons, Inc., USA, 4th edition, 2010.

[6] Liora Benhamou, Maya Bronfeld, Izhar Bar-Gad, and Dana Cohen. Globus Pallidus External Segment Neuron Classification in Freely Moving Rats: A Comparison to Primates. *PLoS ONE*, 7(9), 2012.

[7] Sarah F. Beul, Helen Barbas, and Claus C. Hilgetag. A Predictive Structural Model of the Primate Connectome. *Scientific Reports*, 7:1–30, 2017.

[8] J. P. Bolam, J. J. Hanley, P. A. C. Booth, and M. D. Bevan. Synaptic organisation of the basal ganglia. *Journal of anatomy*, 196:527–542, 2000.

[9] Peter Brown and David Williams. Basal ganglia local field potential activity: character and functional significance in the human. *Clinical neurophysiology*, 116(11):2510–2519, 2005.

[10] S. A. Budoff, J. F. Rodrigues Neto, V. Arboés, M. S. L. Nascimento, C. B. Kunicki, and M. F. P. Araújo. Stereotaxic surgery for implantation of microelectrode arrays in the common marmoset (callithrix jacchus). *J. Vis. Exp.*, 151(e60240), 2019.

[11] György Buzsáki. Neural Syntax: Cell Assemblies, Synapsembles, and Readers. *Neuron*, 68(3):362–385, 2010.

[12] György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes. *Nature reviews. Neuroscience*, 13(6):407–20, 5 2012.

[13] Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, Krishna V. Shenoy, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[14] John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

[15] D. Cyranoski. Marmoset model takes centre stage. *Nature*, 459(492), 2009.

[16] Ted M. Dawson, Todd E. Golde, and Clotilde Lagier-Tourenne. Animal models of neurodegenerative diseases. *Nature Neuroscience*, 21(10):1370–1379, 2018.

[17] G. Du, P. Zhuang, M. Hallett, Y. Q. Zhang, J. Y. Li, and Y. J. Li. Properties of oscillatory neuronal activity in the basal ganglia and thalamus in patients with Parkinson's disease. *Translational Neurodegeneration*, 7(1):1–13, 2018.

[18] Salvador Dura-Bernal, Benjamin A Suter, Padraig Gleeson, Matteo Cantarelli, Adrian Quintana, Facundo Rodriguez, David J Kedziora, George L Chadderdon, Cliff C Kerr, Samuel A Neymotin, Robert A McDougal, Michael Hines, Gordon MG Shepherd, and William W Lytton. Netpyne, a tool for data-driven multiscale modeling of brain circuits. *eLife*, 8:e44494, apr 2019.

[19] Andreas K. Engel, Pascal Fries, and Wolf Singer. Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, 2(10):704–716, 2001.

[20] Pascal Fries. Rhythms for Cognition: Communication through Coherence. *Neuron*, 88(1):220–235, 2015.

[21] Adriana Galvan, Annaelle Devergnas, and Thomas Wichmann. Alterations in neuronal activity in basal ganglia-thalamocortical circuits in the parkinsonian state. *Frontiers in Neuroanatomy*, 9(February):5, 2015.

[22] Adriana Galvan and Thomas Wichmann. Pathophysiology of parkinsonism. *Clinical neurophysiology*, 119(7):1459–1474, 2008.

[23] Carl Gold, Darrell A Henze, Christof Koch, and Gyorgy Buzsaki. On the origin of the extracellular action potential waveform: a modeling study. *Journal of neurophysiology*, 95(5):3113–3128, 2006.

[24] Kevin Gurney, Tony J. Prescott, Jeffery R. Wickens, and Peter Redgrave. Computational models of the basal ganglia: from robots to membranes. *Trends in Neurosciences*, 27(8):453 – 459, 2004.

[25] Jorge Gámez, Germán Mendoza, Luis Prado, Abraham Betancourt, and Hugo Merchant. The amplitude in periodic neural state trajectories underlies the tempo of rhythmic tapping. *PLOS Biology*, 17(4):1–32, 04 2019.

[26] P. Halje, Ivani Brys, Juan J. Mariman, Claudio Da Cunha, Romulo Fuentes, and Per Petersson. Oscillations in cortico-basal ganglia circuits: Implications for parkinson's disease and other neurologic and psychiatric conditions. *Journal of Neurophysiology*, 122(1):203–231, 2019.

[27] Craig Denis Hardman, Jasmine Monica Henderson, David Isaac Finkelstein, Malcolm Kenneth Horne, George Paxinos, and Glenda Margaret Halliday. Comparison of the basal ganglia in rats, marmosets, macaques, baboons, and humans: Volume and neuronal number for the output, internal relay, and striatal modulating nuclei. *Journal of Comparative Neurology*, 445(3):238–255, 2002.

[28] Michael Hines, Andrew Davison, and Eilif Muller. Neuron and python. *Frontiers in Neuroinformatics*, 3:1, 2009.

[29] Allan L Hodgkin and Andrew F Huxley. Currents carried by sodium and potassium ions through the membrane of the giant axon of loligo. *The Journal of physiology*, 116(4):449–472, 1952.

[30] A. J. N. Holgado, J. R. Terry, and R. Bogacz. Conditions for the Generation of Beta Oscillations in the Subthalamic Nucleus-Globus Pallidus Network. *Journal of Neuroscience*, 30(37):12340–12352, 2010.

[31] M. D. Humphries, R. D. Stewart, and K. N. Gurney. A Physiologically Plausible Model of Action Selection and Oscillatory Activity in the Basal Ganglia. *Journal of Neuroscience*, 26(50):12921–12942, 2006.

[32] Mark Humphries and Kevin Gurney. Network effects of subthalamic deep brain stimulation drive a unique mixture of responses in basal ganglia output. *The European journal of neuroscience*, 36:2240–51, 07 2012.

[33] Mark D. Humphries, Jose Angel Obeso, and Jakob Kisbye Dreyer. Insights into Parkinson's disease from computational models of the basal ganglia. *Journal of Neurology, Neurosurgery and Psychiatry*, 40, 2018.

[34] E. M. Izhikevich. Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, 14(6):1569–1572, 2003.

[35] Geetika Kharkwal, Karen Brami-Cherrier, José E. Lizardi-Ortiz, Alexandra B. Nelson, Maria Ramos, Daniel Del Barrio, David Sulzer, Anatol C. Kreitzer, and Emiliana Borrelli. Parkinsonism driven by antipsychotics originates from dopaminergic control of striatal cholinergic interneurons. *Neuron*, 91(1):67–78, 2016.

[36] Noriyuki Kishi, Kenya Sato, Erika Sasaki, and Hideyuki Okano. Common marmoset as a new model animal for neuroscience research and genome editing technology. *Development Growth and Differentiation*, 56(1):53–62, 2014.

[37] Lucas A. Koelman and Madeleine M. Lowery. Beta-band resonance and intrinsic oscillations in a biophysically detailed model of the subthalamic nucleus-globus pallidus network. *Frontiers in Computational Neuroscience*, 13:77, 2019.

[38] James B. Koprich, Lorraine V. Kalia, and Jonathan M. Brotchie. Animal models of α-synucleinopathy for Parkinson disease drug development. *Nature Reviews Neuroscience*, 18(9):515–529, 8 2017.

[39] Jeffrey L. Krichmar. Neurorobotics—a thriving community and a promising pathway toward intelligent cognitive robots. *Frontiers in Neurorobotics*, 12:42, 2018.

[40] Karthik Kumaravelu, David T. Brocker, and Warren M. Grill. A biophysical model of the cortex-basal ganglia-thalamus network in the 6-OHDA lesioned rat model of Parkinson's disease. *Journal of Computational Neuroscience*, 40(2):207–229, 4 2016.

[41] Peter Lakatos, Joachim Gross, and Gregor Thut. A New Unifying Account of the Roles of Neuronal Entrainment. *Current Biology*, 29(18):R890–R905, 2019.

[42] José Lanciego, Natasha Luquin, and José Obeso. Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine*, 2, 10 2012.

[43] Xiaoyu Li, Ping Zhuang, and Yongjie Li. Altered neuronal firing pattern of the basal ganglia nucleus plays a role in levodopa-induced dyskinesia in patients with parkinson's disease. *Frontiers in Human Neuroscience*, 9:630, 2015.

[44] Jean Liénard and Benoît Girard. A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection. *Journal of Computational Neuroscience*, 36(3):445–468, 2014.

[45] Mikael Lindahl and Jeanette Hellgren Kotaleski. Untangling basal ganglia network dynamics and function: Role of dopamine depletion and inhibition investigated in a spiking network model. *eNeuro*, 3(6), 2016.

[46] M. Lu, X. Wei, Y. Che, J. Wang, and K. A. Loparo. Application of reinforcement learning to deep brain stimulation in a computational model of parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1):339–349, 2020.

[47] Mario Manto, Donna L. Gruol, Jeremy D. Schmahmann, Noriyuki Koibuchi, and Ferdinando Rossi. Handbook of the cerebellum and cerebellar disorders. *Handbook of the Cerebellum and Cerebellar Disorders*, pages 1–2424, 2013.

[48] Daniel H. Mathalon and Vikaas S. Sohal. Neural oscillations and synchrony in brain dysfunction and neuropsychiatric disorders it's about time. *JAMA Psychiatry*, 72(8):840–844, 2015.

[49] Matthew M. McGregor and Alexandra B. Nelson. Circuit Mechanisms of Parkinson's Disease, 3 2019.

[50] Robert Merrison-Hort, Nada Yousif, Andrea Ferrario, and Roman Borisyuk. *Oscillatory Neural Models of the Basal Ganglia for Action Selection in Healthy and Parkinsonian Cases*, pages 149–189. Springer International Publishing, Cham, 2017.

[51] Cory T. Miller, Winrich A. Freiwald, David A. Leopold, Jude F. Mitchell, Afonso C. Silva, and Xiaoqin Wang. Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron*, 90(2):219–233, 2016.

[52] Jonathan W. Mink. Basal ganglia mechanisms in action selection, plasticity, and dystonia. *European Journal of Paediatric Neurology*, 22(2):225 – 229, 2018. Movement Disorders.

[53] Jude F. Mitchell and David A. Leopold. The marmoset monkey as a model for visual neuroscience. *Neuroscience Research*, 93:20 – 46, 2015. Marmoset Neuroscience.

[54] Jan Morén, Jun Igarashi, Osamu Shouno, Junichiro Yoshimoto, and Kenji Doya. Dynamics of basal ganglia and thalamus in parkinsonian tremor. In *Multiscale Models of Brain Disorders*, pages 13–20. Springer, 2019.

[55] Vignayanandam Ravindernath Muddapu, Alekhya Mandali, V. Srinivasa Chakravarthy, and Srikanth Ramaswamy. A computational model of loss of dopaminergic cells in parkinson's disease due to glutamate-induced excitotoxicity. *Frontiers in Neural Circuits*, 13:11, 2019.

[56] Simon Musall, Anne E Urai, David Sussillo, and Anne K Churchland. Harnessing behavioral diversity to understand neural computations for cognition. *Current Opinion in Neurobiology*, 58:229 – 238, 2019. Computational Neuroscience.

[57] Meinard Müller. *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[58] Atsushi Nambu, Hironobu Tokuno, and Masahiko Takada. Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research*, 43(2):111–117, 2002.

[59] Eva M Navarro-Lopez, Utku Çelikok, and Neslihan S Şengör. A dynamical model for the basal ganglia-thalamo-cortical oscillatory activity and its implications in parkinson's disease. *Cognitive Neurodynamics*, pages 1–28, 2020.

[60] Christian Nowke, Sandra Diaz-Pier, Benjamin Weyers, Bernd Hentschel, Abigail Morrison, Torsten W. Kuhlen, and Alexander Peyser. Toward rigorous parameterization of underconstrained neural network models through interactive visualization and steering of connectivity generation. *Frontiers in Neuroinformatics*, 12:32, 2018.

[61] Jose A. Obeso, Concepcio Marin, C. Rodriguez-Oroz, Javier Blesa, B. Benitez-Temiño, Juan Mena-Segovia, Manuel Rodríguez, and C. Warren Olanow. The basal ganglia in Parkinson's disease: Current concepts and unexplained observations. *Annals of Neurology*, 64(S2):S30–S46, 1 2009.

[62] Harilal Parasuram, Bipin Nair, Egidio D'Angelo, Michael Hines, Giovanni Naldi, and Shyam Diwakar. Computational modeling of single neuron extracellular electric potentials and network local field potentials using lfpsim. *Frontiers in Computational Neuroscience*, 10:65, 2016.

[63] Alex Pavlides, S. John Hogan, and Rafal Bogacz. Computational models describing possible mechanisms for generation of excessive beta oscillations in parkinson's disease. *PLOS Computational Biology*, 11:1–29, 12 2015.

[64] George Paxinos, Charles Watson, Michael Petrides, Marcello Rosa, and Hironobu Tokuno. *The marmoset brain in stereotaxic coordinates*. Academic Press, 2012.

[65] Rodrigo Pena, Cesar Ceballos, Júnia de Deus, Antonio Roque, Norberto Garcia-Cairasco, Ricardo Leao, and Alexandra Cunha. Modeling hippocampal ca1 gabaergic synapses of audiogenic rats. *International Journal of Neural Systems*, 30:2050022, 03 2020.

[66] Rolf Pfeifer and Josh C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books)*. The MIT Press, 2006.

[67] Werner Poewe, Klaus Seppi, Caroline M. Tanner, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette Eleonore Schrag, and Anthony E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3:1–21, 2017.

[68] Tony Prescott, Fernando González, Kevin Gurney, Mark Humphries, and Peter Redgrave. A robot model of the basal ganglia: Behavior and intrinsic processing. *Neural networks : the official journal of the International Neural Network Society*, 19:31–61, 02 2006.

[69] Tony J. Prescott, Fernando M. Montes González, Kevin Gurney, Mark D. Humphries, and Peter Redgrave. A robot model of the basal ganglia: Behavior and intrinsic processing. *Neural Networks*, 19(1):31–61, 1 2006.

[70] Savva Pronin, Liam Wellacott, Jhielson M Pimentel, Renan C Moioli, and Patricia A Vargas. Neurorobotic Models of Neurological Disorders: A Mini Review. *Frontiers in Neurorobotics*, 15:26, 2021.

[71] Kenneth Price Rainer Storn. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11:341–359, 1997.

[72] C. M. Ranieri, R. C. Moioli, R. A. F. Romero, M. F. P. de Araújo, M. B. De Santana, J. M. Pimentel, and P. A. Vargas. Unveiling parkinson's disease features from a primate model with deep neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[73] K. Deergha Rao, M.N.S. Swamy, K. Deergha Rao, and M.N.S. Swamy. Spectral analysis of signals. In *Digital Signal Processing*, pages 721–751. Springer Singapore, 2018.

[74] Peter Redgrave, Manuel Rodriguez, Yoland Smith, Maria C. Rodriguez-Oroz, Stephane Lehericy, Hagai Bergman, Yves Agid, Mahlon R. Delong, and Jose A. Obeso. Goal-directed and habitual control in the basal ganglia: Implications for Parkinson's disease. *Nature Reviews Neuroscience*, 11(11):760–772, 2010.

[75] M. R. Romano, R. C. Moioli, and L. A. Elias. Evaluation of frequency-dependent effects of deep brain stimulation in a cortex-basal ganglia-thalamus network model of parkinson's disease*. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 3638–3641, 2020.

[76] J.R. Rosenberg, A.M. Amjad, P. Breeze, D.R. Brillinger, and D.M. Halliday. The fourier approach to the identification of functional coupling between neuronal spike trains. *Progress in Biophysics and Molecular Biology*, 53(1):1 – 31, 1989.

[77] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[78] Jonathan Rubin and David Terman. High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model. *Journal of computational neuroscience*, 16:211–35, 05 2004.

[79] Jonathan E Rubin, Cameron C McIntyre, Robert S Turner, and Thomas Wichmann. Basal ganglia activity patterns in parkinsonism and computational modeling of their downstream effects. *European Journal of Neuroscience*, 36(2):2213–2228, 2012.

[80] Abigail A. Russo, Ramin Khajeh, Sean R. Bittner, Sean M. Perkins, John P. Cunningham, Laurence F. Abbott, and Mark M. Churchland. Neural trajectories in the supplementary motor area and primary motor cortex exhibit distinct geometries, compatible with different classes of computation. *bioRxiv*, 2019.

[81] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[82] Terence D. Sanger. A computational model of deep-brain stimulation for acquired dystonia in children. *Frontiers in Computational Neuroscience*, 12:77, 2018.

[83] M. Santana, T. Palmér, H. Simplício, R. Fuentes, and P. Petersson. Characterization of long-term motor deficits in the 6-OHDA model of Parkinson's disease in the common marmoset. *Behavioural Brain Research*, 290:90–101, 2015.

[84] Maxwell B. Santana, Pär Halje, Hougelle Simplício, Ulrike Richter, Marco Aurelio M. Freire, Per Petersson, Romulo Fuentes, and Miguel A.L. Nicolelis. Spinal cord stimulation alleviates motor deficits in a primate model of Parkinson Disease. *Neuron*, 84(4):716–722, 11 2014.

[85] Luis Santos, Javier Fernandez-Rio, Kristian Winge, Beatriz Barragán-Pérez, Vicente Rodríguez-Pérez, Vicente González-Díez, Miguel Blanco-Traba, Oscar E. Suman, Charles Philip Gabel, and Javier Rodríguez-Gómez. Effects of supervised slackline training on postural instability, freezing of gait, and falls efficacy in people with parkinson's disease. *Disability and Rehabilitation*, 39(16):1573–1580, 2017. PMID: 27416005.

[86] Henning Schroll and Fred Hamker. Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy. *Frontiers in Systems Neuroscience*, 7:122, 2013.

[87] A. Sharott, C. K. E. Moll, G. Engler, M. Denker, S. Grun, and A. K. Engel. Different Subtypes of Striatal Neurons Are Selectively Modulated by Cortical Oscillations. *Journal of Neuroscience*, 29(14):4571–4585, 2009.

[88] Osamu Shouno, Yoshihisa Tachibana, Atsushi Nambu, and Kenji Doya. Computational model of recurrent subthalamo-pallidal circuit for generation of parkinsonian oscillations. *Frontiers in neuroanatomy*, 11:21, 2017.

[89] J. Shulman, P. De Jager, and M. Feany. Parkinson's disease: Genetics and pathogenesis. *Parkinson's Disease: Genetics and Pathogenesis*, pages 1–386, 2011.

[90] W. Singer. Neuronal Synchrony: A Versatile Code for the Definition of Relations? *Neuron*, 24:49–65, 1999.

[91] Anke H. Snijders and Bastiaan R. Bloem. Cycling for freezing of gait. *New England Journal of Medicine*, 362(13):e46, 2010. PMID: 20357278.

[92] H. Stephan, G. Baron, and W. K. Schwerdtfeger. *The Brain of the Common Marmoset (Callithrix jacchus)*. Springer-Verlag Berlin Heidelberg, 1980.

[93] D. James Surmeier, Steven M. Graves, and Weixing Shen. Dopaminergic modulation of striatal networks in health and Parkinson's disease. *Current Opinion in Neurobiology*, 29:109–117, 2014.

[94] Shreyas M. Suryanarayana, Jeanette Hellgren Kotaleski, Sten Grillner, and Kevin N. Gurney. Roles for globus pallidus externa revealed in a computational model of action selection in the basal ganglia. *Neural Networks*, 109:113–136, 2019.

[95] David Sussillo. Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology*, 25:156–163, 2014.

[96] D Terman, J E Rubin, A C Yew, and C J Wilson. Activity Patterns in a Model for the Subthalamopallidal Network of the Basal Ganglia. *The Journal of Neuroscience*, 22(7):1–14, 2002.

[97] Gerd Tinkhauser, Alek Pogosyan, Huiling Tan, Damian M. Herz, Andrea A. Kühn, and Peter Brown. Beta burst dynamics in Parkinson's disease off and on dopaminergic medication. *Brain*, 140(11):2968–2981, 11 2017.

[98] Meropi Topalidou, Daisuke Kase, Thomas Boraud, and Nicolas P. Rougier. A computational model of dual competition between the basal ganglia and the cortex. *eNeuro*, 5(6), 2018.

[99] Adriano BL Tort, Maximilian Hammer, Jiaojiao Zhang, Jurij Brankačk, and Andreas Draguhn. Causal relations between cortical network oscillations and breathing frequency. *bioRxiv*, 2020.

[100] Peter J. Uhlhaas and Wolf Singer. Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, 11(2):100–113, 2010.

[101] S. J. van Albada and P. A. Robinson. Mean-field modeling of the basal ganglia-thalamocortical system. I. Firing rates in healthy and parkinsonian states. *Journal of Theoretical Biology*, 257(4):642–663, 2009.

[102] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, and Nidhi Gupta. A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications (IJERA)*, 2(3):1379–1384, 2012.

[103] Doris D. Wang, Coralie de Hemptinne, Svjetlana Miocinovic, Jill L. Ostrem, Nicholas B. Galifianakis, Marta San Luciano, and Philip A. Starr. Pallidal deep-brain stimulation disrupts pallidal beta oscillations and coherence with primary motor cortex in parkinson's disease. *Journal of Neuroscience*, 38(19):4556–4568, 2018.

[104] Doris D Wang, Coralie de Hemptinne, Svjetlana Miocinovic, Salman E Qasim, Andrew M Miller, Jill L Ostrem, Nicholas B Galifianakis, Marta San Luciano, and Philip A Starr. Subthalamic local field potentials in parkinson's disease and isolated dystonia: an evaluation of potential biomarkers. *Neurobiology of disease*, 89:213–222, 2016.

[105] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[106] Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. PMID: 19357332.

CHAPTER

5

APPLICATION SCENARIO

In this chapter, a robot simulation is presented, which integrates the activity recognition framework with heuristics-based and neurorobotics approaches for behaviour selection. The activity recognition module, presented in Chapter 2 and Chapter 3, was employed to feed a behaviour selection mechanism of a simulated robot, which resembles the LARa robot (RANIERI *et al.*, 2018), in a home environment. A mechanism based on simple heuristics was implemented, and compared to a more sophisticated architecture based on the neurophysiological advances presented in Chapter 4 (i.e., the neurorobotics approach). This chapter presents the third research question of section 1.1. A paper on these developments was written, in which the robot simulation and the strategies for behaviour selection are presented in detail, along with the results. This paper, attached to the following pages, was uploaded to arXiv.

Ranieri, C. M.; Moioli, R. C.; Vargas, P. A.; Romero, R. A. F. "A Neurorobotics Approach to Behaviour Selection based on Human Activity Recognition". *arXiv preprint arXiv:2107.12540*, 2021.

Contribution statement: all authors contributed to the design of the experiments. Ranieri provided the specific methods and implementations, performed the experiments and analysed the results. Moioli, Vargas and Romero revised the methods and results presented, contributing to the discussion. Ranieri written the draft of the paper, revised by the other authors.

A NEUROROBOTICS APPROACH TO BEHAVIOUR SELECTION BASED ON HUMAN ACTIVITY RECOGNITION

Caetano M. Ranieri
Institute of Mathematical and Computer Sciences
University of Sao Paulo
Sao Carlos, SP, Brazil
cmranieri@alumni.usp.br

Renan C. Moioli
Digital Metropolis Institute
Federal University of Rio Grande do Norte
Natal, RN, Brazil
renan.moioli@imd.ufrn.br

Patricia A. Vargas
Edinburgh Centre for Robotics
Heriot-Watt University
Edinburgh, Scotland, UK
p.a.vargas@hw.ac.uk

Roseli A. F. Romero
Institute of Mathematical and Computer Sciences
University of Sao Paulo
Sao Carlos, SP, Brazil
rafrance@icmc.usp.br

July 28, 2021

ABSTRACT

Behaviour selection has been an active research topic for robotics, in particular in the field of human-robot interaction. For a robot to interact effectively and autonomously with humans, the coupling between techniques for human activity recognition, based on sensing information, and robot behaviour selection, based on decision-making mechanisms, is of paramount importance. However, most approaches to date consist of deterministic associations between the recognised activities and the robot behaviours, neglecting the uncertainty inherent to sequential predictions in real-time applications. In this paper, we address this gap by presenting a neurorobotics approach based on computational models that resemble neurophysiological aspects of living beings. This neurorobotics approach was compared to a non-bioinspired, heuristics-based approach. To evaluate both approaches, a robot simulation is developed, in which a mobile robot has to accomplish tasks according to the activity being performed by the inhabitant of an intelligent home. The outcomes of each approach were evaluated according to the number of correct outcomes provided by the robot. Results revealed that the neurorobotics approach is advantageous, especially considering the computational models based on more complex animals.

***K*eywords** Behaviour selection · Human activity recognition · Robot simulation · Neurorobotics · Bioinspired computational model

1 Introduction

Truly autonomous behaviour is still not the norm for robots designed to interact socially with humans [6]. In general, behaviour selection has been an active research topic for robotics in general, and human-robot interaction in particular [19]. In this context, the need for a real-time understanding of human actions is of paramount importance for the robotic agent to behave proactively and effectively. Such a requirement could be achieved with techniques for human activity recognition [32].

When dealing with complex modalities (e.g., videos or data from inertial units), activity recognition approaches often rely on machine learning. For instance, video-based activity recognition have been approached by architectures based on convolutional and recurrent neural networks [15, 29]. For inertial data, similar architectures have been proposed, processing either raw data [35, 10] or descriptors obtained through feature extraction methods [47, 1]. To provide a

wider range of possibilities, robots may act symbiotically with other pervasive devices, such as wearable technologies or ambient sensors in intelligent environments, which may provide additional capabilities for sensing and acting based on application-specific components [2]. When synchronised data from different sensors are available, activity recognition techniques may rely on multiple sensor modalities to provide more accurate results, giving rise to techniques for multimodal activity recognition [28, 17, 43].

Although human activity recognition has been a quite fertile field of research, few approaches have been developed to link the outputs from those algorithms into actual response behaviours from a robot. Related works usually consist of direct associations between the recognised activities and the response behaviours [11, 25, 44]. One of the possibilities consist of combining computational neuroscience to the robotics scenarios, characterising the field of neurorobotics [51], which may build upon different biological aspects that influence the behaviour of living beings.

Li *et al.* [24] provided a comprehensive survey on neurorobotics systems (NRS) and the different components that may integrate them. According to the authors, a generalised framework can be depicted for most NRSs in the literature, composed of a *simulated brain*, which is fed with sensory signals from a *body* and turns them into control signals for a *hierarchical controller*, responsible for decoding these signals into control commands for the body, which actuates and senses an external environment. Bioinspired strategies may be introduced to different aspects of the framework, according to the required task of a particular study.

The basal ganglia, a group of subcortical nuclei present in the vertebrate's brain, is known to have an important role in action selection mechanisms, especially regarding striatal circuits [30]. The so-called direct and indirect pathways are characterised by competitive or complementary functions that mediate the excitation of the motor system based on inputs from the motivational system of an individual, deciding whether to "go" or to "stop" performing a certain behaviour [4]. The potential roles of such a mechanism in robotic frameworks have also been evaluated, including simulations in which bioinspired networks receiving different stimuli are expected to respond with different behaviours, resulting in cooperative interactions that produce robot behaviours [3].

In this paper, we present a neurorobotics model which embeds computational models of the basal ganglia-thalamus-cortex (BG-T-C) circuit [22, 40] to provide a decision-making mechanism for a robot - in this context, we may call it a *neurorobot*. The neurorobotics approach has been proposed for enhancing the decision-making mechanism, as suggested by related researches in neurorobotics [26, 33]. It consisted of simulating neurophysiological aspects within a cognitive framework, in which different stimuli was introduced to certain brain structures within the circuit, according to real-time outputs of the activity recognition module. The resulting spike trains from the neurorobotics model were then converted to neural firing frequencies across brain regions, which would be further decoded using convolutional neural networks, in order to infer the most suitable response behaviour for the robot.

The application scenario is a simulated smart home, in which an activity recognition model, presented in [41] for the HWU-USP activities dataset [38], was employed in human-robot interaction tasks, using a mobile robot. In summary, the robot needs to produce response behaviours according to the contextual information inferred by the user (i.e., the recognised activity).

The neurorobotics approach, which is the central contribution of this work, was compared to a heuristics approach, in which a deterministic behaviour selection mechanism was considered using simple heuristics that associate recognised activities to robot behaviours. This neurorobotics approach embedded two computational models, one that resembled neurophysiological data of rodents (i.e., the rat model), and one that resembled data from marmoset monkeys (i.e., the primate model).

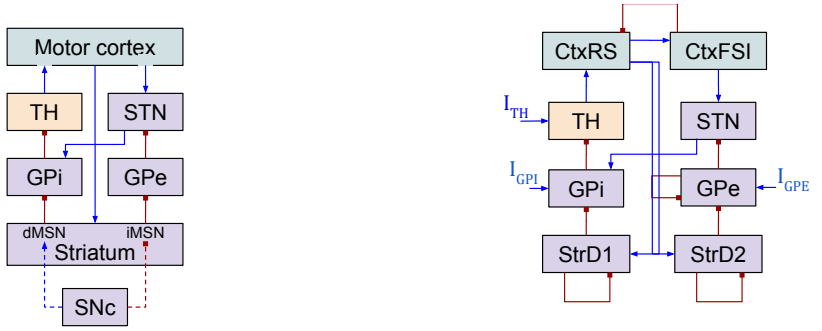
The different factors considered for this study were evaluated according to the relative number of correct outcomes of the robot simulation. Considering the activity recognition framework, the results have confirmed that more accurate classifiers for the activity recognition module led to a greater number of robot tasks successfully completed. Although the performances of the heuristic and neurorobotics approaches varied according to the computational model embedded, the study confirmed that the most complex neurorobotics model (i.e., the marmoset-based model of the BG-T-C circuit) led to an increased performance in relation to the heuristic approaches when a more accurate activity recogniser was considered (i.e., the video-based classifier).

The remainder of this paper is organised as follows. The brain structures considered for this neurorobotics approach and the computational modelling adopted are presented in Section 2. In Section 3, are presented the general aspects of the robotic system, and the integration between each of its modules. In Section 4, the neurorobotics approach is detailed. In Section 5, the methods and implementations are depicted. The corresponding results are presented in Section 6 and discussed in Section 7. Finally, the concluding remarks and directions for future research are provided in Section 8.

2 The BG-T-C Circuit and Original Computational Models

In this section, we present the basic concepts on the brain structures present in the basal ganglia-thalamus-cortex (BG-T-C) circuit, and the original computational modelling. The BG-T-C circuit, illustrated in Figure 1, is formed by the *motor cortex* (M1), the *thalamus* (TH), and the *basal ganglia* (BG), the latter composed of a subset of structures: the *striatum* (Str), the *globus pallidus*, divided into *pars interna* (GPi) and *pars externa* (GPe), the *subthalamic nucleus* (STN), and the *substantia nigra*, divided into *pars compacta* (SNc) and *pars reticulata* (SNr).

In [31], is provided a discussion about the mechanisms of this circuit and presented models to describe it. The most useful model to explain the connections within this circuit, especially those affected by PD, is the so-called classic model, illustrated in Figure 1a.



(a) Classic model, as described by [31].

(b) Computational model, as designed by [22] for rodent data, and adapted by [40] for primate data.

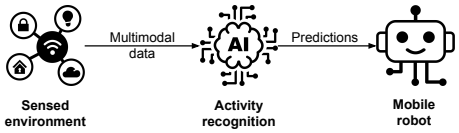
Figure 1: Schematic representations of the classic and computational models of the BG-T-C circuit. In the connections, excitatory synapses are shown as blue arrows, and inhibitory synapses, as red squares.

The pathways start with an excitatory connection from the cortex to the striatum, which projects its output neurons, named *medium spiny neurons* (MSN), to other structures inside the BG. In the direct pathway, the direct MSN (dMSN) inhibits the GPi, which reduces its inhibition to the TH. Then, it excites the motor cortex. In the indirect pathway, the indirect MSN (iMSN) inhibits the GPe, which reduces its inhibition to the STN, which excites the GPi. Thus, this results on inhibition of the TH and absence of excitatory outputs to the motor cortex. In other words, the direct pathway excites the cortex (i.e., positive feedback loop), while the indirect pathway inhibits it (i.e., negative feedback loop).

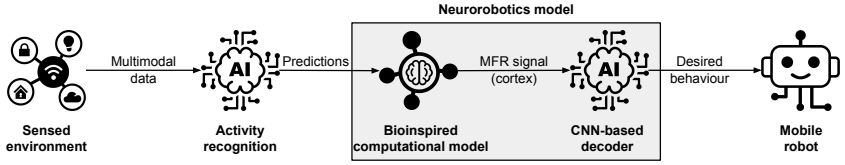
In [22], a computational model of the BG-C-T circuit, originaly developed to study the underlying mechanisms of Parkinson's Disease (PD), was proposed and implemented based on neural data from healthy and PD-induced (i.e., 6-OHDA lesioned) rats [18]. Eight brain structures were modelled and connected based on a simplified version of the classic model (see Figure 1b). In particular, the direct and indirect pathways were modelled separately representing the MSN modulation by D1 and D2 dopamine receptors in the striatum (i.e., StrD1 and StrD2, respectively). The cortex is represented by regular spiking (RS) excitatory neurons and fast spiking (FSI) inhibitory interneurons (i.e., CtxRS and CtxFSI, respectively). A bias current was added in the TH, GPe, and GPi, accounting for the inputs not explicitly modelled. This model was designed with the ability to shift from the simulation of healthy to the PD status, which is done by altering certain conductances.

Although all mammals have a similar set of BG structures that are similarly connected with thalamic and cortical structures, subtle differences between species may be found, with primates being more similar to humans than rodents [27, 21, 7]. A data-driven approach was proposed in [40] to obtain a primate-based computational model of the BG-T-C circuit and the mechanisms of PD. The resulting marmoset computational model was evaluated based on the differences between healthy and PD individuals, with respect to the spectral signature of the brain activity [49], the dynamics of the firing rates of neurons across brain regions [50], and the coherence between spike trains [14].

The implementation used in [40] built on a Python translation of the original computational model of [22], originally made by [46] using the NetPyNE framework and the libraries from the NEURON simulator [8]. Based on the results of the machine learning framework, a practical setup of either the rat or marmoset computational models was made available. The adaptations performed in this work to the original computational models (see Subsection 4.1) were based on the code made available by the authors. For all neurorobotics model evaluations, we considered both the rat and primate computational models, always with the healthy state set on.



(a) Heuristics approach: the predictions from the activity recognition module are fed directly to the mobile robot.



(b) Neurorobotics approach: the predictions are used as stimuli for the embedded, bioinspired computational model of the BG-T-C circuit, which simulates neural activity that is further interpreted by a CNN-based decoder, responsible for deciding the behaviour to be performed by the robot. Both the bioinspired computational model and the CNN-based decoder compose the neurorobotics model presented in this research.

Figure 2: Interaction between modules for the application scenario proposed.

3 Integrated System

The modules of the application scenario, and the interactions between them, are illustrated in Figure 2. In this scenario, the human activities are inferred by a machine learning algorithm, and the supporting behaviours are performed by a mobile robot placed in a simulated environment, composing an ambient assisted living (AAL) application [5].

The general information flow was: given the multimodal data provided by a set of sensors within a *sensed environment*, apply an *activity recognition module* to classify such data into a set of predefined human activities, and produce correspondent response behaviours for a *mobile robot*.

The neurorobotics approach was compared to a heuristics approach. The heuristics approach (Figure 2a) consisted of associating the predictions of the activity recognition module to response behaviours based on simple heuristics, presented in Subsection 5.4. In the neurorobotics approach (Figure 2b), the predictions from the activity recognition module were employed to stimulate a *bioinspired computational model*, whose outputs (i.e., neural firing frequencies of brain simulated regions) were decoded by a *CNN-based decoder*, which provided the decisions for the mobile robot.

More specifically, a *sensed environment* consisted of a previously collected dataset [38], composed by a set of recording sessions X, with each $x \in X$ associated to an activity $a \in A = \{a_1, \ldots, a_{N_A}\}$, where N_A is the number of classes (i.e., labels) considered for this dataset. The function describing these associations is given by Equation 1.

$$f_A : X \to A \iff f_A(x) = a, \quad x \in X, \quad a \in A \tag{1}$$

Each data tuple $x(t)$ comprises a segment, with a previously defined length, of a recording session x starting at timestep $t \in T_x = \{1, \ldots, N_{T_X}\}$, equally spaced among them, to be segmented from x. The *activity recognition* module is a machine learning classifier $g \in G$, which might associate a recording session $x \in X$ at timestep $t \in T_X$ to an activity $a \in A$, through a prediction vector $y_x^t \in Y$ (see Equation 2). In other words, considering that a is unknown at inference time, the inference model g, learned from labelled samples, provides a prediction vector y_x^t, where $y_x^t(a)$ is the probability that a given input $x(t)$ corresponds to activity a.

$$g : X \times T \to Y \iff g(x,t) = y_x^t, \quad x \in X, \quad t \in T, \quad y_x^t \in Y \tag{2}$$

The application scenario was designed so that each activity in A was associated to a desired response for the mobile robot. We defined a set of response behaviours $B = \{b1, \ldots, b_{N_B}\}$, so that each human activity $a \in A$ can be, but not

necessarily is, associated to a response behaviour of the robot. The "no action" behaviour is denoted as $b_\emptyset$. Hence, the function that associates recognised activities to response behaviours is given by Equation 3.

$$f_B : A \to B \cup \{b_\emptyset\} \iff f_B(a) = b, \quad a \in A, \quad b \in B \cup \{b_\emptyset\} \tag{3}$$

The robot simulation would be considered successfully completed if:

- For an activity a being performed in the environment in a session x, the robot completed an expected response behaviour $b \in B$ before x was finished; or
- No response behaviour was expected (i.e., $f_B(a) = b_\emptyset$) and the robot did not complete any of the behaviours in B.

It is worth to notice that, according to this evaluation policy, besides an accuracy requirement (i.e., the correct behaviour must be given in response to a human activity), there was also a time constraint that must be satisfied (i.e., if required, the response behaviour must be completed while the human is still performing the given activity).

Since, by definition, $f_A(x) = a$ is not known at runtime, and can only be inferred by a classifier $g \in G$ as successive prediction vectors y_x^t are provided, a decision-making mechanism was needed to perform adaptive decisions based on partial, time-localised predictions. To this aim, we proposed the neurorobotics model presented in Section 4, and compared it to a simple heuristics-based approach as described in Section 5.

4 The Neurorobotics Model

The neurorobotics model embeds the bioinspired computational model and the CNN-based decoder (see Figure 2). It consists of simulating and decoding the neurophysiological mechanisms of the basal ganglia-thalamus-cortex (BG-T-C) circuit in mammals (see Section 2), responsible for abilities such as motor control, decision-making, and learning [12, 26, 33]. As already stated in Section 2, both the rat-based [22] and the marmoset-based [40] computational models were evaluated as a decision-making mechanism of the neurorobotics model.

4.1 Bioinspired Computational Model

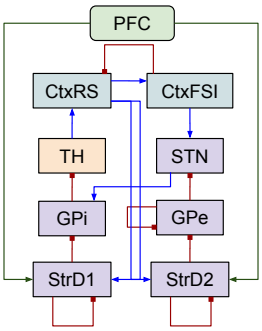
Motivated by the work of [33], two key modifications were introduced to the computational models of the BG-T-C circuit adopted in this work [22, 40]. First, an additional structure, called prefrontal cortex (PFC), was included as a variable source of excitatory stimuli towards the striatum (see Figure 3a). Second, $N_C = N_B$ populations of neurons were implemented as independent channels $c \in C$, each associated to exactly one response behaviour $b \in B$ (see Figure 3b), as defined in Equation 4.

$$f_C : B \to C \iff f_C(c) = b, \quad b \in B, \quad c \in C \tag{4}$$

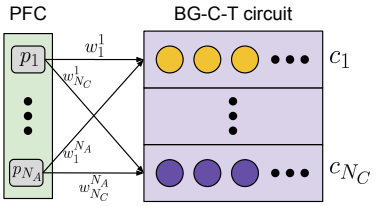
At each timestep, the channels $c \in C$ received a stimulus $s \in S = \{s_1, \ldots, s_{N_C}\}$, whose intensity was based on the linear combination between a prediction vector y_x^t and a weight function f_W, given by Equation 5. The actual value of s is given by function f_S, defined as in Equation 6.

$$f_W : C \times A \to \{0, 1\} \iff f_W(c, a) = \begin{cases} 1, & \text{if } f_B(a) = f_C(c) \\ 0, & \text{otherwise} \end{cases} \quad c \in C, \quad a \in A \tag{5}$$

$$f_S : C \times A \to S \iff f_S(c, a) = s = \sum_{a \in A} f_W(c, a) \cdot y_x^t(a) \quad c \in C, \quad a \in A, \quad s \in S \tag{6}$$



(a) Schematic representation of the computational model as adapted in this work. In the connections, excitatory synapses are shown as blue or green arrows, and inhibitory synapses, as red squares. The blue arrows and red squares correspond to the original synapses as designed by [22] and adapted by [40], while the green arrows are the adaptations provided in this work to allow the stimulation of the circuit in the context of the application scenario proposed.



(b) Predictions from the activity recognition module are interpreted as stimulation originated on the prefrontal cortex (PFC), which selectively stimulates different populations of the computational model, each associated to one response behaviour of the robot.

Figure 3: Adapted version of the computational model of the BG-T-C circuit.

Considering that, as ensured by the softmax activation on the classifiers, $\sum_{a \in A} y_x^t(a) = 1$, then $s \in [0, 1]\mu A/cm^2$, which has shown to be a stable, biologically plausible interval. For a recording sequence x, the set of prediction vectors y_x is employed to update periodically each stimulus $s \in S$, during the course of a corresponding simulation of the computational model (not to be confused with the robot simulation). For each simulation, $N_{T_{\text{sim}}}$ subsequent updates would be done for all $s \in S$, computed for the timesteps in y_x.

A simulation, after finished, produced a spike train for each of the brain regions modelled, contemplating all its length (i.e., all $N_{T_{\text{sim}}}$ updates were considered). The neural firing frequencies were computed according to [23], with the parameters detailed in Subsection 5.3, and summed across each region of each channel, resulting in $N_R \cdot N_C$ output signals for each simulation, each with length L_U, where $N_R = 8$ is the number of regions (see Figure 1b).

Formally, let $u_x^{g,m} \in U$ be defined as the output for a given simulation, where $x \in X$ is a recording session, $g \in G$ is the classifier employed for activity recognition, and $m \in M$, a computational model. Therefore, let a simulation be defined as f_U (see Equation 7), whose output is as a multivariate time-series with $N_R \cdot N_C$ variables and L_U timesteps.

$$\begin{aligned} f_U : X \times G \times M \to U \iff f_U(x, g, m) = u_x^{g,m} \\ (x, g, m) \in X \times G \times M, \quad u \in U \end{aligned} \tag{7}$$

After the simulations were completed, the spike trains at the cortex populations were converted into temporal signals (i.e., neural firing frequencies) based on the mean firing rates across brain regions [23]. The resulting signals were segmented in smaller windows and applied to train and evaluate a convolutional neural network (CNN), which would be employed to determine the decision of the robot at each timestep of the robot simulation (i.e., the *CNN-based decoder*). More details on the implementation of the CNN-based decoder are presented in the next section.

4.2 CNN-Based Decoder

Each simulation of the computational model provided the summed neural firing frequencies of each channel and brain region, generating a data structure $u_x^{g,m}$, associated to the whole recording session that generated it. As a requirement to provide a realistic scenario for the robot simulation, time-localised decisions were required, which must be taken based only in past events. In other words, at a timestep $t_{\text{robot}} = i$ of the robot simulation, only predictions obtained on timesteps $t_j, j < i$ could be taken into account when providing a response behaviour to the robot.

To fulfil this requirement, each instance $u_x^{g,m}$, correspondent to the recording session $x \in X$ in the set of conditions $g \in G$ and $m \in M$ (see Equation 7), was segmented in windows of N_V timesteps, with partial superposition, producing

N_{segs} segments. Considering N_X recording sessions in a given set of conditions, the function f_v would generate a total of $N_X \cdot N_{\text{segs}}$ instances $v \in V$, as defined in Equation 8.

$$\begin{aligned} f_v : U \times T \to V \iff f_v(u_x^{g,m}, t) = v = u_x^{g,m}[t, t + N_V] \\ u_x^{g,m} \in U, \qquad t \in T \quad | \quad t + N_V < L_U, \qquad v \in V \end{aligned} \tag{8}$$

The resulting segments were employed to train a machine learning decoder (i.e., the CNN-based decoder). We considered only the cortex regions to compose the input tuples for the decoder, aiming to preserve biological plausibility regarding this aspect. Given that each channel of the computational model has two cortex regions (i.e., cortex RS and FSI), and that the experiments were performed with N_C channels, associated to the response behaviours $b \in B$, the resulting instances v had shape $N_V \times 2N_C$. The decoder f_Q might be trained to provide a decision vector q_x^t, which corresponds to the probability that a given segment of cortex firing frequencies, given by$v = f_v(u_x^{g,m}, t)$, might be associated to a behaviour in $B \cup \{b_\emptyset\}$. This decoding function may be defined as in Equation 9.

$$\begin{aligned} f_Q : V \to Q \iff f_Q(v) = q_x^t, \\ v \in V, \quad q_x^t \in Q \end{aligned} \tag{9}$$

We have adopted a one-dimensional convolutional neural network (CNN) as decoder, which has shown to provide state-of-the-art results in related work [42] (for the architectural choices and hyperparameter settings, see Subsection 5.3). Classification metrics were provided considering that the categorical output is chosen according to Equation 10, where d_Q corresponds to a response behaviour.

$$d_Q : Q \to B \iff d_Q(q_x^t) = argmax(q_x^t) \tag{10}$$

Finally, the decisions decoded would be fed to the robot simulation and turned into commands, as discussed in Subsection 5.4.

5 Methods

In Figure 4, the different factors assessed in this work, already mentioned, are illustrated. Both the heuristics and the neurorobotics approaches were evaluated with two different models of the activity recognition module: the IMU + ambient and the video-based (see Subsection 5.1). For the heuristics approach, a couple heuristics was considered and compared: the window and the exponential (see Subsection 5.4). For the neurorobotics approach, the rat and marmoset computational models were assessed (see Subsections 4.1 and 4.2).

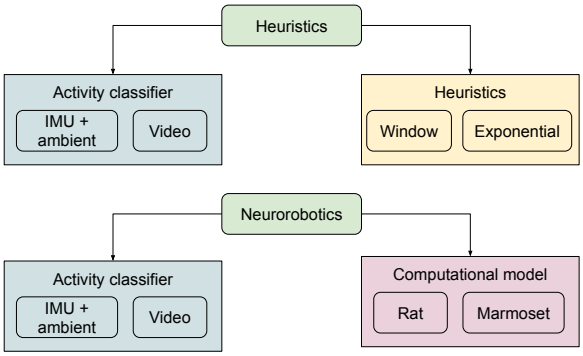


Figure 4: Factors and conditions analysed for the heuristics and neurorobotics approaches. For both approaches, two models for the activity recognition module were considered: the IMU + ambient and the video-based. For the heuristics approach, two approaches were analysed for the decision-making mechanism: window or exponential (see Subsection 5.4). For the neurorobotics approach, two computational models of the BG-T-C circuit were considered: the rat-based and the marmoset-based.

All code was developed in Python language. The machine learning techniques presented for the activity recognition and the CNN-based decoder were implemented with the Tensorflow/Keras framework. The computational models were

implemented using the NetPyNE platform [8]. The robot simulation was implemented in the Gazebo simulator [20] with the Robot Operating System (ROS) [37] as a middleware. The next subsections will provide the implementation details of this work.

5.1 Dataset and Classifiers

We have adopted the HWU-USP activities dataset [38], a multimodal and heterogeneous dataset of human activities recorded in the Robotic Assisted Living Testbed (RALT), at Heriot-Watt University (UK). It is composed by readings of ambient sensors (e.g., switches at wardrobes and drawers, presence detectors, power measurements), inertial units attached to the waist and to the dominant wrist of the subjects, and videos. A set of nine well-defined, pre-segmented activities of daily living was performed by the 16 participants of the data collection. A total of $N_X = 144$ recording sessions were provided, all of them pre-segmented and labelled (i.e., X, A and f_A were provided). The length of the recording sessions varied from less than 25 to over 100 seconds, with high variance either between-classes and between-subjects.

As the activity recognition module, we have employed the framework presented and evaluated in [41]. This was composed by different time-localised classifiers based on artificial neural networks, each focused on a particular modality (i.e., set of similar sensors) or set of modalities. We adopted a couple pre-trained classifiers (i.e., the *IMU + ambient* and the *video*-based classifiers) from the framework to provide the prediction vectors, respecting the between-subjects 8-fold approach for training and evaluating. Let those classifiers be denoted by $g_{\text{I+A}} \in G$ and $g_{\text{video}} \in G$, respectively. Although both classifiers were described in [41], we give a brief presentation of their architectures in the next paragraphs, for the sake of completeness.

Classifier $g_{\text{I+A}}$ was fed with two parallel inputs: a two-seconds-long (i.e., 100 timesteps-long) time-window with the raw signals from the inertial sensors, and the mean values of the ambient sensors in the correspondent timestamps. The inertial data was processed by a one-dimensional Convolutional Neural Network (CNN) [53], composed of two convolutional layers interspersed with pooling layers, followed by a Long Short-Term Memory (LSTM) recurrent layer [16], generating the feature vector v_1. The ambient data was processed by a single fully-connected layer, generating the feature vector v_2. Both v_1 and v_2 were concatenated and sent to a softmax output layer.

Classifier g_{video} has taken, as input, a sequence of 25 optical flow pairs, correspondent to two seconds of video, computed with the TVL1 algorithm [52]. The InceptionV3 CNN architecture [48] was trained to classify each optical flow pair individually. The CNN-LSTM architecture, adopted by the authors, consisted of feeding each optical flow pair within a sequence to this pre-trained InceptionV3 module, and feeding the resulting features as inputs to each timestep of a LSTM layer, whose outputs were connected to a softmax output layer.

Both above-mentioned classifiers were endowed with softmax activation in their outputs, which ensured that the prediction vector respects a valid probability distribution. To provide the prediction vectors, we split each recording session in $N_T = 140$ timesteps, regardless to its original length, and used the referred framework to provide the predictions on each of those timesteps. The effect is to assume that all activities have similar length, a simplification that allowed the design of more uniform and comparable experiments related to the bioinspired computational models (Subsection 4.1), and the robot simulation (Subsection 5.4).

The output of the activity recognition module is, for a whole recording session x processed by a classifier g, a total of $N_X = 144$ sets of prediction vectors $y_x = \{y_x^1, \ldots, y_x^{N_T}\}$, with $N_T = 140$. Outputs from both classifiers $g_{\text{I+A}}$ and g_{video} were applied to all simulations, as described in the following subsections.

5.2 Heuristics Model Implementation

Two policies H were considered for the heuristics model, named *window* or *exponential*, that is, $H = \{h_{\text{window}}, h_{\text{exponential}}\}$. This experimental setup resulted in a total of four conditions for evaluation in the neurorobotics approach, given by the space $G \times H$.

The window policy consisted of deriving a wider prediction vector y_w^t, correspondent to N_{r_w} timesteps. This was done by averaging the N_{r_w} most recent prediction vectors in y_x, from the activity recognition module, as in Equation 11. We have set $N_{r_w} = 8$, which corresponds to windows of four seconds from the recording sessions, because this was the length of the segments considered for the neurorobotics approach (see Subsection 4.2). On the other hand, the exponential policy consisted of deriving a prediction vector y_e^t that considered the whole sequence of previous prediction vectors in y_x, with an exponential decay across iterations, as in Equation 12. If $R = \{r_w, r_e\}$ is the set of the functions to compute y_w^t and y_e^t, then the decision d_r of the heuristics approach, for either the window or exponential policies, is given by Equation 13. It is important to note that, for the window policy of the heuristics approach, as in the

neurorobotics approach, the robot can only begin to move after the first four seconds of each simulation, in which it is gathering the number of prediction vectors necessary to compute the first decision.

$$r_w : Y \times T \to Y \iff y_w^t = r_w(y_x, t) = \frac{\sum_{i=0}^{N_{r_w}-1} y_x^{t-i}}{N_{r_w}} \\ y_x \in Y, \quad t \in T_X | t > N_{r_w}, \tag{11}$$

$$r_e : Y \times T \to Y \iff y_e^t = \begin{cases} r_e(y_x, t=0) = y_x^t \\ r_e(y_x, t>0) = 0.9 \cdot r_e(y_x, t-1) + y_x^t \end{cases} \\ y_x \in Y, \quad t \in T_X \tag{12}$$

$$d_R : Y \times T \times R \to B \iff d_R(y_x, t, r) = f_B(argmax[r(y_x, t)]) \\ y_x \in Y, \quad t \in T_X, \quad r \in R \tag{13}$$

As a reference, we introduced an additional approach, a control condition in which the ground truth labels are directly fed to the robot simulation, providing a unique decision d_{GT} every timestep, as shown by Equation 14.

$$d_{GT} : X \to B \iff d_{GT}(x) = f_B(f_A(x)), \\ x \in X \tag{14}$$

5.3 Neurorobotics Model Implementation

Let M be the bioinspired computational model, which can be rat-based or marmoset-based, that is, $M = \{m_{\text{rat}}, m_{\text{marmoset}}\}$. This experimental setup resulted in a total of four conditions for evaluation in the neurorobotics approach, given by the space $G \times M$. Each independent simulation of the computational model (not to be confused with the robot simulation) was ran for each of the $N_X = 144$ recording sessions under each condition being evaluated, that is, the simulations of the computational models were required to contemplate all instances in the space $X \times G \times M$. Hence, a total of 576 simulations of the computational model was performed.

Each of those simulations ran for 70 seconds with sampling rate of $1,000$ Hz. The stimuli set S was updated every 0.5 second (i.e., update frequency of 2 Hz). This led to an adaptive dynamic that would respond to successive prediction vectors y_x^t, $t_{\text{sim}} \in \{1, \ldots, N_{T_{\text{sim}}}\}$, with $N_{T_{\text{sim}}} = 140$, according to the confidence of each response behaviour. The resulting spike trains in each neuron population were converted to neural firing frequencies (for details, see [23]), with bins of size 20, which resulted in sequences of length $N_U = 3,500$. As stated in Subsection 5.1, for the experiments reported in this work, $N_B = 2$, hence $N_C = 2$. Considering that $N_R = 8$, the multivariate time-series $u_x^{g,m} \in U$ had $N_R \cdot N_C = 16$ variables and $L_U = 3,500$ timesteps, composing a data structure with dimensions $3,500 \times 16$,

The segments for the decoder we set to $N_V = 200$ timesteps (i.e., four-seconds-long) with 75% superposition (i.e., a one-second-long step between the beginning of each segment), resulting in 66 segments. Considering the each condition was composed of $N_x = 144$ recording sessions, these simulations of the computational models led to a total of $144 \cdot 66 = 9,504$ instances $v \in V$, for each $(g, m) \in G \times M$.

The CNN architecture for decoding these time-series into response behaviours is depicted in Table 1. It was composed of two convolutional layers, with 128 and 256 filters, respectively, interspersed with max-pooling layers. A global average pooling operation preceded the softmax output layer, which produced the decision vector q_x^t.

For each set of conditions, the CNN was trained in a cross-subject 8-fold cross-validation scheme, similar to the one adopted for the activity recognition module [41]. The input data was linearly normalised to the range $[0, 1]$, and the classification models were trained for 40 epochs with batch size 32. The ADAM algorithm was employed, with learning rate 10^{-3}, to optimise the categorical cross-entropy loss function. The outputs of the evaluations were stored and organised, in order to serve as inputs to the next steps. The resulting sequences $u_x^{g,m}$ were then introduced to the decision-making mechanism.

5.4 Robot Behaviours

The behaviours $b \in B$ consisted of transporting an object $o \in O$, from a starting position $z \in Z$ to a fixed destination z_{dest}. This task was adopted because it comprises a basic and generic functionality for a mobile robot in a home

Table 1: Layers in the CNN-based decoder. The inputs to the neural network are windows of 200 timesteps from the four cortex channels of the output signals (i.e., neural firing frequencies) of the simulations under a given condition. The output is a decision vector q_x^t with the confidences for each response behaviour.

Layer	Type	Output shape	Free parameters
1	Input	200×4	-
2	Conv1D	200×128	$3,712$
3	MaxPool1D	100×128	-
4	Conv1D	100×256	$229,632$
5	MaxPool1D	50×256	-
6	Global Average Pooling	256	-
7	Softmax	3	-

environment. The associations between the behaviours and the objects are given by Equation 15, while the associations between the objects and their starting positions in the map are given by Equation 16.

$$f_O : B \to O \iff f_O(b) = o, \quad b \in B, \quad o \in O \tag{15}$$

$$f_Z : O \to Z \iff f_Z(o) = z, \quad o \in O, \quad z \in Z \tag{16}$$

At each timestep t_{robot}, a decision $d \in B \cup \{b_\emptyset\}$ (i.e., a response for each recording session $x \in X$ of the activity recognition module) was sent to the robot simulation, composed of a mobile social robot in a home environment (for details on the platforms and implementations employed, see Subsection 5.5). For the neurorobotics approach, this decision is given by Equation 10, already presented in Subsection 4.2. For the heuristics approach, the two policies mentioned (i.e., window and exponential) were evaluated.

The decisions were turned into commands to the robot following a table of rules, depicted in Table 2. A decision d is sent to the robot at each timestep. This decision can be one of the behaviours in $b \in B$ or the "no action" behaviour $b_\emptyset$. Let o_c be the object being carried by the robot at a certain timestep. Two types of situations might be considered: $d \in B$ or $d = b_\emptyset$.

Table 2: Table of rules associating a response behaviour $f(d) = b$ to an output command at each timestep t_{robot} of the robot simulation, considering the object being carried and the current robot position.

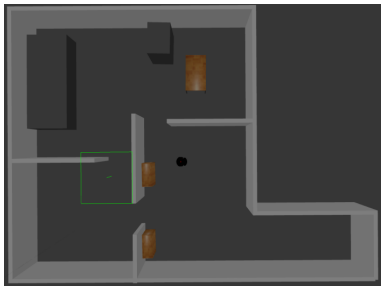
Decision	Object carried	Robot position	Output command
$b \in B$	$o_c = \emptyset$	$z \neq f_Z[f_O(b)]$	Move towards $f_Z[f_O(b)]$
$b \in B$	$o_c = \emptyset$	$z = f_Z[f_O(b)]$	Set $o_c = f_O(b)$
$b \in B$	$o_c = f_O(b)$	$z \neq z_{\text{dest}}$	Move towards z_{dest}
$b \in B$	$o_c = f_O(b)$	$z = z_{\text{dest}}$	Finish behaviour
$b \in B$	$o_c = o_k \in O \mid o_k \neq f_O(b)$	$z \neq f_Z[o_k]$	Move towards $f_Z[o_k]$
$b \in B$	$o_c = o_k \in O \mid o_k \neq f_O(b)$	$z = f_Z[o_k]$	Set $o_c = \emptyset$
$b_\emptyset$	$o_c = o_k \in O$	$z \neq f_Z[o_k]$	Move towards $f_Z[o_k]$
$b_\emptyset$	$o_c = o_k \in O$	$z = f_Z[o_k]$	Set $o_c = \emptyset$
$b_\emptyset$	$o_c = \emptyset$	$\forall z \in Z$	Wait

The first type of situation is characterised by $d = b_\emptyset$, in which the robot must return any object that it may be carrying to the corresponding position, and then stand still, waiting for any further commands. Otherwise, $d = b \in B$, the second type of situation, in which the robot is supposed to grab an object $o_c = f_O(b) \in O$ from position $f_Z[f_O(b)]$ to a destination z_{dest}. If the robot is not carrying any object, that is, $o_c = \emptyset$, then it must move to $f_Z[f_O(b)]$ and take the object. If it is already carrying the correct object, then it must move towards the destination z_{dest}. If it is carrying the wrong object, it is, $o_c = o_k \in O \mid o_k \neq f_O(b)$, then it must return it to $f_Z[o_k]$.

5.5 Robot Simulator

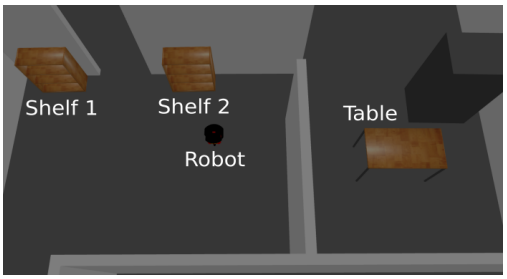
The simulator adopted for the robotics experiments was previously made available as part of the LARa framework [39], consisted of a robot and a software library. The LARa robot was a mobile social robot built on the top of a Pioneer P3-DX platform, endowed with a Hokuyo laser, a mini computer, a Microsoft Kinect sensor, a microphone,

a screen, and a speaker. The LARa library was a set of functionalities implemented to control the robot based on high-level software interfaces, integrated within the Robot Operating System (ROS) [37]. Besides navigation skills and a framework for human-robot interaction, this included a platform for simulation, under conditions that resembled those of the actual robot, deployed to allow offline experiments. The Gazebo simulator [20] was employed, and a map of a typical home environment was designed, as reproduced in Figure 5a. The simulated robot - a simplified version of the LARa robot - is shown in Figure 5b, while the pieces of furniture employed in the experiments are shown in Figure 5c.



(a) Map of the whole home environment employed for the experiments.

(b) Simulated mobile robot.



(c) In a different camera angle, the section of the map in which the robot behaviours were performed, with the indications of the robot and the pieces of furniture involved in the tasks (i.e., shelf 1, shelf 2 and table).

Figure 5: Virtual environment for the robot simulation, using the Gazebo platform.

This setting comprised a realistic environment, which provided several challenging aspects resembling those of a real-world scenario, such as sensors' noise, communication delays and mechanical issues. The ROS platform was employed to connect this simulated environment to a navigation stack, which provided a 2D occupancy grid in which each position (i.e., cell) might be considered empty, navigable or obstacle. This representation was generated previously to the robot simulations reported here, via the GMapping algorithm [13] for Simultaneous Localisation and Mapping (SLAM). The mapping algorithm ran while the robot was teleoperated through the whole environment, with the laser readings and the wheels' encoders combined to gradually compose the occupancy grid. Once the grid was created, the Augmented Monte Carlo Localisation (AMCL) and A* algorithms could be employed as a global planner to perform autonomous navigation. The navigation package was also endowed with a local planner, responsible for creating adaptable short-term paths for obstacle avoidance and environmental changes.

For this work, a set of two response behaviours was defined as $B = \{b_1, b_2\}$. In Table 3, are shown the set of daily activities from the dataset (i.e., $a_p \in A$, $p \in \{1, \ldots, N_A\}$), and the expected response behaviours associated to each of those activities (i.e., $f_B(a_p)$). These were chosen respecting semantic relationships between the activities (i.e., b_1 is the desired response when the user is preparing meals, and b_2, when he is quietly consuming or exchanging information).

These behaviours were based on the assumption that the user is located in the kitchen, and that the human activities are being monitored by sensors that are not affected by the robot actions. The starting position for only the first robot simulation in a battery of experiments is given in Figure 5. However, it had negligible effect in the overall results, since this position was not reset for each simulation, as we discuss later in this subsection.

As shown in Figure 5c, three pieces of furniture are considered. These are *shelf 1*, associated to the robot position $z_{s1} = f_Z(o_1)$, $o_1 \in O$; *shelf 2*, associated to the robot position $z_{s2} = f_Z(o_2)$, $o_2 \in O$; and *table*, the destination, associated to the robot position z_{dest}. The two specific behaviours considered for the experiments performed, b_1 and b_2, consist, respectively, of transporting object o_1 from z_1 (i.e., shelf 1) to z_{dest} (i.e., the table), and transporting object o_2 from z_2 (i.e., shelf 2) to z_{dest} (i.e., the table). Considering that *shelf 2* is closer to the *table* than *shelf 1*, then the

Table 3: List of activities provided by the HWU-USP activities dataset, and expected response behaviours in the application scenario.

Activity a	Description	Response behaviour b
a_1	making a cup of tea	b_1
a_2	making a sandwich	b_1
a_3	making a bowl of cereals	b_1
a_4	using a laptop	b_2
a_5	using a phone	b_2
a_6	reading a newspaper	b_2
a_7	setting the table	$b_\emptyset$
a_8	cleaning the dishes	$b_\emptyset$
a_9	tidying the kitchen	$b_\emptyset$

distances required for b_1 are larger than those for b_2. As a consequence, it was expected that, on average, b_1 required more time to be completed than b_2.

The maximum robot simulation time was set to $N_{T_{\text{robot}}} = 140$ seconds, with each timestep $t_{\text{robot}} \in \{1, \ldots, N_{T_{\text{robot}}}\}$ corresponding to one second in the simulation. Consequently, an expected response behaviour had to be finished within $N_{T_{\text{robot}}}$ seconds to be considered successfully completed. We configured $N_{T_{\text{robot}}} = 140$, which in exploratory experiments has shown to give a reasonable margin for the robot simulations.

A total of $N_X = 144$ robot simulations was performed for each condition analysed. The first simulation for each approach began with the robot positioned as in Figure 5c. All the next simulations began without resetting the robot position after the ending of the previous one, with only the object flag, corresponding to the object o_c being carried by the robot, being cleared. In this scenario, each simulation could be started with the robot in any of the positions in Z, or in locations belonging to the path between them.

6 Results

Concerning the activity recognition module, its classification results are presented in [41]. The overall accuracy registered for the classifiers were computed by taking a set of 25 prediction vectors obtained for a recording session and averaging it. A categorical classification was provided by returning the $argmax$ element in the averaged vector. A cross-validation approach, following the same cross-subject partitioning adopted for evaluating the CNN-based decoder in this work, have been performed. The accuracy reported for the modalities considered for the experiments reported here was 74.30% for g_{I+A}, and 93.75% for g_{video}.

The other modules in this work relied on important adaptations to frameworks previously implemented in related work, as happened to the computational models and the robot simulation, or to components developed from scratch, case of the CNN-based decoder. The corresponding results are shown in the following subsections. The classification metrics from the neural firing frequencies synthesised with the bioinspired computational models are presented in Subsection 6.1. The outcomes of the robot simulations, in all conditions analysed, are presented in Subsection 6.2.

6.1 Simulated Neural Firing Frequencies

A sample of the segments $v \in V$, provided in the simulations of the computational models, is shown in Figure 6. This was generated from a rat model, being stimulated according to an *IMU + ambient* classifier as the activity recognition module. A larger stimulus introduced to the striatum is expected to increase neural firing rates in the BG-T-C circuit, which might be propagated to the cortex.

The overall accuracy and F1-score of the decoder, trained and evaluated according to the 8-fold cross-subject approach described, are shown in the bars plot of Figure 7. The classifier used in the activity recognition module and the computational model employed are shown side-by-side.

The decoder was applied as a part of the decision-making mechanism, responsible for providing decision vectors for the robot simulation. Hence, its results might be correlated to the correct outcomes of the decisions made during the robot simulation. In other words, a good accuracy of the decoder might result in more correct decisions of the robot, which may more often complete the tasks with the correct outcome. The next subsection will present the

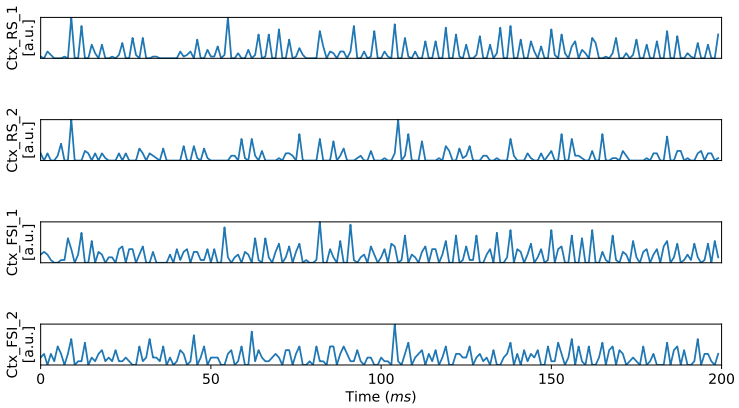


Figure 6: Sample output from the bioinspired computational model of the BG-C-T circuit. For the motor cortex of each channel, RS and FSI, we considered the overall mean firing rates computed with time bins of size 20 milliseconds, and evaluated on two-seconds-long time windows. This data was used as input for the CNN-based decoder, in the next step of the bioinspired pipeline.

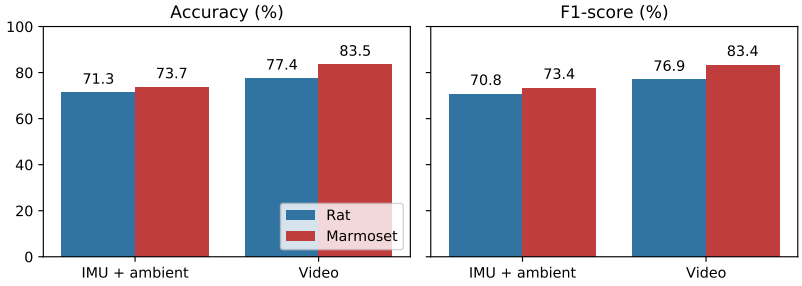


Figure 7: Accuracy and F1-score for the CNN-based decoder in classifying a MFR signal into a set of three possible decisions: B1, B2 or "no action". On choosing the models for evaluation, two factors were analysed: the modalities and models employed for activity recognition (IMU + ambient sensors or video-based) [41], and the computational model considered (rat-based or marmoset-based) [22, 40].

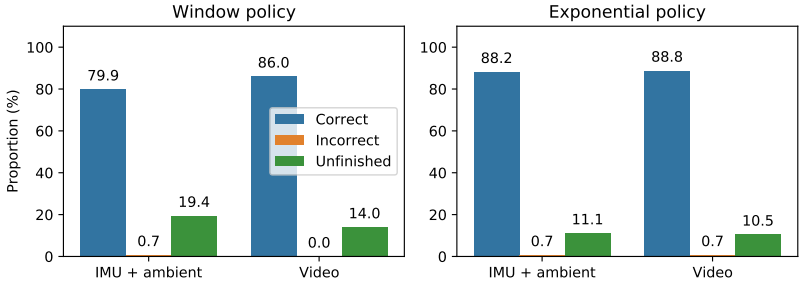
experiments performed to validate this statement. These are the outcomes of the robot simulation not only for each of those conditions, but also for each policy employed for the heuristics approach.

6.2 Outcomes of the Robot Simulations

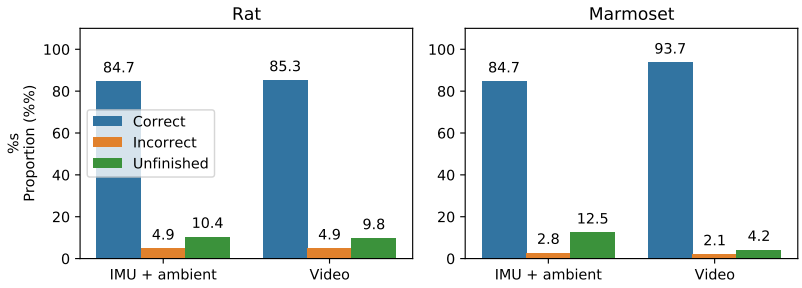
As it was mentioned before, three possible outcomes were considered for the robot simulations, with $f_A(x) = a$ being the activity associated to a recording session $x \in X$:

- *Correct*, if $f_B(a) \in B$ and the activity was completed before the end of the simulation, or if $f_B(a) = b_\emptyset$ and no behaviour was completed;
- *Incorrect*, if the robot completed a behaviour $b_{\text{robot}} \in B$ different from $f_B(a)$, i.e., $b_{\text{robot}} \neq f_B(a)$;
- *Unfinished*, if a response behaviour $b \in B$ was expected from the robot, but no behaviour was completed before the end of the simulation.

In Subsection 5.5, a control condition was introduced, with ground truth decisions being sent for the robot. For this approach, as it was expected, all robot simulations let to the *correct* outcome. In Figure 8a, the outcomes for the heuristics approach are presented, with each of the policies analysed (i.e., window and exponential) being represented in different plots, each illustrating the outcomes for each classifier considered for the activity recognition module. The outcomes for the neurorobotics approach are shown in Figure 8a, with the classifiers for activity recognition (IMU + ambient or video) and the computational models (rat or marmoset) being represented.



(a) Outcomes for the robot simulations performed with the heuristics approach. Four batteries of simulations were performed, considering two factors: the classifiers employed for activity recognition (IMU + ambient sensors and video-based) and the policy for the decision-making mechanism (window or exponential) (see Figure 4a).



(b) Outcomes for the robot simulations performed with the neurorobotics approach. Four batteries of simulations were performed, related to two factors analysed: the modalities and models employed for activity recognition (IMU + ambient sensors or video-based) [41], the computational model considered (rat-based or marmoset-based) [40] (see Figure 4b).

Figure 8: Outcomes for the robot simulations. Three possible outcomes were considered: the robot completed the expected (*correct*) behaviour; the robot concluded the *incorrect* behaviour; no behaviour was completed (*unfinished*), although an action was required from the robot.

The times elapsed for providing the *correct* outcome, when a response behaviour was expected from the robot, were also recorded. The mean and standard deviations, within all simulations performed for each condition, are represented in Figure 9. Two separate plots were provided, separating the classifiers employed for the activity recognition module. The ground truth approach was reproduced in both of them, since it does not depend on prediction vectors, but in the ground truth activities.

This metric considers only the outcomes completed successfully. An approach that provides a fast response with poor accuracy would provide a low time response, though it would not necessarily provide the correct response behaviours very often. Hence, the fact that the heuristics approach with the window policy led to a faster average response than the ground-truth condition is consistent. Since incorrect and unfinished outcomes were not considered for the computation of this mean value, this result only shows that, for this model, the correct outcomes were mostly associated to activities that could be completed in less time (e.g., the behaviour b_2).

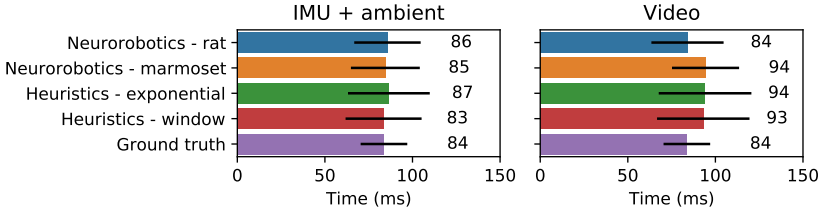


Figure 9: Average times elapsed across the 144 sequences on each simulation in which the *correct* behaviour was performed. Incorrect and unfinished outcomes, as well as correct outcomes when no action was required from the robot, were not considered in this evaluation. All simulated models, basend on either heuristics or neurorobotics, are shown.

7 Discussion

The results from the CNN-based decoder, shown in Figure 7, confirmed some expectations regarding the output signals produced by the simulations of the computational models according to the stimuli provided: it performed better for the the video-based classifier than for the IMU + ambient, and for the marmoset-based model, compared to the rat-based. The accuracy and F1-score metrics were very close, which considering a strictly balanced dataset, points that the results were not affected by any serious issues regarding the trade-off between precision and recall.

All evaluations led to an accuracy measure of over 70% for three classes (i.e., response behaviours b_1, b_2 or $b_\emptyset$). It is important to consider that the stimuli came from noisy prediction vectors from activity recognition algorithms, whose accuracy is variable across successive segments [41], with overall accuracy values of 74.30%, for g_{I+A}, and 93.75%, for g_{video}. These results show that the neural activity provided by the computational models could be reliably interpreted by the proposed decoder, even considering segments of limited length (i.e., four-seconds-long segments within a 70-seconds-long sequence). Hence, this particular technique for brain signals, analysed in previous studies for processing related neuronal data of the BG-T-C circuit [34, 42], has shown to be suitable for the decision-making approach proposed.

Since the accuracy measure of the classifier for activity recognition was significantly higher for the video-based classifier than for the IMU + ambient, it was expected that it could be more easily decoded by the neural network, which was confirmed by the decoder results (see Figure 7). Also, the marmoset-based model led to better decoding performances than the rat-based model, which also meets the expectations, considering a more sophisticated morphology and dynamics in the underlying brain structures in primates than in rodents [27].

Regarding the robot simulations, heuristics approaches were evaluated in parallel to the neurorobotics approaches. In most experiments performed in this work, better performances were found for the models fed by the video-based classifier than those fed by the IMU + ambient classifier, which was expected, since the video classifier is expressively more accurate [41]. As shown in Figure 8a, the window policy led to a lower number of successfully completed response behaviours, especially when fed with prediction vectors coming from the IMU + ambient classifier (less accurate). This condition may be the fairest comparison to the neurorobotics approach since it limits its decisions to data from the four-seconds-long segment that precedes a given decision, the same constraint applied to the CNN-based decoder.

In this context, the neurorobotics approach has shown to provide more accurate outcomes in most conditions, especially for the marmoset model. For the IMU+ambient modality of the activity recogniser, the window policy of the heuristics approach led to 79.9% of correct outcomes, which was surpassed by the 84.7% result for either the rat or marmoset models. For the video modality, the window policy of the heuristics approach led to 86.0% of correct outcomes, which was only slightly above the rat model, which hit 85.3%, and expressively below the marmoset model, which hit 93.7%. These results point that the proposed neurorobotics approach, in the conditions analysed in this study, may lead to better outcomes than a simple heuristics for a real-time task of an autonomous robot.

For the exponential policy of the heuristics approach, a particularity was found: it led to similar results for either the video and IMU + ambient conditions (i.e., 88.2% and 88.8% of correct outcomes, respectively), both with more correct outcomes than those of the window policy. This result is relevant, since it reveals that, by performing a long-term aggregation of prediction vectors obtained subsequently from a single recording session, it may be possible to compensate lower accuracy values provided by certain classifiers that work with different sets of sensors. This possibility might be considered in practical applications, in which more informative modalities that usually lead to high accuracy, such as videos, may be either difficult to be obtained, due to privacy concerns [9], or unfeasible to provide real-time outputs, due to the high computational cost inherent to the operations required for processing them [45].

Regarding the different conditions considered for the neurorobotics approach (i.e., the activity recogniser and the computational model), the expectation was that, when applied to the robot simulation, the number of correct outcomes would be comparatively proportional to the accuracy measures of the decoder (see Figure 7). As shown in Figure 8b, this expectation was met for most conditions, although some exceptions were found.

Better results for the marmoset model were expected, since the number of neurons and the connectivity are larger [36, 21]. The results of the decoder, previously discussed, corroborate to this hypothesis. For the robot simulations, considering the video modality, the marmoset model led to the best results found among all of the simulations, with 93.7% of correct outcomes, against 85.3% achieved by the rat model. However, for the IMU + ambient modality, the results were similar for both models. A possible explanation for this result is that such an increased capacity could compensate the mistakes for a more accurate activity recognise. In other words, the prediction vectors across successive segments could assign higher confidence values (i.e., probabilities) to the expected label (i.e., the ground-truth activity)

for the video-based classifier than for the IMU + ambient, and the marmoset-based model, more sophisticated, was able to take more advantage on it than the rat-based.

By measuring the time elapsed in the robot simulations with correct outcomes, we can see only modest variations across conditions. An important observation regarding this metric is that a fast response is not necessarily an indication of a good performance, since this result is affected not only for the assertiveness of the correct outcomes (i.e., few changes of decision within a simulation), but also to the accuracy of the simulations in a given set of conditions. For instance, a given condition may lead to fast response when it provides the correct outcome, but most simulations may lead to an incorrect or unfinished outcome.

For the neurorobotics approach, the times were approximately similar between both classifiers, except for the marmoset model, which took significantly longer to finish, on average, when fed with the video-based classifier. Considering the heuristics approach, the video-based classifier led to clearly longer times for completing the behaviours, which was probably because some of the changes in decisions (i.e., the robot is performing behaviour b_1, but the decision-making mechanism changes it to b_2 after receiving new, updated prediction vectors) within the simulations allowed for completing more simulations with the correct outcome. The same reason explains why the correct outcomes of the window policy for the IMU + ambient classifier led to a faster response, on average, than the ground-truth value.

8 Conclusions and Future Work

In this paper, we employed a neurorobotics approach based on the embodiment of validated computational models of brain structures for creating a decision-making mechanism to provide effective response behaviours to a mobile robot in a simulated environment.

The chosen application scenario was a simulated smart home where data from the sensed environment was processed with a previously designed activity recognition framework. The neurorobotics approach was compared to some heuristics. For this, two simple heuristics were proposed and evaluated to provide real-time decisions based on the outputs from an activity recognition classifier.

The neurorobotics model used computational models (CM) of the basal ganglia-thalamus-cortex (BG-T-C) circuit, originally designed to study the underlying mechanisms of Parkinson's Disease. The CM were adapted, so that the outputs from the activity recognition module were applied as stimuli to the striatum of the circuit, and spike activity at the cortex was decoded with a convolutional neural network (CNN) to provide decisions to the robot simulation. Different conditions were analysed, including whether the computational models were based on rodent or primate models.

Results were reported with respect to the accuracy obtained for the CNN-based decoder in each condition for the computational model, and to the outcomes of the robot simulations, considering the neurorobotics and the heuristics approaches. The expectations were met for most of the different conditions regarding the neurorobotics approaches. The primate-based computational model led to the best outcomes between the simulations analysed.

Hence, one can conclude that the proposed neurorobotics approach is promising not only as an embedded tool for understanding the neurophysiological aspects of animal behaviour, but also as a practical component to integrate decision-making mechanisms for action selection in mobile robots engaged in human-robot-interaction scenarios.

Future work may consist of providing a real-time simulation of the proposed application scenario, with a robot placed in a physical environment in which human participants may be performing activities. This would require the integration among the different modules shown in the pipelines presented, thus ensuring that all of them can work in real-time. Such an experiment may validate our approach in even more challenging conditions and scenarios, which may foster a wide range of applications.

Acknowledgement

This work was funded by the Sao Paulo Research Foundation (FAPESP), grants 2017/02377-5, 2017/01687-0 and 2018/25902-0, and the Neuro4PD project - Royal Society and Newton Fund (NAF\R2\180773). Moioli acknowledge the support from the Brazilian institutions: INCT INCEMAQ of the CNPq/MCTI, FAPERN, CAPES, FINEP, and MEC. This research was carried out using the computational resources from the CeMEAI funded by FAPESP, grant 2013/07375-0. Additional resources were provided by the Robotics Lab within the ECR, and by the Nvidia Grants program.

References

[1] Sara Ashry, Tetsuji Ogawa, and Walid Gomaa. CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network Using IMU Sensors of Smartwatch. *IEEE Sensors Journal*, 20(15):8757–8770, 8 2020.

[2] Davide Bacciu, Maurizio Di Rocco, Mauro Dragone, Claudio Gallicchio, Alessio Micheli, and Alessandro Saffiotti. An ambient intelligence approach for learning in smart robotic environments. *Computational Intelligence*, pages 1–28, 7 2019.

[3] Jyotika Bahuguna, Philipp Weidel, and Abigail Morrison. Exploring the role of striatal D1 and D2 medium spiny neurons in action selection using a virtual robotic framework. *Wiley Online Library*, 49(6):737–753, 3 2018.

[4] S. Bariselli, W. C. Fobbs, M. C. Creed, and A. V. Kravitz. A competitive model for striatal action selection, 6 2019.

[5] Davide Calvaresi, Daniel Cesarini, Paolo Sernani, Mauro Marinoni, Aldo Franco Dragoni, and Arnon Sturm. Exploring the ambient assisted living domain: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 8(2):239–257, 2017.

[6] Caitlyn Clabaugh and Maja Matarić. Escaping Oz: Autonomy in Socially Assistive Robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):33–61, 5 2019.

[7] Ted M. Dawson, Todd E. Golde, and Clotilde Lagier-Tourenne. Animal models of neurodegenerative diseases. *Nature Neuroscience*, 21(10):1370–1379, 2018.

[8] Salvador Dura-Bernal, Benjamin A Suter, Padraig Gleeson, Matteo Cantarelli, Adrian Quintana, Facundo Rodriguez, David J Kedziora, George L Chadderdon, Cliff C Kerr, Samuel A Neymotin, Robert A McDougal, Michael Hines, Gordon MG Shepherd, and William W Lytton. Netpyne, a tool for data-driven multiscale modeling of brain circuits. *eLife*, 8:e44494, apr 2019.

[9] Francisco Erivaldo Fernandes Junior, Guanci Yang, Ha Manh Do, and Weihua Sheng. Detection of Privacy-sensitive Situations for Social Robots in Smart Homes. In *Automation Science and Engineering (CASE), 2016 IEEE International Conference on*, pages 727–732. IEEE, 2016.

[10] Felipe Aparecido Garcia, Caetano Mazzoni Ranieri, and Roseli Aparecida Francelin Romero. Temporal approaches for human activity recognition using inertial sensors. In *Proceedings - 2019 Latin American Robotics Symposium, 2019 Brazilian Symposium on Robotics and 2019 Workshop on Robotics in Education, LARS/SBR/WRE 2019*, pages 121–125. Institute of Electrical and Electronics Engineers Inc., 10 2019.

[11] Ilche Georgievski, Tuan Anh Nguyen, Faris Nizamic, Brian Setz, Alexander Lazovik, and Marco Aiello. Planning meets activity recognition: Service coordination for intelligent buildings. *Pervasive and Mobile Computing*, 38:110–139, 7 2017.

[12] B. Girard, N. Tabareau, Q. C. Pham, A. Berthoz, and J. J. Slotine. Where neuroscience and dynamic system theory meet autonomous robotics: A contracting basal ganglia model for action selection. *Neural Networks*, 21(4):628–641, 5 2008.

[13] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE transactions on Robotics*, 23(1):34–46, 2007.

[14] P. Halje, Ivani Brys, Juan J. Mariman, Claudio Da Cunha, Romulo Fuentes, and Per Petersson. Oscillations in cortico-basal ganglia circuits: Implications for parkinson's disease and other neurologic and psychiatric conditions. *Journal of Neurophysiology*, 122(1):203–231, 2019.

[15] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4–21, 4 2017.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[17] Javed Imran and Balasubramanian Raman. Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208, 1 2020.

[18] Hitoshi Kita and Takako Kita. Cortical stimulation evokes abnormal responses in the dopamine-depleted rat basal ganglia. *Journal of Neuroscience*, 31(28):10311–10322, 2011.

[19] Woo Ri Ko and Jong Hwan Kim. Behavior Selection of Social Robots Using Developmental Episodic Memory-Based Mechanism of Thought. In *2018 IEEE International Conference on Consumer Electronics - Asia, ICCE-Asia 2018*. Institute of Electrical and Electronics Engineers Inc., 11 2018.

[20] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. IEEE, 2004.

[21] James B. Koprich, Lorraine V. Kalia, and Jonathan M. Brotchie. Animal models of alpha-synucleinopathy for Parkinson disease drug development. *Nature Reviews Neuroscience*, 18(9):515–529, 8 2017.

[22] Karthik Kumaravelu, David T. Brocker, and Warren M. Grill. A biophysical model of the cortex-basal ganglia-thalamus network in the 6-OHDA lesioned rat model of Parkinson's disease. *Journal of Computational Neuroscience*, 40(2):207–229, 4 2016.

[23] Petr Lánský, Roger Rodriguez, and Laura Sacerdote. Mean Instantaneous Firing Frequency is Always Higher Than the Firing Rate. *Neural Computation*, 16(3):477–489, 3 2004.

[24] Junjun Li, Zhijun Li, Fei Chen, Antonio Bicchi, Yu Sun, and Toshio Fukuda. Combined Sensing, Cognition, Learning, and Control for Developing Future Neuro-Robotics Systems: A Survey. *IEEE Transactions on Cognitive and Developmental Systems*, 11(2):148–161, 6 2019.

[25] Kang Li, Jinting Wu, Xiaoguang Zhao, and Min Tan. Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-Mimicking Decision Mechanism. In *8th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2018*, pages 498–503. Institute of Electrical and Electronics Engineers Inc., 4 2019.

[26] Yabin Liang, Zikai Yan, Qi Zhang, Hongyu Liang, Xiyu Ji, Yin Liu, and Rong Liu. A decision-making model based on basal ganglia account of action prediction. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2019*, pages 1705–1710. Institute of Electrical and Electronics Engineers Inc., 12 2019.

[27] Jean Liénard and Benoît Girard. A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection. *Journal of Computational Neuroscience*, 36(3):445–468, 2014.

[28] Yantao Lu and Senem Velipasalar. Autonomous Human Activity Classification from Wearable Multi-Modal Sensors. *IEEE Sensors Journal*, 19(23):11403–11412, 12 2019.

[29] Chih Yao Ma, Min Hung Chen, Zsolt Kira, and Ghassan AlRegib. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 71:76–87, 2 2019.

[30] Jeffrey E. Markowitz, Winthrop F. Gillis, Celia C. Beron, Shay Q. Neufeld, Keiramarie Robertson, Neha D. Bhagat, Ralph E. Peterson, Emalee Peterson, Minsuk Hyun, Scott W. Linderman, Bernardo L. Sabatini, and Sandeep Robert Datta. The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. *Cell*, 174(1):44–58, 6 2018.

[31] Matthew M. McGregor and Alexandra B. Nelson. Circuit Mechanisms of Parkinson's Disease, 3 2019.

[32] Roghayeh Mojarad, Ferhat Attal, Abdelghani Chibani, Sandro Rama Fiorini, and Yacine Amirat. Hybrid Approach for Human Activity Recognition by Ubiquitous Robots. In *IEEE International Conference on Intelligent Robots and Systems*, pages 5660–5665. Institute of Electrical and Electronics Engineers Inc., 12 2018.

[33] Garrett Mulcahy, Brady Atwood, and Alexey Kuznetsov. Basal ganglia role in learning rewarded actions and executing previously learned choices: Healthy and diseased states. *PLoS ONE*, 15(2):1–26, 2020.

[34] Shu Lih Oh, Yuki Hagiwara, U. Raghavendra, Rajamanickam Yuvaraj, N. Arunkumar, M. Murugappan, and U. Rajendra Acharya. A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neural Computing and Applications*, pages 1–7, 2018.

[35] Francisco Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 1 2016.

[36] Tony J. Prescott, Fernando M. Montes González, Kevin Gurney, Mark D. Humphries, and Peter Redgrave. A robot model of the basal ganglia: Behavior and intrinsic processing. *Neural Networks*, 19(1):31–61, 1 2006.

[37] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[38] Caetano M. Ranieri, Scott MacLeod, Mauro Dragone, Patricia A. Vargas, and Roseli A. F. Romero. Human activities with videos, inertial units and ambient sensors. *Dryad Digital Repository*, 2021.

[39] Caetano M. Ranieri, Guilherme Nardari, Adam H.M. Pinto, Daniel C. Tozadore, and Roseli A.F. Romero. LARa: A robotic framework for human-robot interaction on indoor environments. In *Proceedings - 15th Latin American Robotics Symposium, 6th Brazilian Robotics Symposium and 9th Workshop on Robotics in Education, LARS/SBR/WRE 2018*, pages 383–389. Institute of Electrical and Electronics Engineers Inc., 12 2018.

[40] Caetano M. Ranieri, Jhielson M. Pimentel, Marcelo R. Romano, Leonardo A. Elias, Roseli A. F. Romero, Michael Lones, Mariana F. P. Araujo, Patricia A. Vargas, and Renan C. Moioli. A data-driven biophysical computational model of parkinson's disease based on marmoset monkeys. *Manuscript submitted for publication.*, 2021.

[41] Caetano Mazzoni Ranieri, Scott MacLeod, Mauro Dragone, Patricia Amancio Vargas, and Roseli Aparecida Francelin Romero. Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors. *Sensors*, 21(3):768–, 1 2021.

[42] C.M. Ranieri, R.C. Moioli, R.A.F. Romero, M.F.P. De Araujo, M.B. De Santana, J.M. Pimentel, and P.A. Vargas. Unveiling Parkinson's Disease Features from a Primate Model with Deep Neural Networks. In *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[43] C.M. Ranieri, P.A. Vargas, and R.A.F. Romero. Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach. In *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[44] Francisco J. Rodriguez Lera, Francisco Martín Rico, Angel Manuel Guerrero Higueras, and Vicente Matellán Olivera. A context-awareness model for activity recognition in robot-assisted scenarios. *Expert Systems*, 37(2):e12481, 4 2020.

[45] Itsaso Rodríguez-Moreno, José María Martínez-Otzeta, Basilio Sierra, Igor Rodriguez, and Ekaitz Jauregi. Video Activity Recognition: State-of-the-Art. *Sensors*, 19(14):3160, 7 2019.

[46] Marcelo R. Romano, Renan C. Moioli, and Leonardo A. Elias. Evaluation of Frequency-Dependent Effects of Deep Brain Stimulation in a Cortex-Basal Ganglia-Thalamus Network Model of Parkinson's Disease. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2020-July, pages 3638–3641. Institute of Electrical and Electronics Engineers Inc., 7 2020.

[47] Odongo Steven Eyobu and Dong Han. Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable IMU Sensor Data Using a Deep LSTM Neural Network. *Sensors*, 18(9):2892, 8 2018.

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA, 2016. IEEE.

[49] Gerd Tinkhauser, Alek Pogosyan, Huiling Tan, Damian M. Herz, Andrea A. Kühn, and Peter Brown. Beta burst dynamics in Parkinson's disease off and on dopaminergic medication. *Brain*, 140(11):2968–2981, 11 2017.

[50] S. J. van Albada and P. A. Robinson. Mean-field modeling of the basal ganglia-thalamocortical system. I. Firing rates in healthy and parkinsonian states. *Journal of Theoretical Biology*, 257(4):642–663, 2009.

[51] Patrick Van Der Smagt, Michael A. Arbib, and Giorgio Metta. Neurorobotics: From vision to action. In *Springer Handbook of Robotics*, pages 2069–2094. Springer International Publishing, 1 2016.

[52] C. Zach, T. Pock, and H. Bischof. A duality based approach for real-time TV-L 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[53] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS:818–833, 2014.

CHAPTER

6

CONCLUSION

This thesis aimed at contributing to the design of intelligent environments, by providing a contextualised framework of multimodal techniques for human activity recognition and employing them to an application scenario. The techniques for activity recognition were designed based on different types of sensors within an environment (i.e., videos, inertial units and ambient sensors), with multiple deep leaning techniques and datasets. In parallel, developments on the field of computational neuroscience, specifically regarding the neurophysiological aspects of Parkinson's Disease and its underlying mechanisms, were performed during the course of the research, and integrated to a decision-making framework as a neurorobotics approach. The application scenario consisted of a robot simulation in which a mobile social robot might provide responses to human activities, which could be done based on heuristic approaches, consisting of simple heuristics, or neurorobotics approaches, consisting of adaptations of brain simulations.

The objectives were met, initially by a comparative work involving different deep learning techniques for activity recognition, based on videos and inertial sensors. This was done by adopting the UTD-MHAD and Egocentric Multimodal datasets, and applying them to train CNN, LSTM and TCN-based models. Despite important architectural differences, the LSTM and TCN methods led to compatible results for most conditions analysed. The method for feature-level fusion performed well for the UTD-MHAD dataset, though it led to overfitting for the Egocentric Multimodal dataset, which may indicate that this technique require more precise synchronisation to provide good generalisation.

Methods implemented during these analyses were employed in a second set of experiments. This time, a new dataset was collected in the Robotic Assisted Living Testbed, at Heriot-Watt University, which provided daily activities with variable length that could include long-term dependencies, similar to what could be expected in an actual home environment. Also, ambient sensors were recorded and synchronised to the inertial and video data. The deep learning techniques were adapted to this new dataset, and evaluations have shown increased performances when the ambient sensor data was combined to the inertial units. The best results, nonetheless,

were obtained when the video modality was present.

The application scenario was designed as a simulation environment with the Gazebo platform. The home environment and robot are in line with the developments that have been taken place in the Robots Learning Laboratory (LAR), at ICMC-USP, especially regarding the LARa framework. The decision-making strategies were enhanced due to the introduction of the neurorobotics approach, based on the computational models implemented and adapted during the scholarship developed in the Robotics Laboratory at Heriot-Watt University, under supervision of professor Patricia Amancio Vargas. This was based on partial results provided by the Neuro4PD project, also addressed in a dedicated chapter of this thesis. The neurorobotic models behaved as expected for most conditions, with video-based classifiers performing better than IMU + ambient classifiers, and primate models performing better than rodent models. An interesting result was that the marmoset models based on video classifiers performed significantly better than the heuristic approaches.

6.1 Final Considerations

Regarding activity recognition, the objectives were met, initially, by a comparative research involving different deep learning techniques for activity recognition, based on videos and inertial sensors. This was presented in Chapter 2. Focus was given to modelling the temporal dependencies either in sequences of inertial data, in features extracted from optical flow (i.e., video data), and in fusion between those approaches. To this aim, LSTM and TCN were designed and evaluated in parallel. The use of LSTM modules in this type of application has been quite common. The TCN architecture was also considered, because, when this stage of the thesis was being developed, its introduction to the domain of video-based human activity recognition was still a novelty. The experiments have confirmed a compatibility between LSTM and TCN as modules of the neural networks applied for temporal modelling in both modalities considered (i.e., videos and inertial units). Besides the comparison between LSTM and TCN for activity recognition, the study also considered two approaches for fusion, it is, late fusion and feature-level fusion, the latter depending on time-synchronised snippets across modalities to provide accurate results.

Two public datasets were adopted for evaluations: the egocentic multimodal dataset, and the UTD-MHAD. A CNN-based model for individual RGB frames was also considered, with output features from a spatial CNN combined to the other models through weighted averaging (i.e., fusion between a model for inertial data, a model for videos based on sequences of previously computed optical flow maps, and a model for RGB frames extracted from videos), achieving accuracies up to 80.62% for the egocentric multimodal dataset, and 85.47% for the UTD-MHAD without considering depth data. This result for the UTD-MHAD was below state-of-the-art for these modalities, and was improved in the next step of the research presented in this thesis (see

Chapter 3). Nevertheless, it was constrained to reproduce the same architecture for both the late fusion and feature-fusion approaches, and also across datasets (i.e., to allow a fair collection of results, the same architectures and hyperparameters were used for all model calibrations and evaluations). Rather than improving the overall accuracy for a particular dataset, the aim of this study was to provide a comparison between the different conditions analysed, and it was successful in this regard, in which similar architectures endowed with either LSTM or TCN modules for temporal modelling led to compatible results for a couple representative datasets.

Two other limitations of the above-mentioned developments might be pointed. First, the temporal video-based models, which led to the best results, have the drawback of being heavily resource-consuming, and depending of the computation of optical flow maps as a preprocessing step, which makes it difficult to implement on real-time scenarios. In these situations, the less expensive approaches analysed, such as the inertial-based models, may be preferred despite of its lower accuracy. Second, the fusion approaches rely on synchronised data from multiple modalities, which is not always trivial to implement for real-time scenarios.

Part of the methods implemented during these analyses were employed in the next experiments, considering also data from ambient sensors. This step of the work was described in Chapter 3. This time, a new dataset was built from scratch as part of this work: the HWU-USP activities dataset, collected at the RALT lab in Edinburgh at Heriot-Watt University. More specifically, the dataset was composed of RGB and depth videos from the camera of a TIAGo robot, data from IMU sensors attached to the users' wrist and waist, and a set of ambient sensors (i.e., switches at the doors of wardrobes and drawers, motion sensors and power measurements) from a smart home. The objective was to build and study a multimodal dataset composed of RGB and depth videos, inertial units and ambient sensors from a smart home in the context of activities of daily living, all of them sharing a kitchen environment and performed in the context of a regular breakfast. A set of 16 participants performed 9 activities, resulting in a total of 144 instances that composed 116 min of recordings in total. All data were stored, made anonymous, and made available to the research community.

This dataset allowed for the proposal of multimodal approaches involving not only videos and data from inertial sensors, but also ambient sensors. To the best of our knowledge, this was the first public multimodal activities dataset that provides these three modalities altogether and synchronously. A deep learning framework was also proposed to perform experiments in such a multimodal scenario. It was based on two-dimensional CNN modules for feature extraction on RGB frames, depth images and optical flow pairs, and LSTM layers for temporal modelling, when applicable. Data from inertial sensors were fed to a similar architecture, with a one-dimensional CNN being applied to extract features to be modelled by a LSTM module. For these modalities, the same experiments were performed on both the HWU-USP and the UTD-MHAD datasets. Results varied from one modality to another, especially for the HWU-USP, in which the architectures based on computer vision, specifically after computing dense optical

flow, performed significantly better. These differences were smaller for the UTD-MHAD dataset.

The data from the ambient sensors, present only on the new HWU-USP, were introduced as an additional channel of information on the neural network that processed the inertial data, with no feature extraction: the binary variables were fed to a fully-connected layer whose output was concatenated to the IMU features extracted by the CNN-LSTM modules. As expected, the introduction of this modality led to expressive improvements in accuracy. The best multimodal model led to a very high accuracy, which points to the relevance of considering different sources of data to perform activity recognition tasks.

As already mentioned, the real-time acquisition of synchronised data across modalities is often challenging. Hence, all scenarios might be considered as valid approaches, even those with lower accuracy. For instance, it may be feasible to implement a model based only on inertial and ambient sensors, however the introduction of the video modality, which could expressively enhance the classification accuracy and confidence, may be out of question. This means that the accuracies achieved for the IMU-only and IMU + ambient scenarios are important even if they are below those of the video-based models, since they can be easily introduced to real-time applications in the future.

Another aspect of these results that need to be interpreted cautiously is that the HWU-USP dataset was constrained to a single, specific environment. Therefore, the models trained and evaluated using this dataset would not necessarily generalise to other environments, with different layout, furniture, and appearance. This is particularly critical for the video modality, based on recordings made from the same point of view. However, the other modalities may also be affected. For instance, for the inertial sensors, the displacements made by the participants were dependent on the positions of the furniture and objects present in the environment, and different motion patterns may arise in other scenarios. Nonetheless, if fed by a larger dataset or retrained to other specific environments, the architectures presented here are likely to provide results compatible to those presented here, an hypothesis that is reinforced by the results with the other datasets analysed (see Chapter 2).

The already mentioned studies in computational neuroscience were performed during the internship of the candidate at the Heriot-Watt University, Edinburgh, Scotland, leading to two outcomes in this context. A detailed description of the motivations, methods and results coming from this part of the work is present in Chapter 4. First, a deep framework was proposed to extract features related to Parkinson's Disease (PD) from Local Field Potential (LFP) brain signals of a marmoset monkey dataset. Different neural networks were applied as machine learning techniques, both as classifiers and autoencoders, and results were reported in terms of accuracy and properties of the representations learnt by each model. The deep networks presented classification metrics higher than the shallow networks, with accuracy up to 99.80% for the ConvLSTM model. The autoencoder embedding has shown to be informative of the PD-related features, with clustering approaches reaching homogeneity up to 91.27%, and higher

classification metrics when fed to a fully-connected network, in comparison to the raw input (e.g., 95.76% accuracy, against 93.65%). Pre-training the CNN, on the other hand, had little effect compared to training from scratch. Even though the convolutional networks extract features in the time domain, the input segments with higher attributions presented an enhanced peak at the beta frequency range of the average spectrum of the PD individuals when compared to the healthy ones. Regarding the intermediate representations of the convolutional layers, we have analysed the average power spectra at five frequency bands of feature maps with the highest attributions. The proposed methods and analysis may contribute for a better understanding of the mechanisms underlying Parkinson's disease.

Second, a realistic biophysical computational model was presented to resemble data from the basal ganglia-thalamus-cortex circuit from the marmoset monkey brain in both healthy and Parkinson's Disease (PD) conditions. To this aim, a data-driven strategy was designed based on the local field potential (LFP) dataset previously collected (the same data employed for the deep learning framework) from five adult marmosets, including healthy and 6-OHDA models of PD. Model optimisation and validation was accomplished with evolutionary algorithms. The proposed modelling strategy produced computational models that resembled both single-neuron mean firing rates and spectral LFP characteristics found in healthy and PD marmosets models. To the best of our knowledge, this was the first computational model of PD based on simultaneous, multisite electrophysiological recordings from a primate model of the disease. This work can facilitate the investigation of the mechanisms of PD and support the development of techniques that can inform new PD therapies. Also, the approach proposed can be potentially applied to other neural engineering problems where biological data can be used to fit multiscale models of brain circuits.

After this stage was completed, involving computational neuroscience developments, an application scenario was presented in the context of human-robot interaction. As presented and discussed in Chapter 5, data from a sensed environment (i.e., the HWU-USP activities dataset), processed with the activity recognition framework, was employed in a decision-making mechanism to provide response behaviours in a robot simulation. Two approaches were considered: a heuristics-based approach, in which two simple heuristics were proposed and evaluated to provide real-time decisions based on outputs from an activity recognition classifier, and a neurorobotics approach. The computational model of the BG-T-C circuit, originally designed to study the underlying mechanisms of PD, was adapted, so that the outputs from the activity recognition module were applied as stimuli to the striatum of the circuit, and spike activity at the cortex was decoded with a convolutional neural network (CNN) to provide decisions to the robot simulation. Within this bioinspired setting, an additional factor was analysed: whether the computational model was based on a primate or rodent model.

Results were reported with respect to the accuracy obtained for the CNN-based decoder in each condition for the computational model, and to the outcomes of the robot simulations,

considering both approaches proposed (i.e., the heuristics and the neurorobotics approaches). The expectations were met for most of the different conditions regarding the neurorobotics approaches. The marmoset-based computational model led to the best outcomes between the simulations analysed, which points that the proposed approaches are promising not only as computational models for understanding the neurophysiological aspects of animal behaviour, but also as practical components that may integrate decision-making mechanisms for action selection.

The fact that the application scenarios were implemented on a simulated environment allowed for separate executions of each of its steps. The activity recognition module ran on its own, generated its predictions, which were stored and, only in a future step, fed to the following modules. In the case of the heuristics approaches, these prediction vectors were directly fed to the robot simulation. For the neurorobotics approach, the computational model took these stored outputs as stimuli for the simulation of the brain activity, and, again its results were stored. The decoder was trained and evaluated upon the activity generated from the computational model, and its predictions were stored. Finally, the robot simulation ran using the prediction vectors stored from the decoder.

This step-by-step approach is unfeasible for real-time systems, which would be the case for a real-world application. In this case, the integration between those components would be challenged by bottlenecks on the execution time of several modules, especially the activity recognition. Such an integration effort would be far from trivial. It was not analysed in the present work because, with the introduction of the neurorobotics aspects to the simulation, the scope of this thesis was delineated to the reasoning aspect of the scenario. This includes the steps between the data gathering and the decision-making approach, but not the integration aspects needed for an actual end-to-end implementation. Such a development may be explored in future research.

6.2 Future Research

Following the results directly provided in this thesis, two directions may be depicted for future work. The first direction refers to the enhancement of the multimodal activity recognition methods, and the second one, to application scenarios that may be developed based on the results from this thesis. The next subsections discuss these possibilities.

6.2.1 Activity Recognition Methods

Regarding the activity recognition methods, the models explored in this thesis were focused on deep neural networks. This approach was in line with most recent advances on the literature. Nevertheless, more accurate techniques may be considered for each modality, and lead to other fusion approaches. Although the video modality has already provided accurate results,

close to 100% in some cases, there is still room for improvement regarding the modalities other than the videos, which is important for actual AAL applications, since video data might not always be available due to privacy issues. If those videos will be gathered by a robot's camera, the positioning of the robot, which may not always be facing the inhabitants of the environment, may be another limitation.

The neural networks employed for the inertial modality consisted of introducing the raw data, with minimum preprocessing, to combinations of one-dimensional convolutional or recurrent networks, for feature extraction and temporal modelling. Other types of preprocessing, however, may turn sequences of raw data into other representations, as provided in related work. This would be compatible to what has been done for the videos, which were classified by models that required the previous computation of optical flow maps.

As already stated, the field of human activity recognition has been fertile, giving rise to different, accurate approaches for each modality. The HWU-USP activities dataset allowed for the development of techniques that process sequences with long-term dependencies, differently from most work on the literature. This new benchmark may be employed to evaluate techniques that model explicitly these longer-term dependencies. Some activities have a sequence of more fine-grained actions (e.g., to prepare a sandwich, the participant had to place a plate on the board, take the ingredients from the fridge, take the bread from the cupboard, assemble the sandwich, and so on). Although such fine-grained annotations were not provided in the dataset, they may be introduced based on the videos, without the need of more data collection.

A new session of data collection could also be targeted. An aspect that may be improved is the introduction of additional points of view for the camera within an environment. Besides, other environments, with different background, furniture and objects may be considered, resulting in a more heterogeneous dataset that can be actually employed to train generalised models for activity recognition.

6.2.2 Application Scenarios

The scenarios considered for this thesis were implemented in a simulated environment, with each module processing its inputs separately, without concerns about real-time constraints. A challenging scenario for future research consists of the implementation of all the modules described in a real-time application, which could recognise activities online, based on human users within the same environment as a physical robot is present. In this scenario, the robot behaviours could affect the environment, closing the loop for a complete neurorobotics approach focused on human-robot interaction.

The real-time constraints would have to be considered when choosing between frameworks for activity recognition, models for behaviour selection, and other functions eventually introduced. Such an end-to-end approach would require participants performing activities in

an experimental setup, computers processing sensors inputs and algorithms for both activity recognition and behaviour selection. The physical robot, which could be any mobile robot, would then perform actual tasks in response to the users' actions. Besides evaluations based on the performance of the system, the participant's perceptions on the robot behaviour may also be applied to measure the social acceptability of the approach.

The LARa robot, or other mobile social robot, may be introduced not only as an actuator to accomplish tasks based on classification results, but also as an additional source of data, considering for example its own camera, microphone, and other sensors. To allow for such an experiment, an additional data collection procedure could be designed, beginning with a teleoperated robot in an inhabited environment (i.e., Wizard-of-Oz approach). Such a dataset would contemplate limitations regarding the availability of data, which are expected in real-world scenarios, and could lead to a challenging benchmark for human activity recognition in home environments.

6.2.3 Neurorobotics

Within the context of the the computational neurascience developments, future research may be directed to the evaluation of machine-learning methods to brain signals, in tasks such as diagnosis based on different sensing information, or on the replication of neurophysiology aspects, such as the computational model presented. New advances on computational models that resemble brain regions related to movement or decision-making may lead to neurorobotics models that can be employed to study different biological phenomena, or even to provide more biologically plausible robot behaviours.

Neurorobotics approaches focused on studying the underlying mechanisms of Parkinson's Disease may also be devised based on the bioinspired models proposed and assessed here. In this work, the role of the basal ganglia-thalamus-cortex circuit in decision-making processes was assumed as a mediator for behaviour selection. Research focused on Parkinson's Disease may build on the outcomes of this work to connect firing frequencies of a computational model to movements in robotic models, composing a synthetic testbed for studies on motor symptoms.

BIBLIOGRAPHY

BAHUGUNA, J.; WEIDEL, P.; MORRISON, A. Exploring the role of striatal D1 and D2 medium spiny neurons in action selection using a virtual robotic framework. **Wiley Online Library**, Blackwell Publishing Ltd, v. 49, n. 6, p. 737–753, 3 2018. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.14021>. Citation on page 21.

BAI, S.; KOLTER, J. Z.; KOLTUN, V. Convolutional sequence modeling revisited. In: **6th International Conference on Learning Representations, ICLR 2018, , April 30 - May 3, 2018, Workshop Track Proceedings**. Vancouver, BC, Canada: OpenReview.net, 2018. Citation on page 20.

BAKAR, U. A.; GHAYVAT, H.; HASANM, S. F.; MUKHOPADHYAY, S. C. Activity and anomaly detection in smart home: A survey. In: **Next Generation Sensors and Systems**. [S.l.]: Springer International Publishing, 2016. v. 16, p. 191–220. ISBN 9783319216713. Citation on page 18.

BARISELLI, S.; FOBBS, W. C.; CREED, M. C.; KRAVITZ, A. V. **A competitive model for striatal action selection**. [S.l.]: Elsevier B.V., 2019. 70–79 p. Citation on page 21.

CALVARESI, D.; CESARINI, D.; SERNANI, P.; MARINONI, M.; DRAGONI, A. F.; STURM, A. Exploring the ambient assisted living domain: a systematic review. **Journal of Ambient Intelligence and Humanized Computing**, Springer, v. 8, n. 2, p. 239–257, 2017. Citation on page 17.

CARREIRA, J.; NOLAND, E.; HILLIER, C.; ZISSERMAN, A. A Short Note on the Kinetics-700 Human Action Dataset. **http://arxiv.org/abs/1907.06987**, 7 2019. Citation on page 18.

CHAARAOUI, A. A.; CLIMENT-PÉREZ, P.; FLÓREZ-REVUELTA, F. A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living. **Expert Systems with Applications**, Elsevier Ltd, v. 39, n. 12, p. 10873–10888, 9 2012. ISSN 09574174. Citation on page 17.

CHAVARRIAGA, R.; SAGHA, H.; CALATRONI, A.; DIGUMARTI, S. T.; TRÖSTER, G.; MILLÁN, J. d. R.; ROGGEN, D. The Opportunity challenge: a benchmark database for on-body sensor-based activity recognition. **Pattern Recognition Letters**, v. 34, n. 15, p. 2033–2042, 2013. Citation on page 18.

CHEN, C.; JAFARI, R.; KEHTARNAVAZ, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: **Image Processing (ICIP), 2015 IEEE International Conference on**. Quebec City, QC, Canada: IEEE, 2015. p. 168–172. Citations on pages 18 and 20.

COOK, D. J.; CRANDALL, A. S.; THOMAS, B. L.; KRISHNAN, N. C. CASAS: A Smart Home in a Box. **Computer**, v. 46, n. 7, p. 62–69, 2013. Citation on page 18.

FOSTER, M. E.; ALI, S.; LITWIN, S.; PARKER, J.; PETRICK, R. P.; SMITH, D. H.; STINSON, J.; ZELLER, F. Using ai-enhanced social robots to improve children's healthcare experiences. In:

SPRINGER. **International Conference on Social Robotics**. [S.l.], 2020. p. 542–553. Citation on page 18.

GEORGIEVSKI, I.; NGUYEN, T. A.; NIZAMIC, F.; SETZ, B.; LAZOVIK, A.; AIELLO, M. Planning meets activity recognition: Service coordination for intelligent buildings. **Pervasive and Mobile Computing**, Elsevier B.V., v. 38, p. 110–139, 7 2017. ISSN 15741192. Citation on page 18.

GOMEZ-DONOSO, F.; ESCALONA, F.; RIVAS, F. M.; CAÑAS, J. M.; CAZORLA, M. Enhancing the ambient assisted living capabilities with a mobile robot. **Computational intelligence and neuroscience**, Hindawi, v. 2019, 2019. Citation on page 17.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu, v. 9, n. 8, p. 1735–1780, 11 1997. Citation on page 20.

IGLESIAS, A.; JOSÉ, R. V.-A.; PEREZ-LORENZO, M.; TING, K. L. H.; TUDELA, A.; MARFIL, R.; DUEÑAS, Á.; BANDERA, J. P. Towards long term acceptance of socially assistive robots in retirement houses: use case definition. In: IEEE. **2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)**. [S.l.], 2020. p. 134–139. Citation on page 17.

KHAN, H. T. Population ageing in a globalized world: Risks and dilemmas? **Journal of evaluation in clinical practice**, Wiley Online Library, v. 25, n. 5, p. 754–760, 2019. Citation on page 17.

KOENIG, N.; HOWARD, A. Design and use paradigms for gazebo, an open-source multi-robot simulator. In: IEEE. **2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)**. [S.l.], 2004. v. 3, p. 2149–2154. Citation on page 21.

KUEHNE, H.; JHUANG, H.; GARROTE, E.; POGGIO, T.; SERRE, T. HMDB: a large video database for human motion recognition. In: **2011 IEEE International Conference on Computer Vision (ICCV)**. Barcelona, Spain: IEEE, 2011. p. 2556–2563. Citation on page 18.

KUMARAVELU, K.; BROCKER, D. T.; GRILL, W. M. A biophysical model of the cortex-basal ganglia-thalamus network in the 6-OHDA lesioned rat model of Parkinson's disease. **Journal of Computational Neuroscience**, Springer New York LLC, v. 40, n. 2, p. 207–229, 4 2016. ISSN 15736873. Citations on pages 21 and 69.

LERA, F. J. R.; RICO, F. M.; HIGUERAS, A. M. G.; OLIVERA, V. M. A context-awareness model for activity recognition in robot-assisted scenarios. **Expert Systems**, Blackwell Publishing Ltd, v. 37, n. 2, p. e12481, 4 2020. ISSN 0266-4720. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12481>. Citation on page 18.

LI, J.; LI, Z.; CHEN, F.; BICCHI, A.; SUN, Y.; FUKUDA, T. Combined Sensing, Cognition, Learning, and Control for Developing Future Neuro-Robotics Systems: A Survey. **IEEE Transactions on Cognitive and Developmental Systems**, Institute of Electrical and Electronics Engineers Inc., v. 11, n. 2, p. 148–161, 6 2019. ISSN 23798939. Citation on page 18.

LI, K.; WU, J.; ZHAO, X.; TAN, M. Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-Mimicking Decision Mechanism. In: **8th Annual IEEE International Conference on Cyber Technology in Automation, Control**

and Intelligent Systems, CYBER 2018. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. p. 498–503. ISBN 9781538670569. Citation on page 18.

LIMA, W. S.; SOUTO, E.; EL-KHATIB, K.; JALALI, R.; GAMA, J. Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview. **Sensors**, MDPI AG, v. 19, n. 14, p. 3213, 7 2019. ISSN 1424-8220. Available: <https://www.mdpi.com/1424-8220/19/14/3213>. Citation on page 20.

LIU, J.; SHAHROUDY, A.; PEREZ, M. L.; WANG, G.; DUAN, L.-Y.; CHICHUNG, A. K. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 5 2019. ISSN 0162-8828. Citation on page 18.

MANZI, A.; MOSCHETTI, A.; LIMOSANI, R.; FIORINI, L.; CAVALLO, F. Enhancing activity recognition of self-localized robot through depth camera and wearable sensors. **IEEE Sensors Journal**, IEEE, v. 18, n. 22, p. 9324–9331, 2018. Citation on page 18.

MARKOWITZ, J. E.; GILLIS, W. F.; BERON, C. C.; NEUFELD, S. Q.; ROBERTSON, K.; BHAGAT, N. D.; PETERSON, R. E.; PETERSON, E.; HYUN, M.; LINDERMAN, S. W.; SABATINI, B. L.; DATTA, S. R. The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. **Cell**, Cell Press, v. 174, n. 1, p. 44–58, 6 2018. ISSN 10974172. Citation on page 21.

MCCANN, P. Urban futures, population ageing and demographic decline. **Cambridge Journal of Regions, Economy and Society**, Oxford University Press UK, v. 10, n. 3, p. 543–557, 2017. Citation on page 17.

MELO, L. A. d.; FERREIRA, L. M. d. B. M.; SANTOS, M. M. d.; LIMA, K. C. d. Socioeconomic, regional and demographic factors related to population ageing. **Revista Brasileira de Geriatria e Gerontologia**, SciELO Brasil, v. 20, n. 4, p. 493–501, 2017. Citation on page 17.

MOJARAD, R.; ATTAL, F.; CHIBANI, A.; FIORINI, S. R.; AMIRAT, Y. Hybrid Approach for Human Activity Recognition by Ubiquitous Robots. In: **IEEE International Conference on Intelligent Robots and Systems**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2018. p. 5660–5665. ISBN 9781538680940. ISSN 21530866. Citation on page 17.

MULCAHY, G.; ATWOOD, B.; KUZNETSOV, A. Basal ganglia role in learning rewarded actions and executing previously learned choices: Healthy and diseased states. **PLoS ONE**, v. 15, n. 2, p. 1–26, 2020. ISSN 19326203. Available: <http://dx.doi.org/10.1371/journal.pone.0228081>. Citation on page 21.

NI, B.; WANG, G.; MOULIN, P. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: **2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)**. Barcelona, Spain: IEEE, 2011. p. 1147–1153. ISBN 9781467300629. Citation on page 18.

OBESO, J. A.; MARIN, C.; RODRIGUEZ-OROZ, C.; BLESA, J.; BENITEZ-TEMIÑO, B.; MENA-SEGOVIA, J.; RODRÍGUEZ, M.; OLANOW, C. W. The basal ganglia in Parkinson's disease: Current concepts and unexplained observations. **Annals of Neurology**, v. 64, n. S2, p. S30–S46, 1 2009. Citation on page 20.

PAL Robotics. **TIAGo Handbook version 1.7.1**. 2017. <www.pal-robotics.com>. Citation on page 20.

PATEL, A.; SHAH, J. Smart ecosystem to facilitate the elderly in ambient assisted living. In: SPRINGER. **Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications**. [S.l.], 2021. p. 501–510. Citation on page 17.

PETRICK, R. P.; FOSTER, M. E. Knowledge engineering and planning for social human–robot interaction: a case study. In: **Knowledge Engineering Tools and Techniques for AI Planning**. [S.l.]: Springer, 2020. p. 261–277. Citation on page 18.

PRONIN, S.; WELLACOTT, L.; PIMENTEL, J. M.; MOIOLI, R. C.; VARGAS, P. A. Neurorobotic Models of Neurological Disorders: A Mini Review. **Frontiers in Neurorobotics**, Frontiers, v. 15, p. 26, 2021. ISSN 1662-5218. Citation on page 21.

RANIERI, C.; MOIOLI, R.; ROMERO, R.; ARAUJO, M. D.; SANTANA, M. D.; PIMENTEL, J.; VARGAS, P. Unveiling Parkinson's Disease Features from a Primate Model with Deep Neural Networks. In: **Proceedings of the International Joint Conference on Neural Networks**. [S.l.: s.n.], 2020. ISBN 9781728169262. Citation on page 69.

RANIERI, C.; VARGAS, P.; ROMERO, R. Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach. In: **Proceedings of the International Joint Conference on Neural Networks**. [S.l.: s.n.], 2020. ISBN 9781728169262. Citation on page 25.

RANIERI, C. M.; MACLEOD, S.; DRAGONE, M.; VARGAS, P. A.; ROMERO, R. F. Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 21, n. 3, p. 768–, 1 2021. Citation on page 35.

RANIERI, C. M.; MACLEOD, S.; DRAGONE, M.; VARGAS, P. A.; ROMERO, R. A. F. Human activities with videos, inertial units and ambient sensors. **Dryad Digital Repository**, Dryad, 2021. Citation on page 35.

RANIERI, C. M.; NARDARI, G.; PINTO, A. H.; TOZADORE, D. C.; ROMERO, R. A. LARa: A robotic framework for human-robot interaction on indoor environments. In: **Proceedings - 15th Latin American Robotics Symposium, 6th Brazilian Robotics Symposium and 9th Workshop on Robotics in Education, LARS/SBR/WRE 2018**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2018. p. 383–389. ISBN 9781538677612. Citations on pages 21 and 105.

RANIERI, C. M.; PIMENTEL, J. M.; ROMANO, M. R.; ELIAS, L. A.; ROMERO, R. A. F.; LONES, M.; ARAUJO, M. F. P.; VARGAS, P. A.; MOIOLI, R. C. Towards a marmoset computational model of parkinson's disease. **Manuscript submitted for publication**, 2021. Citation on page 69.

REISS, A.; STRICKER, D. Introducing a new benchmarked dataset for activity monitoring. In: **2012 16th International Symposium on Wearable Computers**. Newcastle, UK: IEEE, 2012. p. 108–109. ISBN 978-0-7695-4697-1. Citation on page 18.

SANDEEPA, C.; MOREMADA, C.; DISSANAYAKA, N.; GAMAGE, T.; LIYANAGE, M. An emergency situation detection system for ambient assisted living. In: IEEE. **2020 IEEE International Conference on Communications Workshops (ICC Workshops)**. [S.l.], 2020. p. 1–6. Citation on page 17.

SANTANA, M. B.; HALJE, P.; SIMPLÍCIO, H.; RICHTER, U.; FREIRE, M. A. M.; PETERSSON, P.; FUENTES, R.; NICOLELIS, M. A. Spinal cord stimulation alleviates motor deficits in a primate model of Parkinson Disease. **Neuron**, Cell Press, v. 84, n. 4, p. 716–722, 11 2014. Citations on pages 20 and 69.

SONG, S.; CHEUNG, N.-M.; CHANDRASEKHAR, V.; MANDAL, B.; LIRI, J. Egocentric activity recognition with multimodal fisher vector. In: **2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Shanghai, China: IEEE, 2016. p. 2717–2721. ISBN 978-1-4799-9988-0. Citations on pages 18 and 20.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In: **Computer Vision and Pattern Recognition (CVPR)**. Providence, RI, USA: [s.n.], 2012. Citation on page 18.

United Nations. **World Population Prospects 2019 - Population Division - United Nations**. 2019. Available: <https://population.un.org/wpp/DataQuery/>. Citation on page 17.

WEI, H.; JAFARI, R.; KEHTARNAVAZ, N. Fusion of Video and Inertial Sensing for Deep Learning–Based Human Action Recognition. **Sensors**, MDPI AG, v. 19, n. 17, p. 3680, 8 2019. ISSN 1424-8220. Available: <https://www.mdpi.com/1424-8220/19/17/3680>. Citation on page 18.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.]: Springer Verlag, 2014. v. 8689 LNCS, n. PART 1, p. 818–833. ISBN 9783319105895. ISSN 16113349. Citation on page 20.

APPENDIX

A

ETHICS CLEARANCE FOR THE HWU-USP ACTIVITIES DATASET

HERIOT WATT UNIVERSITY

MULTIMODAL ACTION RECOGNITION DATASET

INFORMATION SHEET

1. GENERAL INFORMATION

This experiment consists of a data collection based on multimodal data from individuals performing activities of daily living. It will consider inertial data from wearable devices, RGB and depth videos, as well as data from environmental sensors. All participants are expected to be adults without incapacitant physical or cognitive disabilities.

The data recorded is intended to be used in action recognition tasks using data-driven approaches, specifically machine learning. In order to provide our further experiments with this data with transparency and reproductivity, as well as to allow other research groups to come with different approaches for dealing with the same problem, we intend that our dataset will be made publicly available.

2. PREPARATION

Experiments will be performed at the ambient assisted living laboratory, Lyell Centre, Heriot-Watt University, Edinburgh Campus. The data collection for which you are being invited includes sensing inertial information using a device that will be attached to your waist using the clip shown in Figure 1. To allow it to be placed correctly, we kindly ask that you come to the session wearing trousers that allow such placement.

Figure 1. Waist clip with the inertial sensor.

When you arrive at the ambient assisted laboratory, you will be asked to wear the wristband shown in Figure 2, equipped with an inertial measurement unit, in your dominant arm. After that,

you will be asked to wear the waist clip shown in Figure 1, equipped with a similar device than the one in the wristband. If you feel any kind of discomfort or uneasiness due to the use of those devices, please tell any member of the staff, so that the referred device may be removed, or the experimental session may be canceled.

Figure 2. Wristband with the inertial sensor.

3. DATA ACQUISITION

Besides the inertial sensors presented in the previous sessions, RGB and depth videos will also be collected using a Microsoft Kinect sensor. Environmental sensors present at the laboratory, such as presence detectors, switches, and pressure sensors, may also be considered. Audio data will **not** be kept on the records. All data will be anonymized so that no records of your name or other data that may allow your identification will be kept. Regarding the RGB videos, all human faces will be blurred. The resulting dataset will be made available for the research community.

4. EXPERIMENTAL PROCEDURE

Once you are ready, the recording procedures will be prepared by the staff. You will be asked to perform a set of predefined activities. Each activity should be recorded twice, providing two samples of 10 seconds. The positions of the participant or the camera will be changed for each record. Each sample may be recorded more than once, in order to ensure the quality of the data collected. For each participant, the recording session may last up to 1h30. The activities will be the following:

a) *Searching object*: walk randomly around the house, as you were searching for a lost object. The position of the Kinect sensor will vary on each recording.

b) *Talking on the phone*: pretend to talk on a wireless phone (e.g., a cell phone), both while standing/wandering and sitting.
c) *Using a laptop*: you will be given a laptop to perform random tasks on a laptop, such as checking the news on the Internet. You will use it both on your lap, while you sit on a sofa, and in a table.
d) *Manipulating cell phone:* perform any action either on your own smartphone (preferably) or on a smartphone provided by the staff, both sitting and standing.
e) *Reading in paper*: read a book or paper magazine for a while.
f) *Cleaning surface*: you will be given a dry cloth, with which you will be asked to pretend you are cleaning surfaces, such as a table and other furniture.
g) *Cleaning floor*: you will be given a broom or a vacuum cleaner, with which you will be asked to pretend you are cleaning the floor of the house.
h) *Washing the dishes*: pretend you are washing the dishes, which will be performed without neither water or cleaning products.
i) *Preparing a meal*: pretend to prepare a meal, with non-harmful utensils and fake ingredients.
j) *Having a meal*: pretend you are having a meal, without actually eating.

Please note that other activities could be amended during the trial to comply with the data collection, in case any of the above shows to be inadequate. You can freely refuse to participate in the experiment, as well as withdraw your consent at any time during the experiment or in the future, without any kind of negative consequence.

5. CONTACT INFORMATION

This is a joint project between the Heriot-Watt University at Edinburgh, Scotland, and the University of Sao Paulo, Brazil. The responsible researchers are Caetano Mazzoni Ranieri (ICMC-USP) and Patricia Amancio Vargas (MACS-HWU), with whom you can get in touch through their respective email addresses: cmranieri@usp.br and p.a.vargas@hw.ac.uk. Doubts regarding this experiment, currently or in the future, as well as requests for withdraw of consent, might be directed to any of these researchers.

6. FUNDING

This project is being funded by the Sao Paulo Research Foundation (FAPESP), grants 2018/25902-0 and 2017/02377-5.

Grant award: Sao Paulo Research Foundation, grants 2018/25902-0 and 2017/02377-5.

Participant identification number for this session: S_____(e.g., S01, S02)

CONSENT FORM

Project: Multimodal Action Recognition Dataset

Name of researchers: Patricia Amancio Vargas; Caetano Mazzoni Ranieri

☐ I confirm that I have read and understood the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

☐ I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. My legal rights are not going to be affected.

☐ I understand that the data collected during the study may be looked at by individuals from the Heriot-Watt University and the University of Sao Paulo, and will be further made publicly available. I give permission for my records to be made available on these terms.

☐ The procedures regarding confidentiality have been clearly explained (e.g. use of names, anonymization of data, etc.) to me.

☐ The use of the data in research, publications, sharing and archiving has been explained to me.

☐ I agree to take part in the above study.

_____	_____
Name of participant	Date and signature
_____	_____
Name of person taking consent	Date and signature

You are logged in as **cm264** (student) Caetano Mazzoni Ranieiri

Research Projects Project System: Edit Project Details

Research Projects Academic year: 2016/17

Title	Multimodal Action Recognition Dataset
Student	Caetano Mazzoni Ranieiri
Supervisor	Patricia Vargas
Second Reader	
Third Reader	
Abstract:	The experiment consists of data collection based on multimodal data from individuals performing activities of daily living. It will consider inertial data from wearable devices, RGB and depth videos, as well as data from environmental sensors. The data recorded is intended to be used in action recognition tasks using data-driven approaches, specifically machine learning. Although there are some datasets relying on multimodal data, to the best of our knowledge, this will be the first dataset with video and inertial data consisting of daily living activities in indoor environments, especially with data from environmental sensors.
Ethics Approval	Approved
	Edit/View Ethics Form
	Update Current Project

Use the form below if when you are ready to upload your dissertation, or its associated code. Irrespective of your names for the files they will receive names based on your login.

Document Upload	
Dissertation Document (PDF)	Choose file No file chosen
Documents & Code (ZIP)	Choose file No file chosen Already uploaded as cm264_code.zip
Upload document(s)	

APPENDIX

B

PUBLICATIONS

During the development of PhD project, besides the papers included in the collection that composed this thesis, other papers were published by the candidate, either as main author or co-author. The conference papers published are the following (the list includes the two papers published at IJCNN 2020 and reproduced in Chapter 2 and Chapter 4.

- TOZADORE, DANIEL C. ; PINTO, ADAM H. M. ; RANIERI, CAETANO M. ; BATISTA, MURILLO R. ; ROMERO, ROSELI A. F. . ”Tablets and humanoid robots as engaging platforms for teaching languages”. In: 2017 Latin American Robotics Symposium (LARS) and 2017 Brazilian Symposium on Robotics (SBR), 2017, Curitiba. 2017 Latin American Robotics Symposium (LARS) and 2017 Brazilian Symposium on Robotics (SBR), 2017. p. 1.
- MOREIRA PINTO, ADAM HENRIQUE ; MAZZONI RANIERI, CAETANO ; VINCENTIN NARDARI, GUILHERME ; CARNIETO TOZADORE, DANIEL ; FRANCELIN ROMERO, ROSELI APARECIDA . ”Users’ Perception Variance in Emotional Embodied Robots for Domestic Tasks”. In: 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), 2018, Joao Pessoa. 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), 2018. p. 476.
- MAZZONI RANIERI, CAETANO; VICENTIM NARDARI, GUILHERME ; MOREIRA PINTO, ADAM HENRIQUE ; CARNIETO TOZADORE, DANIEL ; FRANCELIN ROMERO, ROSELI APARECIDA . ”LARa: A Robotic Framework for Human-Robot Interaction on Indoor Environments”. In: 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), 2018, Joao Pessoa. 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE), 2018. p. 376.

- TOZADORE, DANIEL ; RANIERI, CAETANO ; NARDARI, GUILHERME ; GUIZILINI, VITOR ; ROMERO, ROSELI . ”Effects of Emotion Grouping for Recognition in Human-Robot Interactions”. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), 2018, Sao Paulo. 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), 2018. p. 438.

- APARECIDO GARCIA, FELIPE ; MAZZONI RANIERI, CAETANO ; APARECIDA FRANCELIN ROMERO, ROSELI . ”Temporal Approaches for Human Activity Recognition Using Inertial Sensors”. In: 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), 2019, Rio Grande. 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), 2019. p. 121.

- RANIERI, CAETANO M.; VARGAS, PATRICIA A. ; ROMERO, ROSELI A. F. . ”Uncovering Human Multimodal Activity Recognition with a Deep Learning Approach”. In: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, Glasgow. 2020 International Joint Conference on Neural Networks (IJCNN), 2020. p. 1.

- RANIERI, CAETANO M.; MOIOLI, RENAN C. ; ROMERO, ROSELI A. F. ; DE ARAUJO, MARIANA F. P. ; DE SANTANA, MAXWELL BARBOSA ; PIMENTEL, JHIELSON M. ; VARGAS, PATRICIA A. . ”Unveiling Parkinson’s Disease Features from a Primate Model with Deep Neural Networks”. In: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, Glasgow. 2020 International Joint Conference on Neural Networks (IJCNN), 2020. p. 1.

Two papers were accepted and published at peer-reviewed journals. The first of them, which described an important part of the work developed here and is reproduced in Chapter 3, was published to the Sensors journal by MDPI. The second one, co-authored during the scholarchip period at Heriot-Watt University, was published to Frontiers in Neurorobotics. The references are reproduced as follows.

- RANIERI, CAETANO MAZZONI; MACLEOD, SCOTT ; DRAGONE, MAURO ; VARGAS, PATRICIA AMANCIO ; ROMERO , ROSELI APARECIDA FRANCELIN . ”Activity Recognition for Ambient Assisted Living with Videos, Inertial Units and Ambient Sensors”. SENSORS, v. 21, p. 768, 2021.

- PIMENTEL, JHIELSON M.; MOIOLI, RENAN C.; DE ARAÚDO, MARIANA F. P.; RANIERI, CAETANO M.; ROMERO, ROSELI A. F.; BROZ, FRANK; VARGAS, PATRICIA A. . ”Neuro4PD: An Initial Neurorobotics Model of Parkinson’s Disease”. FRONTIERS IN NEUROROBOTICS, 2021.

Two preprints have been uploaded to arXiv, and are currently being submitted to peer-reviewed journals. Both of them have been included to the collection of papers provided in this thesis (see Chapter 4 and Chapter 5). The references are listed as follows.

- RANIERI, CAETANO M.; PIMENTEL, JHIELSON M.; ROMANO, MARCELO R.; ELIAS, LEONARDO A.; ROMERO, ROSELI A. F.; LONES, MICHAEL; DE ARAUJO, MARIANA F. P.; VARGAS, PATRICIA A.; MOIOLI, RENAN C. . "A Data-Driven Biophysical Computational Model of Parkinson's Disease based on Marmoset Monkeys". *arXiv preprint arXiv:2107.12536*, 2021.
- RANIERI, CAETANO M.; MOIOLI, RENAN C.; VARGAS, PATRICIA A.; ROMERO, ROSELI A. F. . "A Neurorobotics Approach to Behaviour Selection based on Human Activity Recognition". *arXiv preprint arXiv:2107.12540*, 2021.

APPENDIX

C

OTHER ACTIVITIES

During the development of the project described in this thesis, the author participated on the following academic events:

- XXI Congresso Brasileiro de Automatica (CBA), by 3rd to 7th of October 2016, at the Federal University of Espirito Santo, Vitoria, ES, Brazil.
- Data Analysis in HRI Experiments, by 24th of October 2016, lectured by Gabriele Trovato, at the University of Sao Paulo, Sao Carlos campus.
- III Workshop do Centro de Robótica de Sao Carlos, by 27th of October 2016, at the University of Sao Paulo, Sao Carlos campus.
- Sao Paulo School of Advanced Science on Smart Cities, by 24th of July to 4th of August 2017, at the University of Sao Paulo, Sao Paulo campus.
- VII Brazilian Conference on Intelligent Systems (BRACIS), by 22nd to 25th of October 2018, at IBM Brasil, Sao Paulo, SP.
- XV IEEE Latin America Robotics Symposium, by 6th to 10th of November 2018, at Joao Pessoa, PB, Brazil.
- The Robotics Lab Research Open Day 2019, by 21st of June 2019, at Heriot-Watt University, Edinburgh campus.
- The Ambient Assisted Living (AAL) Summer School, sponsored by the Scottish Informatics and Computer Science Alliance (SICSA), by 6th to 8th of August 2019, at Heriot Watt University, Edinburgh campus.
- Tutorial on "Cellular-Automata (CA) Models in Autonomous Robotics: Development and Trends", by 7th and 8th of August 2019, lectured by professor Gina Oliveira, at Heriot-Watt University, Edinburgh campus.

- IEEE World Congress on Computational Intelligence (WCCI) 2020, by 19th to 24th of June 2020, in Glasgow, Scotland, with online participation due to the COVID-19 pandemic.

The author participated to the Teaching Improvement Program (PAE), actuating as a teaching assistant of the subject "Neural Networks" ministred by professor Roseli A. F. Romero, to both undergraduate and graduate students, at ICMC-USP. This participation was performed in the second semester of 2017, 2018, and 2020.


ICMC
USP
SÃO CARLOS