Can The Way You Walk Reflect Your Health?

Pedro Flores de Albuquerque

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Paulo Luís Serras Lobato Correia

Examination Committee
Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino
Supervisor: Prof. Paulo Luís Serras Lobato Correia
Member of the Committee: Prof. Nuno Miguel de Pinto Lobo e Matela

January 2021
Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.
Acknowledgments

First and foremost, I want to thank Professor Paulo Correia for all his support, knowledge and availability throughout this work. Professor Luis Soares for his valuable inputs and João Machado with whom I collaborated and shared important work experiences. I would like to acknowledge the help and feedback from Tanmay and João Loureiro and also their previous work, upon which I was able to develop mine. I would also like to thank Instituto de Telecomunicações, with whom I developed this work.
Resumo

A análise da marcha humana pode providenciar informação relevante ao diagnóstico e recuperação de pacientes com distúrbios na marcha. Como alternativa à prática corrente de análise subjetiva, técnicas modernas de detecção e classificação automática de marcha patológica usam métodos de análise profunda baseados em dados visuais. Desenvolver e testar tais métodos requer a disponibilidade de bases de dados de marcha patológica apropriadas. Contudo, majoritariamente devido a questões de ética e privacidade quanto à partilha de dados de pacientes reais, poucas existem publicamente acessíveis. Além disto, as que são disponibilizadas têm no máximo 10 voluntários a simular marchas patológicas. Neste contexto, foi desenvolvida uma nova base de dados de marcha patológica denominada GAIT-IT. Com 21 sujeitos a simular 4 tipos de marcha patológica e a respetiva marcha normal, GAIT-IT é consideravelmente maior que as bases de dados atualmente disponíveis. Os vídeos adquiridos num estúdio profissional permitiram a extração de silhuetas binárias de alta qualidade que possibilitam melhores condições de treino a algoritmos de aprendizagem automática.

É proposta uma nova abordagem de aprendizagem profunda para classificação de marcha patológica, que combina as capacidades de extração de padrões espaciais e temporais de redes convolucionais e recorrentes, respectivamente. O sistema CNN-LSTM proposto processa sequências de marcha como conjuntos de silhuetas, permitindo a aprendizagem de padrões temporais entre padrões espaciais extraídos de diversos momentos. O sistema proposto é capaz de superar os resultados de soluções recentes, com maior capacidade de generalização sobre novos dados e suporte a alternativas de menor complexidade sem comprometer o desempenho.

Palavras-chave: Análise da marcha, Marcha patológica, Aprendizagem Profunda, Classificação de marcha, Sequências de Vídeo, Base de dados
Abstract

Gait analysis, i.e. the study of human motion, can provide useful information for the diagnosis, monitoring and recovery of gait related pathologies. As an alternative to the current standard practice that relies on subjective assessment, the state-of-the-art vision based approaches for automatic gait pathology detection and classification use deep learning. The development and testing of such solutions requires the availability of suitable pathological gait datasets. However, mainly due to privacy and ethical concerns associated with sharing data from real patients, few gait pathology datasets are publicly available. Furthermore, in those available, volunteers simulate gait affected by specific pathologies, considering at most 10 subjects. To address this, the developed work presents a new pathological gait dataset, GAIT-IT. Captured from 21 subjects simulating 4 gait pathologies and normal gait, GAIT-IT is significantly larger than publicly available gait pathology datasets. With video sequences recorded in a professional studio, nearly perfect binary silhouettes were extracted, allowing the improved training of machine learning algorithms.

A novel deep learning approach for pathological gait classification is proposed, combining the spatial and temporal feature extraction abilities of convolutional and recurrent neural networks, respectively. The proposed CNN-LSTM framework processes gait cycles as a collection of silhouette frames, allowing the system to learn temporal patterns among the spatial features extracted at individual time steps. Trained with gait sequences from GAIT-IT, the proposed system is able to outperform state-of-the-art solutions, achieve a greater generalization capability on cross-dataset tests and support a significant reduction in network complexity without compromising the overall performance.

Keywords: Gait analysis, Pathological gait, Deep learning, Gait classification, Video sequences, Gait dataset
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Resumo</td>
<td>vii</td>
</tr>
<tr>
<td>Abstract</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xvii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xix</td>
</tr>
<tr>
<td>Glossary</td>
<td>1</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 Context and Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Problem Definition</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Objectives</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Contributions</td>
<td>4</td>
</tr>
<tr>
<td>1.5 Thesis Structure</td>
<td>4</td>
</tr>
<tr>
<td><strong>2 Background: Pathological Gait Analysis</strong></td>
<td>5</td>
</tr>
<tr>
<td>2.1 Gait Analysis Concepts</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1 Gait Cycle Components</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2 Gait Parameters</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Pathological Gait Classification</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Evaluation Metrics for Classification Performance</td>
<td>9</td>
</tr>
<tr>
<td><strong>3 State of the Art: Gait Analysis Systems</strong></td>
<td>13</td>
</tr>
<tr>
<td>3.1 Gait Data Acquisition</td>
<td>13</td>
</tr>
<tr>
<td>3.2 Gait Analysis Systems Architecture</td>
<td>14</td>
</tr>
<tr>
<td>3.2.1 Pre-processing: from Video to a Suitable Gait Representation</td>
<td>15</td>
</tr>
<tr>
<td>3.2.2 Feature Extraction: Deep learning in Gait Analysis</td>
<td>20</td>
</tr>
<tr>
<td>3.2.3 Classification: Machine Learning Techniques</td>
<td>29</td>
</tr>
<tr>
<td><strong>4 GAIT-IT Dataset</strong></td>
<td>33</td>
</tr>
<tr>
<td>4.1 Motivation</td>
<td>33</td>
</tr>
<tr>
<td>4.2 Existing Datasets</td>
<td>34</td>
</tr>
<tr>
<td>4.3 Dataset description</td>
<td>34</td>
</tr>
<tr>
<td>4.4 Description of Simulated Gait Types</td>
<td>35</td>
</tr>
<tr>
<td>4.5 Video Data Processing</td>
<td>36</td>
</tr>
<tr>
<td>4.5.1 Background Subtraction</td>
<td>37</td>
</tr>
</tbody>
</table>
# Table of Contents

4.5.2 Skeleton Computation ........................................... 38  
4.5.3 GEIs and SEIs Computation ................................... 40  

5 Influence of the Dataset on Deep Learning Classifiers 45  
5.1 Baseline Solution .................................................. 45  
5.2 Cross-validation Experiments ...................................... 46  
5.3 Cross-dataset Tests .................................................. 48  

6 Proposed Pathological Gait Classification System 51  
6.1 Proposed Gait Analysis Framework ............................... 51  
   6.1.1 Key Frame Selection ........................................... 52  
   6.1.2 CNN Feature Extraction ....................................... 53  
   6.1.3 LSTM Feature Extraction ..................................... 56  
   6.1.4 Classification .................................................. 56  
6.2 Performance Comparison Using Different Convolutional Bases  .............................................. 57  
6.3 Fine-tuning the Convolutional Base ................................ 58  
6.4 Using a Bidirectional LSTM ....................................... 60  
6.5 Building a Convolutional Base For Pathological Gait Classification .............................................. 61  
   6.5.1 Proposed PGait CNN Architecture ............................ 61  
   6.5.2 Performance Evaluation ........................................ 62  

7 Experimental Results 67  
7.1 Cross-validation Results on GAIT-IT ............................ 67  
7.2 Cross-dataset Tests on GAIT-IST .................................. 67  
7.3 Performance Comparison on Different Gait Types .................. 69  

8 Conclusions 71  
8.1 Achievements ....................................................... 71  
8.2 Future Work ........................................................ 71  

Bibliography 73
List of Figures

1.1 Gait analysis applied to (a) clinical practice [6], (b) surveillance [7] and (c) sports [8]. ...
1.2 Marble copy of The Discobulus by Myron [9], portraying body movement. .................
1.3 The Running Lion Tamer, chronophotograph by Marey [13]. ..............................
2.1 Division of the gait cycle in eight functional phases, according to [19]. ............
2.2 Spatial parameters involving feet distances. ........................................
2.3 Illustration of characteristic movements in different types of impaired gait [25]. ....
2.4 Illustration of a confusion matrix for multi-class classification. ..............
3.1 Main modules of a biometric analysis system. ....................................
3.2 Components of background subtraction [35]. (a) background, (b) moving subject in foreground, (c) binary mask of the walking person. ........................
3.3 MEI and MHI examples. (a) Key movement frame, respective (b) MEI and (c) MHI [37]. 16
3.4 Example of binary silhouettes and corresponding GEI [39]. ......................
3.5 Detecting human action using a skeleton based model [48]. .......................... 19
3.6 Illustration of (a) Edge matching error, (b) region matching error, (c) structural model [51]. 19
3.7 (a) OpenPose result rendered on original image, (b) obtained binary skeleton image [27]. 20
3.8 Illustration of the GEINet architecture [59]. ........................................ 21
3.9 Topology of 3D CNN used in gait recognition [64]. ..............................
3.10 DeepGait proposed architecture [72]. ........................................
3.11 Fine-tined VGG-19 architecture [73]. ........................................
3.12 VGG-19 architecture for pathological gait classification [27]. ................
3.13 Illustration of (a) RNN [76] and (b) LSTM [77] architectures. ................
3.14 Illustration of (a) body joint heatmaps extraction, (b) architecture proposed for cross-view gait recognition in [78]. ........................
3.15 Illustration of bidirectional layers in a recurrent neural network [80]. ....
3.16 LSTM architecture adopted in [82]. ........................................
3.17 Illustration of a ConvLSTM cell [84]. ........................................
3.18 Illustration of (a) face features in wrong proportion and orientation, (b) CNN detecting a face based on unmatched face features [86]. ............................
3.19 CapsNet with 3 layers from [88]. ........................................
3.20 Illustration of an Artificial Neural Network [92]. ................................
3.21 Illustration of the LDA method [93]. ........................................
4.1 Subjects age distribution in the GAIT-IT dataset. ...........................
4.2 Simulation example of (a) first severity and (b) second severity of diplegic gait in the GAIT-IT dataset. ........................................
4.3 Simulation example of (a) first severity and (b) second severity of diplegic gait in the GAIT-IT dataset. ........................................
4.4 Simulation example of (a) first severity and (b) second severity of diplegic gait in the GAIT-IT dataset. ................................................................. 36
4.5 Simulation example of (a) first severity and (b) second severity of diplegic gait in the GAIT-IT dataset. ................................................................. 37
4.6 (a) Front view background frame (b) respective HSV colour histogram. ......................... 37
4.7 (a) Side view background frame (b) respective HSV colour histogram. ......................... 38
4.8 (a) HSV histogram of front view frame (b) respective front view frame (c) resulting binary silhouette with red arrows pointing to small noise clusters (d) binary silhouette after morphological operations. .................................................. 38
4.9 (a) HSV histogram of side view frame (b) respective side view frame (c) resulting binary silhouette with red arrow pointing to small noise cluster (d) binary silhouette after morphological operations. .................................................. 39
4.10 CNN architecture from [55]. .................................................................................. 39
4.11 Pose output format of detected body parts using OpenPose [97]. ............................... 40
4.12 (a) Rendered skeleton on a front view frame (b) front view binary skeleton (c) Rendered skeleton on a side view frame (d) side view binary skeleton. ......................... 41
4.13 Bounding box around a detected silhouette in the (a) front view and (b) side view; resized results are presented in: (c) front view and (d) side view. ......................... 41
4.14 Illustration of the method to identify complete cycles from a gait sequence. (a) Representation of the step width measured in a binary frame. (b) Plot of the step width along the frames in a gait sequence, highlighting the maximum distance peaks. Two gait cycles can be determined for the analysed sequence: one cycle between the first and third peaks and a second cycle between the third and fifth peak. ................................. 42
4.15 Examples of different binary silhouette lower portions considered to measure the step width, represented as a green rectangular box. The blue vertical lines represent the left and right foreground edges within those fractions. (a) Binary silhouette from a diplegic gait sequence (b) binary silhouette from a neuropathic gait sequence. ................................. 43
4.16 Example of (a) binary silhouettes in a gait cycle (b) GEI computed from this gait cycle (c) binary skeletons in a gait cycle (d) SEI computed from this gait cycle. ................................. 43
5.1 Illustration of the VGG-19 architecture [69] for the ImageNet challenge [57]. ................ 46
5.2 GAIT-IST 10-fold average training and validation accuracy history using (a) the GEIs as inputs and (b) the SEIs as inputs. ............................................. 47
5.3 GAIT-IT 10-fold average training and validation accuracy history using (a) the GEIs as inputs and (b) the SEIs as inputs. ............................................. 47
5.4 GAIT-IT and GAIT-IST 10-fold average training and validation accuracy history using (a) the GEIs as inputs and (b) the SEIs as inputs. ............................................. 48
6.1 General architecture of the proposed system, divided into 4 main modules. ................ 52
6.2 Example of 9 key silhouette frames representing a complete cycle in a normal gait sequence, along with the eight gait phases in between each consecutive frames. ................................. 53
6.3 Example of 9 key silhouette frames representing a complete cycle in a Parkinsonian gait sequence, extracted at the relative moments determined for normal gait. .......................... 53
6.4 Feet distance plot along the frames highlighting the extracted 9 key frames in a (a) normal gait sequence and a (b) Parkinsonian gait sequence. ............................... 54
6.5 Illustration of the ResNet residual building block [67]. ........................................... 55
6.6 Illustration of the Xception network architecture [66]. ........................................... 55
6.7 Illustration of the temporary effect of dropout in a neural network [103]. .................... 57
6.8 Illustration of the bidirectional LSTM implemented in the proposed system. ............... 60
6.9 Illustration of the proposed PGait CNN architecture for feature extraction in pathological gait classification. ................................................................. 63
6.10 Plot of the 10-fold average (a) training and test accuracies and (b) training and test loss values, along 50 epochs, corresponding to the training process of the proposed PGait CNN for pathological gait classification. ......................................................... 63
7.1 Examples illustrating persistent segmentation errors on the binary silhouettes of some subjects in GAIT-IST [94]. ................................................................. 68
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Mean values of cadence, speed, stride length and stride duration from [21]</td>
<td>8</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of currently available pathological gait datasets</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Openpose keypoint output labels</td>
<td>40</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of all classification accuracies across all datasets using GEIs and SEIs as inputs and re-trained on VGG-19.</td>
<td>48</td>
</tr>
<tr>
<td>6.1</td>
<td>Classification accuracies on the ImageNet challenge [57] and number of parameters of the 4 considered CNNs.</td>
<td>54</td>
</tr>
<tr>
<td>6.2</td>
<td>10-fold cross-validation average classification accuracies of the proposed CNN-LSTM system, using different convolutional bases from top performing pre-trained CNNs.</td>
<td>58</td>
</tr>
<tr>
<td>6.3</td>
<td>Confusion matrix containing the classification accuracies of the proposed system on the GAIT-IT dataset, averaged from the 10-fold cross-validation. These results are obtained by fine-tuning the VGG16 convolutional base of the CNN-LSTM system. The overall accuracy was 96.2%.</td>
<td>59</td>
</tr>
<tr>
<td>6.4</td>
<td>Confusion matrix containing the classification accuracies of the proposed system on the GAIT-IT dataset, averaged from the 10-fold cross-validation. These results are obtained by fine-tuning the VGG19 convolutional base of the CNN-LSTM system. The overall accuracy was 96.2%.</td>
<td>59</td>
</tr>
<tr>
<td>6.5</td>
<td>Confusion matrix containing the classification accuracies of the proposed CNN-BiLSTM system on the GAIT-IT dataset using the fine-tuned VGG16 as the convolutional base. The results are computed as the average accuracies of the 10 folds from the cross-validation and the overall accuracy was 96.5%.</td>
<td>61</td>
</tr>
<tr>
<td>6.6</td>
<td>Confusion matrix with the classification accuracies of the CNN-BiLSTM system on the GAIT-IT dataset using the proposed CNN as the convolutional base. The results are computed as the average accuracies of the 10 folds from the cross-validation and the overall accuracy was 93.8%. It represents a state-of-the-art classification rate achieved with the proposed CNN for gait analysis on 224 × 224 binary frame inputs.</td>
<td>64</td>
</tr>
<tr>
<td>6.7</td>
<td>Comparison of the number of parameters among the different systems presented since Section 5. The CNN-BiLSTM system using the proposed convolutional base achieves a reduction of 36 to 41 times in the number of parameters and model size in megabytes.</td>
<td>64</td>
</tr>
<tr>
<td>6.8</td>
<td>Training and execution duration compared among the different systems presented since Section 5. The CNN-BiLSTM system with the proposed convolutional base is 6 to 8 times faster in training and 6 times faster in yielding predictions than its counterparts.</td>
<td>65</td>
</tr>
<tr>
<td>7.1</td>
<td>Summary of the cross-validation results on GAIT-IT, obtained with the best solutions described in Section 5 and in Section 6. The highest classification rate of 96.5% was achieved by the CNN-BiLSTM with a pre-trained and fine-tuned VGG16 convolutional base.</td>
<td>68</td>
</tr>
</tbody>
</table>
7.2 Summary of the cross-dataset results using GAIT-IT as the training set and GAIT-IST as the testing set. The classification accuracies were obtained with the best solutions described in Section 5 and in Section 6. The highest classification rate of 91.4% was achieved by the CNN-BiLSTM with a pre-trained and fine-tuned VGG16 convolutional base.

7.3 Classification accuracies according to each gait type obtained by the proposed CNN-LSTM systems and the state-of-the-art VGG19 solution.

7.4 Confusion matrix containing the overall 10-fold cross-validation classification accuracies averaged across every system considered in Table 7.3.
**Nomenclature**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>1 Dimensional</td>
</tr>
<tr>
<td>2D</td>
<td>2 Dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>3 Dimensional</td>
</tr>
<tr>
<td>AFI</td>
<td>Average Feet Image</td>
</tr>
<tr>
<td>AOM</td>
<td>Amount Of Movement</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>Bidirectional Long Short-Term Memory</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>ConvLSTM</td>
<td>Convolutional Long Short-Term Memory</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyography</td>
</tr>
<tr>
<td>FC</td>
<td>Fully Connected</td>
</tr>
<tr>
<td>FCCN</td>
<td>Fundação para a Computação Científica Nacional</td>
</tr>
<tr>
<td>FCT</td>
<td>Fundação para a Ciência e Tecnologia</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>GEI</td>
<td>Gait Energy Image</td>
</tr>
<tr>
<td>GTI</td>
<td>Gait Texture Image</td>
</tr>
<tr>
<td>ILSVRC</td>
<td>ImageNet Large Scale Visual Recognition Challenge</td>
</tr>
<tr>
<td>ISCTE-IUL</td>
<td>ISCTE-Instituto Universitário de Lisboa</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short-Term Memory</td>
</tr>
<tr>
<td>MEI</td>
<td>Motion Energy Image</td>
</tr>
<tr>
<td>METI</td>
<td>Mestrado em Engenharia de Telecomunicações e Informática</td>
</tr>
<tr>
<td>MHI</td>
<td>Motion History Image</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MSI</td>
<td>Motion Silhouettes Image</td>
</tr>
<tr>
<td>MSN</td>
<td>Multi-stream Network</td>
</tr>
<tr>
<td>PAF</td>
<td>Part Affinity Fields</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>SEI</td>
<td>Skeleton Energy Image</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TILT</td>
<td>Transform Invariant Low-rank Textures</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>
1 Introduction

In this chapter, the context of this work and the motivation for the development of a gait analysis system is presented. The problem definition is described followed by the main objectives and the structure of this work.

1.1 Context and Motivation

Biometrics is a term related to the analysis and measurements of the human body based on biological and/or behavioral characteristics. Common biometric traits include fingerprint, voice and eye attributes (retina and iris), typically used in security and data protection applications. Gait can be defined as any method of locomotion involving periodic body movements, such as sequences of loading and unloading of the limbs [1]. Therefore, gait is a behavioral biometric trait. It relates to patterns in the performance of activities, as opposed to physiological traits that refer to inherent physical attributes of the human body. Although it usually refers to walking, actions like running, swimming and jumping can fit in the same definition. In the case of walking, a gait cycle starts with the first contact of one foot on the ground, followed by a forward swing motion of each leg separately and ending with the next initial contact of that same foot. Gait can be used as a biometric trait, since the movements involved in locomotion vary between individuals as a function of their skeletal dimensions, muscular composition and body mass, among other physiological characteristics. The distinct combination of these attributes leads to the possibility that in theory, gait is unique to each individual [2]. However, several factors influence the way one walks and can produce considerable gait variations within each person. Examples of these factors are walking surface, footwear, carrying loads (backpack, purse, etc.), fatigue, injury and emotional state.

The information obtained from the study and analysis of gait has a wide range of applications (see Figure 1.1). In medicine, it can contribute to the diagnosis and monitoring of treatment of gait related pathologies [3]. In surveillance and forensics, gait recognition can be done non-intrusively and from a distance, even with poor image resolution, using current image processing techniques [4]. Besides not requiring the subject's cooperation, gait is hard to conceal which is an advantage over traditional biometric traits like the face, that can be hidden with a mask, and the fingerprint, which can't be found if the subject wears gloves [5]. Gait analysis is also a field of interest in sports where it can be used to improve training and evaluate performance.

![Figure 1.1: Gait analysis applied to (a) clinical practice [6], (b) surveillance [7] and (c) sports [8].](image)
Interest in human posture and motion is at least as old as the period of ancient Egypt, as manifested for instance in the art of representing people in historic events, religious stories and performing activities like agriculture and construction. In ancient Greece, the study of body movements was mainly in the context of fighting and sports techniques, while other applications include sculptures portraying movement, as the example in Figure 1.2, which also demonstrates knowledge in human kinematics.

![Figure 1.2: Marble copy of The Discobulus by Myron [9], portraying body movement.](image)

The first records regarding gait analysis can be traced as far back as 350 BC, written by Aristotle. However, it was not until the 17th century that the study of human locomotion had further advancements through the work of Giovanni Borelli, describing the displacement of the body's centre of gravity, to maintain balance in forward motion [10]. Later in 1836, the German brothers Wilhelm and Eduard Weber made significant progress in the quantitative measurement of locomotion and performed experiments on individuals walking and running under different conditions [11]. Following major breakthroughs in photography, the technique of capturing motion in several frames, chronophotography, started being used in medical research. This opened many possibilities in motion analysis, so far mainly limited to human observation. Étienne-Jules Marey developed the chronophotographic gun, capable of capturing twelve frames in a second and record them on a single photographic plate, which he applied in the study of pathological locomotion. One of Marey's photographs of motion is illustrated in Figure 1.3. Albert Londe, working at Salpêtrière hospital in Paris, constructed a camera with twelve lenses to record the movements of many patients with gait related disorders, gathering a significant amount of images displaying different cases of pathological gait [12].

![Figure 1.3: The Running Lion Tamer, chronophotograph by Marey [13].](image)

The work of Braun and Fisher (1895-1904) introduced three-dimensional gait analysis, with methods
to determine the trajectory of the centre of gravity, the kinetics involving the foot and the kinematics of the leg in the swing phase [14]. The combination of technological advancements and needs in medical care that followed the two World Wars was a stimulus for research in biomechanics. In 1953, Inman, Eberhart and Saunders described, in a collaborative work, the primary determinants in normal gait and their assessment in relationship to pathological gait [15]. In 1966 Patricia Murray, a renowned researcher and physical therapy teacher, started a series of collaborations responsible for many important publications regarding clinical gait, including contributions in the study of walking patterns in children, healthy aged people and individuals with specific neurological disorders such as Parkinson’s disease [16–18].

1.2 Problem Definition

Nowadays there are many ways to acquire gait data for analysis, from instrumented rooms to wearable sensors or vision-based systems. For instance, floor based sensors are able to measure pressure and forces exerted on the ground at each step but the amount of data that can be obtained in a single acquisition is limited to the length of the equipment where such sensors are integrated, usually in the form of floor mats. Methods using these sensors require dedicated spaces and specialized technicians to setup the equipment and conduct the acquisition. Wearable sensors provide gait data acquisition over large periods of time without the need for constrained environments. Inertial sensors belong to this category and can measure the acceleration and orientation of the moving body parts. However, due to the fact that such sensors are in direct contact with the body, they require a previous setup often done by specialized personnel and can affect the natural gait of the subject. Acquisition methods using vision-based sensors are unobtrusive and do not require the cooperation of the subject to obtain gait data (or at least the cooperation is much simpler, e.g. limited to walking along a predefined path). Like floor based sensors, these are also limited in the amount of consecutive information that can be obtained from each subject by the angle of view of the optical sensor.

For gait analysis in a clinical context, a professional is required to interpret the obtained data in order to provide a diagnosis, independently of the acquisition method. Therefore, significant contributions in the field of clinical gait analysis can arise from solutions that are able to obtain motion data and automatically provide an initial classification of the observed gait across different gait pathologies. The problem addressed in this work is the development of such systems using deep learning techniques to tackle one of the main difficulties, which is the extraction and classification of discriminant and meaningful features from a given gait representation.

1.3 Objectives

The developed work will present a review of state-of-the-art systems in the context of pathological gait analysis considering their contributions and limitations. Focused on an unobtrusive approach regarding data acquisition, a proposed framework will consider and combine different deep learning methods of feature extraction to perform pathological gait classification on 2D video-based gait data. The acquisition of gait sequences with subjects simulating different pathologies will increase the available data to train
the developed deep learning methods.

1.4 Contributions

This work explores deep learning solutions for the automatic analysis and classification of human walking into different gait related pathologies or healthy gait. Its main contributions can be described as follows:

- Development of a new publicly available pathological gait database. Comprising 21 subjects performing their normal gait and simulating 4 different gait related pathologies with 2 levels of severity, it is characterized as the largest database of 2D video data currently available, with different gait representations, namely, binary silhouettes, binary skeletons, Gait Energy Image (GEI) and Skeleton Energy Image (SEI).

- A new framework for pathological gait classification to process video gait sequences as collections of key frames in each complete gait cycle. It comprises a proposed module for key frame selection followed by a feature extraction module that combines spatial and temporal analysis through the use of convolutional and recurrent neural networks.

- A proposed convolutional neural network (PGait CNN) built for feature extraction in gait analysis using binary images. PGait CNN is combined with a bidirectional LSTM recurrent network, to obtain a powerful pathological gait classification solution. Taking advantage of the developed database, the proposed solution presents an alternative of significantly reduced complexity in comparison with the ones available, which repurpose pre-trained models built for similar tasks to deal with the lack of available gait data.

1.5 Thesis Structure

This work is divided in eight chapters. In this first chapter, the context and motivation are introduced, followed by the problem definition, objectives and contributions of the proposed work. In the second chapter, a theoretical background to gait analysis is provided, describing the main concepts and different gait related pathologies, mentioning those with greatest interest to the research community, and which should be addressed by the proposed solution. In the third chapter, gait data acquisition systems are presented and grouped according to the type of sensors used. State-of-the-art systems will be reviewed and categorized according to the main modules of a general gait analysis system architecture. The developed pathological gait database is presented in chapter four, along with a description of the currently available alternatives. In the fifth chapter, a state-of-the-art deep learning solution is described and used as the baseline to validate the advantages of the new database, with the largest amount of high quality video-based gait data among its counterparts. The sixth chapter will present the proposed solutions, describing the main components of the pathological gait classification system and the experiments conducted to continuously improve its performance. The experimental results are presented in chapter seven, comparing the proposed framework with state-of-the-art solutions in a compilation of cross-validation and cross-dataset tests. Finally, in the eight chapter, the main achievements are summarized along with the primary considerations for future work developments.
2 Background: Pathological Gait Analysis

In this chapter, the fundamentals of gait analysis are discussed. In Section 2.1, the basic main concepts are described, providing the theoretical context for the study of human locomotion. Section 2.2 introduces the concept of pathological gait, referring the main categories in which it can be classified. Finally, Section 2.3 reviews the primary metrics used to evaluate the performance of methods for gait analysis.

2.1 Gait Analysis Concepts

This section presents the fundamental concepts of gait, essential to its qualitative and quantitative assessment. Section 2.1.1 characterizes the key events in a gait cycle, as well as their sequential relations. In Section 2.1.2, the main measurable attributes of motion are discussed.

2.1.1 Gait Cycle Components

The motion patterns involved in human locomotion are associated to a sequence of events that characterize a gait cycle. Jacqueline Perry, an influential researcher in normal and pathological gait, proposed the division of the gait cycle into eight phases, according to their functional significance [19]. An illustration of this division is displayed in Figure 2.1. This type of approach allows for a clearer understanding of the interactions between body parts, and their relative impact in performing the functional objective of each phase. Such method of analysing the walking pattern is important, in order to determine the effects of gait impairments and the relation with the pathologies that may cause them.

The functional division of gait is hierarchical, starting with two general phases, which alternate between the two lower limbs:

- **Stance phase**: It begins when the respective foot first touches the ground (heel contact) and ends when that same foot leaves the floor (toe off). This phase accounts for an approximate 60% of the whole cycle, with two periods of double support where both feet are on the ground, and a period of single support where only the observed foot is on the ground.

- **Swing phase**: Consists of the forward motion by the lower limb. It starts when the foot leaves the ground (toe off), the final moment of the stance phase, and ends with the next contact of that same foot with the ground (heel contact). During this whole period, all the body weight is supported by the contralateral lower limb, and it stands for approximately 40% of the gait cycle.

The stance phase can be further divided in five parts:

1. **Initial contact**: This is the moment the foot first touches the ground, which starts the first double support period. The hip is flexed, the knee is extended, and the heel is the first part to make ground contact, which is the key event in this phase often referred as heel strike or heel contact.

2. **Loading response**: This phase has the objective of absorbing the shock from the weight being transferred to the observed limb. Beginning right after the initial floor contact, the knee is flexed while the hip starts to extend as the upper body is displaced towards the limb. It ends when the contralateral foot leaves the ground.
3. **Mid stance**: This phase is characterized by the functional objective of stabilizing all the weight over the supporting limb. The hip continues to extend until the body lines up with the standing foot, right before the heel starts to rise.

4. **Terminal stance**: This period of this phase starts with the heel leaving the ground and goes until the contralateral foot makes initial floor contact. It serves the purpose of moving the body forward, past the standing limb.

5. **Pre-swing**: When the contralateral foot touches the ground, the weight starts being transferred to the opposite limb until the observed foot leaves the ground, often referred to as toe-off. This is also the second and terminal double support period.

Finally, the swing phase is composed by three sub-phases:

1. **Initial swing**: Beginning when the foot leaves the ground, this period consists of its advancement through hip flexion until it reaches the opposite foot, while also increasingly flexing the knee.

2. **Mid swing**: The flexed knee begins to extend and the limb in forward motion moves past the whole body. This phase ends with the tibia in a vertical plane.

3. **Terminal swing**: The movement of the limb reaches its final stage, as it prepares for foot contact with the floor through further knee extension.

![Figure 2.1: Division of the gait cycle in eight functional phases, according to [19].](image)

### 2.1.2 Gait Parameters

The quantification of gait attributes allows the measurement of variations between individuals, and within the same individual. With sufficient data, an acceptable standardization of certain features can be achieved. In a clinical context, this is fundamental to classify a patient's gait in terms of disability, against established references, and to monitor the progress and effectiveness of treatment [20].

The features that characterize gait can be divided into several categories, such as:

- Kinematic features that describe the motion of the body and its parts.
- Kinetic features such as the ground reaction force exerted by foot during motion.
- Electrical activity produced by the muscles. Electromyography (EMG) is the medical technique for recording and evaluating such electrical signals.
- Pressure exerted by the different areas of the foot, when in contact with the floor.
Body posture and symmetry between each side in locomotion.

Kinematic features obtained in gait analysis can be divided into spatial, temporal and spatiotemporal parameters. The primary spatial parameters are described as follows, along with an illustration in Figure 2.2, regarding those in relation to feet positions.

- **Step length**: The distance between the heel contact of one foot and the following heel contact of the opposite foot.
- **Stride length**: The distance covered by two heel contacts of the same foot.
- **Step width**: The distance between both heels, measured in a line perpendicular to the movement direction.
- **Joint angles**: Describe the angles measured between two body parts as line segments, connected by the same joint. Other angular parameters can involve determining the angles between body parts and other references, such as the angle between the foot palm and a vertical line.
- **Step angle**: Measures the angle between the movement direction and the line that passes through the heel and the index toe. It is also referred as the degree of toe in, if a positive angle represents the index toe being between both heels, and as degree of toe out if otherwise.

![Figure 2.2: Spatial parameters involving feet distances.](image)

Temporal parameters used to describe gait include:

- **Step duration**: The time elapsed between consecutive heel contacts from both feet. This can be used to determine foot symmetry, which represents the duration of the step from each side as a percentage of the gait cycle.
- **Stride duration**: The time elapsed between consecutive heel contacts of the same foot.
- **Stance duration**: The duration of the stance phase. It can also be divided in single and double support duration.
- **Swing duration**: The duration of the swing phase.

The spatiotemporal parameters result from the combination of features from space and time domains, such as:

- **Cadence**: The number of steps during a certain time period. Often represented in steps per minute.
- **Speed**: The distance covered in a time period, usually measured in meters per second.
- **Joint angular speed**: The rate of change of joint angles during a given time period, usually
measured in radians per second.

As an example, a study published in 2011 [21] obtained measurements of spatiotemporal gait parameters from a considerable large amount of test subjects, using a mobile inertial sensor based system, in which each sensor contained three accelerometers and three gyroscopes. Table 2.1 presents mean values of walking speed and cadence, and of stride time and length, obtained from 1860 subjects, with ages ranging from 5 to 100 years old.

<table>
<thead>
<tr>
<th>Speed (meters/second)</th>
<th>Cadence (steps/minute)</th>
<th>Stride Length (meters)</th>
<th>Stride duration (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>1,33</td>
<td>119</td>
<td>1,42</td>
</tr>
<tr>
<td>Men</td>
<td>1,30</td>
<td>113</td>
<td>1,47</td>
</tr>
</tbody>
</table>

Table 2.1: Mean values of cadence, speed, stride length and stride duration from [21]

2.2 Pathological Gait Classification

The ability to walk requires the proper work of many biological systems, such as the nervous, muscular and skeletal systems. Thus, several health conditions can cause gait impairments that are manifested in distinct ways as a function of their primary pathology [19] (see Figure 2.3). Medical research on such conditions, provides information that makes the connection between diseases and specific locomotive disorders. Categorizing the different forms of motion impairment, allows for a better assessment of their underlying cause. The more common types of pathological gait can be categorized as follows [22, 23]:

- **Hemiplegic Gait**: It is a unilateral incapacitation, characterized by the arm being flexed and held close to the torso, the leg in extension and the foot and respective toes also flexed, pointing upwards. While walking, the patient’s leg on the affected side is dragged in a circular motion instead of properly lifted. This pathology is commonly a consequence of a stroke episode.

- **Diplegic Gait** (Scissors gait): Both sides of the body are affected during locomotion, with the effects being more severely reflected in the lower limbs than in the upper ones. The hips are flexed, the knees are bent and close together, and the feet are extended and rotated inwards. Walking involves circumduction of the legs due to foot drop by weakness in the distal muscles, which in more severe cases causes the legs to cross in a scissors resembling movement. This type of pathology is often caused by brain damage, such as cerebral palsy.

- **Neuropathic Gait** (Steppage Gait): The most common attribute of this type of pathological gait is referred to as foot drop, as a lack of flexion strength causes the toes to drag on the floor. Exaggerated lifting of the respective leg (high stepping) if often seen as an attempt to prevent this condition.

- **Myopathic Gait** (Waddling Gait): Characterized by weakness on the hip muscles, causes the pelvis to drop on the opposite side of the affected muscles. A waddling motion is often reported when both sides are affected, with short steps in a swaying motion.

- **Choreiform Gait** (Hyperkinetic Gait): Patients with this type of gait exhibit irregular, involuntary and abrupt movements, typical manifestations of Huntington’s Disease.
• **Ataxic Gait** (Spastic Gait): This gait type is usually referred as unsteady and uncoordinated, comparable to one caused by alcohol intoxication, commonly seen in cerebellar disease.

• **Parkinsonian Gait** (Propulsive Gait): Attributes of this type of gait are a stooped and stiff posture and small steps. It is also characterized by erratic body shaking and difficulty initiating movement.

• **Sensory Gait**: The loss of perception regarding the position and sensorial input from the foot, characterizes this gait with exaggerated movements, such as slamming the foot onto the floor, in order to compensate for this lack of awareness [24].

![Illustration of characteristic movements in different types of impaired gait](image)

In the work to be developed, 4 of these 8 types of pathological gait will be considered for their relevance among the scientific community, namely, hemiplegic, diplegic, neuropathic and Parkinsonian. Testing a system for pathology classification depends on available datasets and on the amount of data they provide. The decision of focusing on these pathologies is also based on the DAI-2 [26] and GAIT-IST databases [27] which provide with gait sequences of subject performing their normal gait as well as sequences where these subjects simulate each of these types of gait pathologies. In the future work, the acquisition of new gait sequences will have the objective of extending this existing database.

### 2.3 Evaluation Metrics for Classification Performance

Pathological gait classification is a multi-class classification problem. As mentioned in 2.2, it consists of outputting predictions that classify the observed gait across different gait related pathologies and normal gait. The performance of such systems is usually described with a confusion matrix, as illustrated in Figure 2.4, which summarizes the prediction results against the ground truth of the input samples. A classification can be referred as Positive if it corresponds to an observed class or Negative otherwise. These results can be further categorized in 4 different types, namely, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN):

- **TP**: The system outputs a class prediction that corresponds to the actual class of the input sample.
- **TN**: The predicted class for a sample and the respective ground truth are different from the observed class.
- **FP**: The prediction belongs to the observed class but the ground truth corresponds to a different
• **FN**: The ground truth of the sample corresponds to the observed class but the systems outputs a prediction belonging to a different class.

Using the confusion matrix, different performance metrics can be computed to evaluate the classification results of a given system.

Precision can be defined as the proportion of Positive predictions that are correctly classified. It focuses on the ability of the system to produce predictions for a class that actually correspond to the ground truth. Precision can be defined as:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  

(2.1)

Recall measures the proportion of ground truth Positives that are correctly identified by the system. In the case of gait classification, optimizing the system according to this metric is to focus on detecting as many patients as possible with a given pathology with some tolerance for healthy patients being wrongly classified, as opposed to optimizing Precision which focuses on producing predictions that are actually correct and may tolerate missing some samples of pathological gait. Recall can be defined as:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]  

(2.2)

The accuracy metric can be used to measure the overall proportion of correct classifications taking in account both True Positive and True Negative classifications. It can be defined as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]  

(2.3)

Each metric is used to evaluate different aspects of a system’s classification performance, which
means that improving its results involves optimizing the system according to a trade-off between different metrics that depend on the context and purpose of the classification task. Tolerating some types of errors can allow the system to be optimized for specific performance characteristics.
3 State of the Art: Gait Analysis Systems

This chapter provides a review of modern gait analysis systems, with a particular focus on solutions using computer-based image processing, and those considering machine learning techniques, such as deep learning, for feature extraction and classification purposes. It provides a survey on prominent methods and analysis systems, describing the main steps from collecting gait data to reaching context related classifications.

3.1 Gait Data Acquisition

Nowadays, a rich variety of gait data can be obtained through the use of different types of sensors, providing the means to measure several types of gait parameters, like those referred in 2.1.2. Modern gait data acquisition can be classified into three main types of systems [28]:

- **Floor sensors based**: These systems are composed of sensors placed along the floor, in the form of a walking platform. The basic analysis done with such systems consists in recording the feet positions over time, directly extracting features such as step width, the length and duration of each step and of the full gait cycle. More advanced systems can be distinguished as force platforms, able to measure force vectors applied in each step, and pressure platforms that can differentiate the pressure exerted on each area under the foot [20].

- **Wearable sensors based**: Different types of sensors can record three-dimensional information related to walking patterns. Sensors like accelerometers and gyroscopes, measure the acceleration, velocity and orientation of the moving object. Electromyography sensors record muscular electrical activity. Wearable sensors can be used to capture information from specific body parts, and have the advantage of portability, allowing the continuous measurement and data report through various activities outside dedicated and constrained spaces. Moreover, most of the common personal electronic devices currently integrate some of these sensors, supporting many gait analysis applications [29]. One aspect to take into consideration is the fact that these sensors may condition the natural movements of the subject, depending on how they are used, which represents a possible disadvantage of such systems when compared to unobtrusive solutions. Some wearable systems include force and pressure sensors placed under the foot, implementing some of the functions described for the floor sensor based systems.

- **Vision based**: Optical sensors can be used to acquire image sequences of the walking person. Using a single camera, 2D data can be acquired from which a wide range of gait features can be extracted, after processing the captured images with techniques such as background subtraction and image segmentation. By using depth cameras, the distance between objects and a specific viewpoint can be measured, allowing to map the image points to a three-dimensional space. This can also be achieved with stereo vision systems, using two or more cameras with different viewpoints, by finding key features in the images and corresponding them between the different cameras. The identification of key features and specific body parts can be improved by the use of active or passive markers, which emit or reflect light signals respectively.
The diversity of gait data provided by floor sensors based systems is relatively limited in terms of the different gait parameters that can be extracted by such systems. The equipment used is usually associated with constrained spaces, dedicated to the acquisition process. Wearable sensor based systems have the advantage of portability which allows for the acquisition of data over longer periods of time and can be used in many types of applications. However, their performance can be conditioned by the way they are placed, adding to the fact that their use can affect the natural gait of the subject. Vision based systems for gait analysis can be unobtrusive and do not require subject cooperation or interaction with any type of equipment. Therefore, this type of system is suited for applications in less constrained environments, such as the integration of gait analysis in a clinical context. In the work to be developed, a gait analysis system will be proposed, based on image data captured from a single 2D camera with a simple setup.

### 3.2 Gait Analysis Systems Architecture

The general architecture of a biometric analysis system comprises three main modules, namely: Pre-processing, Feature extraction and Classification, as illustrated in Figure 3.1.

![Figure 3.1: Main modules of a biometric analysis system.](image)

The main goals of each of these modules can be described as follows:

- **Pre-processing**: The function of this module concerns the treatment of gathered biometric data, to assure its quality before analysis and obtain a relevant representation of the biometric trait. Acquiring data is always associated with some degree of error and uncertainty, involving noisy data, missing values and out-of-bounds measurements [30]. Also, processes like normalization and transformation of data may be required, for example, in cases where values can have different scales. This means that analysing raw information is prone to misleading results, and this challenge is particularly important in machine learning applications, designed to recognize patterns in a dataset and come up with statistical predictions, that rely on valid and meaningful data [31].

- **Feature extraction**: The purpose of this module is to extract relevant information from the obtained data and determine its distinguishing attributes. This can be associated to dimensionality reduction [32], as often inputs are intractable and a large portion of its values are redundant, which requires the reduction of the data into representative features, for its use in practical applications. Subsequently, a good implementation of feature extraction will lead to a better system performance and significant reductions in expended time and computational power.

- **Classification**: It consists in determining the category to which an observation belongs, given a
set of features. A classifier can be implemented, for instance, through algorithms such as Linear regression, Support Vector Machines (SVM), Decision Trees, Linear Discriminant Analysis (LDA), and Artificial Neural Networks. Regarding pathological gait analysis, the classification process consists of associating the quantifiable properties of an individual’s gait to a certain type of gait-related pathology, like those referred in Section 2.2.

The performance of a gait analysis system depends on the combination of methods used for each of these modules. In this section, state-of-the-art contributions will be reviewed and categorized according these modules. The presented solutions will be considered for the development of the proposed system for pathological gait classification.

3.2.1 Pre-processing: from Video to a Suitable Gait Representation

As mentioned in section 3.2, one of the main objectives with the pre-processing module of a gait analysis system is yielding meaningful gait representations from which valuable features can be extracted. Typically, the original images are not suitable for analysis, and background subtraction is usually applied on the acquired image sequences, in order to isolate the observed subject. By first extracting data from the background, the static part of the scene can be acquired, defined by values obtained for depth, colour, light intensity or some combination of these attributes. Then, for sequences containing moving objects, this background model will be subtracted to obtain a binary mask, representing the foreground object [33]. In the case of gait analysis, the observed object is usually a walking individual, and the obtained mask is a human silhouette. For gait analysis in unconstrained environments, some challenges may arise for this task, such as being able to isolate the background, accounting for undesired foreground objects in the acquired sequences and dealing with light reflections and the presence of shadows. In [34], a method based on optical flow is used to improve foreground detection in crowded environments. Figure 3.2 illustrates the described process and serves as an example of a subject’s shadow affecting the detected foreground.

![Figure 3.2: Components of background subtraction [35]. (a) background, (b) moving subject in foreground, (c) binary mask of the walking person.](image)

Regarding vision based analysis, the representations resulting from processing the biometric data can be categorized according to two different approaches: model based and appearance based [36].
Appearance Based

These methods produce gait representations that do not contain any prior knowledge of human motion. Usually, features yielded for extraction with this type of approach are based on binary silhouette shapes, which result from the previously mentioned process of background subtraction. The approach in [37] generates two temporal motion templates illustrated in Figure 3.3, the Motion Energy Image (MEI) and the Motion History Image (MHI), combining them for action recognition. The MEI is a binary cumulative image where a pixel is activated for the current frame depending on its value along a predefined number of previous frames. The MHI implicitly describes the direction of motion by defining the intensity of a pixel as a function of the temporal history of the movement. It generates a grey-level image where brighter pixels represent activity in more recent frames.

![Figure 3.3: MEI and MHI examples. (a) Key movement frame, respective (b) MEI and (c) MHI [37].](image)

Inspired by the MHI representation, the work in [38] proposed the Motion Silhouettes Image (MSI), also a grey-level image, with a reduced computational cost. The MSI is calculated for each frame, iterating through the sequence images from the second to the last, given that each MSI is a product of the pixel values from the current and previous frame.

A widely used representation is the Gait Energy Image (GEI) [39]. Each binary silhouette obtained from an image sequence of a gait cycle is cropped, normalized in size and horizontally aligned, to assure that every frame is centred and has the same dimensions. The pre-processed shapes are then used to compute the GEI, defined as follows in equation 3.1:

\[
GEI(x, y) = \frac{1}{N} \sum_{i=1}^{N} B_i(x, y)
\]  

(3.1)

where \(N\) represents the number of frames in one gait cycle, or in an integer number of gait cycles, and \(B_i(x, y)\) a binary silhouette image, with \(x\) and \(y\) as its pixel coordinates. The result is a grey-level image with the mean values of each pixel across all frames, comprising motion features of a full gait cycle in a single image. This representation is robust against noise in individual frames, as it is shown in Figure 3.4, where particular errors in the data affecting some regions and edges are minimized in the final image.
Based on this representation, the detection of specific gait impairments can be improved by focusing on particular areas of the GEI to retrieve features [40], such as the head and chest for visually impaired gait, and the leg region for leg related gait impairment.

The methods described in [41] are used to compute gait features for abnormal gait detection. The stance phase, swing phase and step length are obtained from the binary silhouette frames, while the intensity, representing the amount of movement in a specific area of the GEI, and the amplitude, representing the limb movement's broadness, are obtained from a GEI. Using the binary silhouettes, a gait cycle can be determined by tracking the distance between the heel of the back foot and the toe of the front foot [42]. This distance peaks twice in every gait cycle, one time for each step, meaning that one cycle occurs for every two consecutive peaks of this distance. This work also produced one of the few available gait databases, the INIT database, with subjects performing a normal gait as well as simulating several types of gait impairments.

In [43], the developed system performs automatic detection and classification of gait impairments, based on two types of features: feet-related and body-related features. In some gait impairments, self-occlusions can occur due to a significantly shorter step size, posing a challenge in the detection of key events in a gait cycle, such as the initial contact, thus preventing the estimation of certain gait parameters. To tackle this problem, this work proposes the detection of “foot flat” instants, where the foot is in complete contact with the ground. An average feet image (AFI) is computed by averaging feet images, extracted from the lower 10% part of the silhouettes, corresponding to an interval between two initial contacts, estimated as in [42]. The AFI highlights the foot when it is in the foot flat position, which can be extracted by applying the Otsu thresholding method. With these positions determined along a video sequence, feet-related features are extracted, namely, step length, normalized step count, normalized speed and foot flat ratio. Regarding the body-related features, the extraction of the amount of movement (AOM) is described, allowing the system to capture movement restrictions while shifting weight onto the impaired side of the body. The system also computes the amount of shift in the centre of gravity in relation to the centre of base of support and the torso orientation, both providing crucial information on how posture can be affected by impairments. With a SVM classifier, a classification rate of 98.8% was achieved on the INIT database [41].

To address the problem of viewpoint changes in gait analysis for recognition, the Gait Texture Image (GTI) representation was proposed in [44]. It is computed in the same way as the GEI, with the difference that it uses the foreground binary mask directly, instead of the cropped and normalized silhouettes, producing an image resembling a chronophotography. This representation is then used as the input of

![Figure 3.4: Example of binary silhouettes and corresponding GEI [39].](image-url)
low-rank texture optimization. The Transform Invariant Low-rank Textures (TILT) method [45] is applied to transform a GTI from any viewpoint into a canonical viewpoint. The work in [46] describes a method for gait recognition using a subject’s shadow, which typically contains significant changes in orientation and perspective along the gait sequence. By obtaining GTI images from shadow silhouettes, the TILT method is applied to normalize them into a canonical view. The normalized shadows are then used to construct GEIs, which minimizes the impact of self-occlusions inherent to the observation from certain viewpoints, since only the visible portion of the shadow will be transformed in its canonical view. Besides recognition, viewpoint changes and non-lateral perspectives in unconstrained environments are also a challenge in gait classification, and analysing shadow silhouettes could prove to be an interesting method to apply in such conditions.

A recent work [27] used the GEI representation to perform gait classification over four different gait related pathologies and normal gait. With original video sequences containing multiple complete gait cycles, each one was identified by tracking the distance between both feet [42]. After this identification process, a GEI was computed for each gait cycle in a video sequence, as well as for the whole sequence. This work also developed a new dataset for abnormal gait classification, the GAIT-IST dataset, providing image data for the previously mentioned five types of gait.

A different approach [47] consists of obtaining the dynamic information of motion by deriving optical flow images from consecutive frames along the sequence. For each of these images, a set of features describing the spatial distribution of the flow can be computed, such as the centroid of moving points and moments of inertia.

**Model based**

Model based gait representations attempt to fit the observation to a model that involves prior knowledge of the human body (structural model) or its motion (motion model) [46]. This type of approach can be robust to background noise, occlusions and camera viewpoint changes, an especially challenging task in gait analysis. However, many methods rely on the accurate detection of certain key body points, meaning that the performance can be severely compromised if existing occlusions coincide with these points.

Fitting an image to a structural model often involves the computation of the correlation between the image and a human skeleton model, or between predefined shapes and the respective image body parts they represent. Using a skeleton human model usually involves the determination of body joints and the location of key body parts, such as the head and torso [48]. The skeletal model results from the connection of these points through line segments, as illustrated in Figure 3.5.
Applying a shape based body model can be achieved using edge correspondence, a measure of the difference between image and model shapes edges, and using region correspondence, which is a measure that takes into account the portion of the model shape that is overlapped with the acquired image and the portion that does not contain the image body part information [49]. An illustration of these two metrics is shown in Figure 3.6.

Motion models are used to provide a template that represents the dynamic characteristics of the body and its parts. This can be done by determining key points and tracking them from one frame to the next to derive motion parameters. The work in [50] presents a system to extract joint angle features from a human model based on the automatic determination of body joints in the image sequences. In [51], the positions of body joints were also used to extract temporal models of joint angles, and combine them with a structural model.

3D models can be developed using multiple calibrated cameras, or cameras equipped with depth sensors such as Kinect cameras. This approach is more robust to camera viewpoint and subject pose variations, since it involves recording motion from different perspectives. The approach from [52] combines 2D and 3D data in frontal gait analysis using the RGB and depth information captured with Kinect. It shows an improvement in recognition rate by taking in account for feature extraction, the 3D coordinates of the subject together with its frontal, top and side-views 2D projections. In [53], the Kinect V2 is
used in a setup where two of these cameras are placed with perpendicular view directions to create a 3D skeleton based database, from which a 3D model of human walking motion is constructed. It combines static features, namely distances between different joints, and dynamic features such as speed, stride length and variation of the body's centre of mass. This work achieved subject recognition accuracy rates on the created database between 88% and 94% for each of the five different viewpoints of 52 subjects. Gait analysis systems using 3D information were also applied to the task of gait classification. The method described in [54] focuses on abnormal gait detection using 3D skeleton model representations, and proposing a new spatiotemporal feature, the joint motion history feature.

The previously mentioned work in abnormal gait classification [27] also developed a new representation, the Skeleton Energy Image (SEI), using the OpenPose system [55] to extract the 2D coordinates of several key parts of the human body. By connecting these parts, a binary image of a skeleton can be obtained, as the one displayed in Figure 3.7. With a skeleton image for each frame, the SEI can be obtained with the same method used for the computation of the GEI. This approach presented an improvement in abnormal gait classification relative to the use of the GEI, arguing that the SEI does not represent the physical constitution of a subject, focusing on the dynamic characteristics of movement.

![Figure 3.7: (a) OpenPose result rendered on original image, (b) obtained binary skeleton image [27].](image)

The application of models in gait analysis requires additional computational resources, especially for 3D models, and often relies on quality image acquisition, as mentioned for the cases based on key point detection. Further disadvantages arise for the use of 3D models, due to the fact that the range of the depth cameras is relatively limited, and that image acquisition requires a constrained environment, previous camera calibration and a setup of multiple cameras that is hard to replicate in most application contexts.

### 3.2.2 Feature Extraction: Deep learning in Gait Analysis

Deep learning is a subtype of machine learning with a new take on learning data representations, based on learning successive layers of increasingly meaningful representations [32]. This can become the method of choice in classification problems, where the features to extract for such purposes are complex and have no intuitive relation with the input data [56].
Convolutional Neural Networks (CNN)

Since the last decade, the CNN model has been widely used for feature extraction in image recognition systems. Its increasing popularity can be attributed to its introduction in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), an annual software contest by the ImageNet project [57], through the AlexNet [58] which won the competition in 2012. While densely connected layers in a deep neural network learn global patterns in the feature space of the input, the convolutional layers in a CNN learn local patterns. In the case of image processing, this can translate to applying 2D filters to find local features in small windows of the inputs [32]. By stacking these convolutional layers, each layer will learn larger patterns with the features obtained from the previous one, giving the CNN architecture the ability to learn multiple features with a high level of abstraction. Pooling layers are often used after a set of convolutional layers for dimensionality reduction, which also makes the learned features less sensitive to their specific location in the input. Classification is achieved by feeding the extracted features into a final set of fully connected layers (FC).

This type of architecture was used in [59], where the proposed CNN architecture (Figure 3.8), the GEINet, is applied to cross-view gait recognition and evaluated with the OU-ISIR dataset [60]. It consists of two blocks of layers, each with one convolutional, one pooling and one normalization layer, followed by two fully connected layers and ending with a softmax layer to output probabilities for each class. In [61], a CNN was used to perform gait identification and tested on the CASIA B dataset [62], containing subjects walking from 11 different camera viewpoints. In this case, recognition was tested for each view separately, with an average accuracy rate of 95.88% over all views.

![Figure 3.8: Illustration of the GEINet architecture [59].](image)

Still in the context of gait recognition, the work in [63] developed a network containing two parallel CNNs sharing the same parameters. In the training stage, pairs of images containing the same person and pairs containing different identities are the input of the system, which will combine the output of each CNN to compute a contrastive loss. With this model, a similarity metric can be learned to simultaneously minimize the distance between image pairs of similar subjects and maximize it for differing pairs.

A 3D convolutional neural network uses 3D convolutions to capture spatiotemporal features, as the one in [64] used for gait recognition in multiple views (Figure 3.9). It uses 3x3x3 convolutional filters, with the three dimensions representing spatial width, spatial height and temporal extent. This allows for the inclusion of temporal information, with a somewhat limited range, as well as more non-linearities between convolutional layers. The performance of this method was evaluated using the CASIA-B gait sequences from 11 different view angles, achieving an average recognition accuracy of 97.35% over all views.
CNN and Transfer Learning

Deep learning is used to solve elaborate problems that require a proper training process to learn complex patterns present in the input space. For this to be possible, a large amount of labeled data is usually required to provide diverse and meaningful examples of input data, giving the model the capacity to generalize over new, unseen data. In the context of pathological gait classification, datasets containing sufficient video sequences of patients affected with gait related pathologies are not available and hard to obtain. Transfer learning is a machine learning method where a pre-trained model built for one task is reused to perform a different task, thus transferring the previously acquired knowledge to solve a new problem. Fine-tuning, a process to effectively repurpose a model, consists of adjusting the parameters of a pre-trained network by training a part of it, usually some of the top layers, with new data specific to the new context.

Competitors on the ImageNet recognition challenge [57] are provided with a subset of the ImageNet database [65] to train their models with over 1.3 million images to classify new images into 1000 different categories. Some of the available top performing models in the task of image classification are: Xception [66], ResNet [67], Inception [68], VGG-16 and VGG-19 [69]. These models pre-trained on ImageNet have been used in many problems related to image analysis, showing great performance improvements.

In the medical field, a pre-trained CNN architecture, the Inception V4 [70], was employed in the detection of Alzheimer’s disease on MRI image data [71], achieving an average of 96.25% detection accuracy with a relatively small dataset. In [72], VGG-19 was used for cross-view gait recognition with Joint Bayesian to model view variance (see illustration in Figure 3.10), supporting the use of CNNs to create effective appearance based gait analysis systems. Tested on OULP dataset [60], it achieved an average accuracy rate of 97.75% in recognition on the same view, and average of 89% with the largest view variance of 30°.

A pathological gait classification system based on GEI gait representations was proposed in [73], using a pre-trained VGG-19 model (Figure 3.11), fine-tuned on the INIT database [41]. Classification could be performed using the VGG-19 by adapting its fully connected layers to output a different number of class predictions. However, the best performance was achieved with only the first FC layer of the VGG-19 as a feature vector, using Principal Component Analysis (PCA) for dimensionality reduction and
Linear Discriminant Analysis (LDA) for classification. This involved fine-tuning the model by replacing the softmax layer, which originally outputs predictions corresponding to 1000 different classes for the ILSVRC challenge, with another softmax layer outputting predictions for the desired number of classes. The proposed system was evaluated on the DAI dataset [42], testing its performance in classifying normal and impaired gait, and on the DAI 2 dataset [26], for the classification of gait across 5 different gait related pathologies, achieving accuracies of 97% and 95% respectively.

The previously mentioned work [27] in section 3.2.1 also used a VGG-19 architecture in the context of pathological gait classification. Fine-tuning was done with the GAIT-IST database to perform classification across 4 different pathologies and normal gait. In the process of fine-tuning, layers can be frozen in order to maintain its parameters unchanged while the rest are trained on new data. This is often used to preserve the general basic features, such as edges, simple colours and curves learned by the first layers. Considering each set of convolutional layers separated by a pooling layer as a convolutional block, the described method to find the optimal fine-tuning starts by first training just the last block, freezing the others and testing the classification results on the validation set. This process is repeated, unfreezing the last frozen block at each time and ending at block 2.
The CNN (Figure 3.12) was tested for classification, by changing the last FC layer. It was replaced with one containing 5 channels, corresponding to 5 gait types in the new classification task, instead of the original layers with 1000 channels.

Two other classifier models were also tested, namely, Support Vector Machine (SVM) and LDA. As in [73], these involved using only the first fully connected layer of the VGG-19 as a feature vector and applying PCA for dimensionality reduction. The proposed systems are used on GEI and SEI representations of the same silhouette images in the GAIT-IST database, comparing the respective classification results. The system with the CNN classifier performed with an overall accuracy of 94.5% and 97.4% using the GEIs and SEIs respectively. For the LDA classifier, an overall accuracy of 89.6% and 96.4% was achieved for the GEIs and SEIs respectively. Although the best results are achieved using the CNN classifier on the GAIT-IST database, the system using the LDA had a better performance in a cross-database scenario where testing was done using the DAI2 dataset [26], with an overall accuracy of 76.7%, while the CNN classifier had a classification rate of 43.3%.

**Recurrent Neural Networks (RNN)**

Recurrent neural networks are characterized by including memory when processing a sequence of input data. This type of architecture processes input series and maintains a state containing information about previous data. The state parameters of an RNN are thus able to capture relationships between inputs in a sequence, meaning that such architectures are useful to process data in a time series and learn possible temporal related patterns. However, long-term dependencies with information being kept for many timesteps proved to be a practical impossibility due to the vanishing gradient problem [74]. During the process of backpropagation (see 3.2.3 for a detailed description) in networks with several layers, the gradient of the loss function with respect to weights at the initial layers is a result of the multiplication of several derivatives that depend on parameters from every layer since the last. For each value smaller than one present in that chained multiplication, the result will decrease exponentially and may practically vanish as it reaches the initial layers, preventing those weights to be effectively updated. This is especially challenging in the case of RNNs since the weights of the network are used recurrently to compute the state at each timestep as a function of the current input and previous state. This means that backpropagating the error at an output of the network to reach previous timesteps involves repeated multiplications of the same weights, resulting in the exponential scaling of small values. Long Short-Term Memory (LSTM) networks [75] were introduced as a solution to this problem, allowing the storage of short and long term dependencies in the state of LSTM cells. This type of architecture (see...
Figure 3.13) presented the novel concepts of gate units and memory cell, which allowed for the ability to regulate the impact of the information running through the network. Usually, a gate can serve one of three different purposes, namely, discarding current information from the cell state, determining new information to add to the cell state or yielding the output of the respective layer based on the current cell state. In the LSTM architecture, the gradient of the loss function backpropagates between timesteps of the entire recurrent chain through the cell state, without multiplications involving the network’s weight parameters. This prevents the gradients from vanishing, thus providing the ability to bridge long time gaps between the inputs.

![Image](image_url)

Figure 3.13: Illustration of (a) standard RNN [76] and (b) LSTM [77] architectures.

This notion of time dependency is applicable to gait analysis since each movement involved in locomotion is significantly related to past movements and will influence the ones after. This led researchers to consider the implementation of RNNs in gait analysis systems. Since representations such as the GEI provide a very compact representation of the temporal information in a gait sequence, it would be desirable to use the images as a direct input to an RNN. However, image inputs are not suited for the existing RNN models and sets of meaningful features extracted using a CNN are often used to reduce the dimension and complexity of the data.

In [78], a new method to learn gait features for recognition was developed to preserve temporal information in gait sequences. It used a CNN and a LSTM model, combined to perform cross-view gait recognition. Although the representation of subjects using binary silhouettes isolates the observed subject, simplifying the image and discarding unnecessary information, it can be associated with segmentation errors and is sensible to clothing and carried object occlusions. This is especially true for gait analysis in unconstrained environments. The solution in this method was to use the CNN to extract body joint heatmaps, as illustrated in Figure 3.14.

The heatmap representation of the body joints is directly fed into the LSTM and the value of the output from the last timestep is used as a gait sequence representation. Training and testing were done with the CASIA-B dataset [62] and the best results were achieved for recognition with a probe view of 54° and gallery views in a 0°-72° range, presenting an average accuracy of 83.8%.
Bidirectional RNNs [79] (see Figure 3.15) extend RNN models and are typically used when the full input sequence is available at prediction time. In these architectures, the input series is processed in its forward and backward order simultaneously to compute the output at each timestep based on past and future information. This increases the amount of information provided to the network which can improve the performance in certain classification problems.

This type of network was used for action recognition [81], preceded by a multi-stream CNN for feature extraction. The multi-stream network (MSN) has four convolutional networks based on the VGG architecture to compute features based on appearance and pixel trajectories (motion). The CNNs outputs are fed to a bidirectional LSTM to analyse long-term temporal dynamics of the actions.

In [82], a bidirectional LSTM (see Figure 3.16) is employed in the task of pathological gait classification. This work also produced a new open-access dataset where subjects performed their natural gait and simulated two types of impairments. The data was acquired with a Kinect camera which provides joint orientation values, used to calculate flexion angles of the lower limbs. These angles are then used
as inputs for the LSTM for classification, with an average accuracy of 82%, tested on the developed database.

Figure 3.16: LSTM architecture adopted in [82].

**Convolutional LSTM**

Recurrent networks process inputs as a time series of one dimensional features. These architectures can also be used for processing spatiotemporal information, such as image sequences, by having a CNN in a first stage of feature extraction as previously referred, and flattening its outputs into a 1D feature vector. The work in [83] proposes a different approach with a convolutional LSTM model that extends the fully connected LSTM, where a unit in each layer is connected to every unit in the previous instance of that layer. In this architecture, convolutional structures are present in input-to-state and state-to-state transitions. The recurrent layers in this model are therefore able to deal with inputs containing 2D spacial data and use it for convolutional operations as in CNNs. This allows to have combined in one model the capacity to process spatial and temporal information. Figure 3.17 illustrates a convolutional LSTM (ConvLSTM) cell.

Figure 3.17: Illustration of a ConvLSTM cell [84].
In [85], this type of architecture was applied to process human motion for gesture recognition. It proposes a model that uses a 3D CNN in a first stage of feature extraction to obtain short-term features of gestures and combined it with a Convolutional LSTM to further extract long-term spatiotemporal features. Considering this architecture for gait analysis can be advantageous since it provides the ability to process the temporal relations between the frames in a gait sequence using the 2D spatial features from each one.

**Capsule Neural Networks**

Convolutional networks can recognize or classify an object by detecting and combining sets of features learned by each of its layers. High level features are a combination of lower level features as a weighted sum, without considering the rotational and translational relationships between them. The example shown in Figure 3.18 illustrates the difficulty, even for top performing CNNs, to detect the difference between the illustrated face and one with the correct disposition of its constituting elements. This can be considered as a drawback in most CNN architectures, especially for applications such as detection with overlapping or distorted objects.

![Figure 3.18: Illustration of (a) face features in wrong proportion and orientation, (b) CNN detecting a face based on unmatched face features [86].](image)

The typical approach to tackle this issue in CNNs is to use pooling layers or successive convolutional layers to reduce the spacial size of the data and allow higher layer units to detect features in a larger region of the input image [87].

Capsule networks (see Figure 3.19) introduce the concept of layers with capsules, which are groups of neurons whose activities represent properties of an entity present in the image. Those properties include different types of instantiation parameters such as size, position, orientation and deformation [88]. This is achieved by replacing scalar-output feature detectors of CNNs with vector-output capsules. Each capsule outputs the probability of the entity being present in the input, encoded as the length of the output vector, as well as a set of instantiation parameters, encoded as that vector’s orientation. This provides a different way of recognizing wholes by their parts since two active capsules may activate a higher level capsule if the visual entities they represent have the expected spatial relationship [89]. The process of ensuring that the output of a capsule is sent to an appropriate parent capsule in the layer above is referred as "routing-by-agreement". The weights of the connections between capsules in consecutive layers will be updated according to the relevance in their coupling, and this dynamic routing
can be viewed as an attention mechanism that allows each capsule to attend to some active capsules at the level below and to ignore others [88]. This is presented as an alternative to the pooling operations often present in CNNs, that are argued to lose important spatial information by making features invariant to small positional changes at the expense of discarding precise relationships between the parts of an object.

Figure 3.19: CapsNet with 3 layers from [88].

Capsule networks have shown promising results in different applications. A framework based on this architecture was developed as a diagnosis tool for the identification of COVID-19 cases from X-ray images [90]. It achieved an accuracy of 95.7% while being able to handle a small dataset, which is a significant capability in the general application of deep learning solutions, and especially in this context where these solutions are developed while facing a sudden emergence of the COVID-19 virus. It also has the advantage of having a significantly lower number of trainable parameters than other methods with CNN models.

In [91], a framework based on capsule networks was used in gait recognition, using input images such as the GEI. The feature extraction using capsule layers can be more robust to affine transformations on input images and the developed models were able to exceed state-of-the-art results, evaluated in different tasks including cross-view, cross-walking and cross-clothing conditions.

3.2.3 Classification: Machine Learning Techniques

The final stage of a gait analysis system is the classification process. Depending on the type of features extracted, different methods should be considered to implement the classifier. As previously mentioned regarding the work from [25], choosing the best classifier can also depend on the datasets used to test the performance of the system. According to the conducted research on state-of-the-art systems, two of the most popular classifiers are presented.

Artificial Neural Networks

Artificial Neural Networks are based on layers of connected units, namely, input layers, hidden layers and output layers, as illustrated in Figure 3.20. A single connection is characterized by a weight parameter and is established from a unit in one layer to another unit belonging to the next layer. Each unit can be connected to multiple units in the previous layer and in next layer. Values from the previous layer are multiplied by the weights of the respective connection and summed at the input of each unit, which will put the result through a nonlinear function to yield an output for the next layer. This allows for the detection of non-linearities in the input data, which makes neural networks suited for learning complex
patterns in classification problems.

The output of a unit in the network can be defined as follows:

$$a_i^j = f \left( Wa_i^{i-1} + b \right)$$  \hspace{1cm} (3.2)

Where $a_i^j$ is the jth unit in the ith layer of the system, $W$ is the weight vector of corresponding to the connections from the previous layer to this unit, $a_i^{i-1}$ is the vector containing the values outputted by all units of the previous layer, $b$ is the bias associated with the current unit and $f$ is a non-linear function usually referred as the activation function. The values yielded at the output layer are the classification predictions, which can be in the form of a probability distribution over each class.

When training these networks, a loss function is associated to the network outputs in order to quantify how close the predictions are relative to the ground truth. The weights of the network are updated according to the loss function to minimize its value, through an algorithm called backpropagation which consists in finding the derivative of the loss function with respect to weights. The partial derivative of the loss with respect to a single weight is calculated using the chain rule:

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial S} \frac{\partial S}{\partial w}$$  \hspace{1cm} (3.3)

Where $L$ represents the loss function, $f$ represents the activation function of the unit, $S$ is the weighted sum of the outputs from units in the previous layer and $w$ is the weight to be updated.

Linear Discriminant Analysis

Linear Discriminant Analysis is a method to find a linear combination of features to perform dimensionality reduction on the input space while maintaining the discriminant relations between the variables, as illustrated in Figure 3.21. The first step is to compute the within class and between classes scatter matrices, which are an estimate of the covariance matrices. The within class scatter matrix is defined as:

$$S_w = \sum_{c=1}^{C} S_c,$$  \hspace{1cm} (3.4)
where \( c \) is the number of classes and \( S_c \) is given by

\[
S_c = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T,
\]

(3.5)

where \( x_i \) is a sample of the class \( c \), \( N \) is the total number of samples in that class and \( \mu \) is the mean value across all samples of that same class.

The between class scatter matrix is given by

\[
S_b = \sum_{i=1}^{N} (\mu_i - \mu)(\mu_i - \mu)^T,
\]

(3.6)

where \( N \) is the number of classes, \( \mu_i \) is the mean across all samples of a class and \( \mu \) is the mean across all samples of all classes.

The main objective of LDA is to find the projection matrix \( P \) that maximizes the ratio between the determinant of \( S_w \) and the determinant of \( S_b \):

\[
P_{LDA} = \arg\max_P \frac{P^T S_b P}{P^T S_w P},
\]

(3.7)

This will provide with a new axis system with a reduced dimensionality to which the inputs can be projected without losing class separability.

(a) ![Figure 3.21: Illustration of the LDA method][93].

---

[93]: #
4 GAIT-IT Dataset

This chapter presents the proposed dataset of video-based pathological gait data, GAIT-IT. Section 4.1 introduces the motivation for its development and Section 4.2 reviews the existing pathological gait datasets publicly available. The description of the proposed dataset and the simulated gait types that it features follows in Sections 4.3 and 4.4. Finally, in Section 4.5, the processing steps performed to obtain the different gait representations in GAIT-IT from the acquired video gait data are described.

4.1 Motivation

Pathological gait has gained increasing interest from researchers and different methods have been employed to create automatic analysis systems. To properly develop and evaluate such methods, it is essential to have significant amounts of relevant data available, especially in the application of deep learning solutions. However, given the constraints of acquiring gait data from real patients and data privacy protection issues involved, currently available datasets for the study of pathological gait are comprised of subjects simulating impaired gaits. These datasets are also characterised by typically containing information from a limited number of subjects, or relatively small amounts of data. For these reasons, the creation of a new pathological gait dataset was included in the scope of this work, in order to increase the available data and support the application of deep-learning techniques. This was achieved in a collaboration with João Pedro Machado, a master’s student of Telecommunications and Computer Engineering (METI) from ISCTE-IUL, and FCT|FCCN (Fundação para a Ciência e a Tecnologia), which provided access to a professional studio and video acquisition resources.

4.2 Existing Datasets

From the publicly available vision based pathological gait data, four datasets can be distinguished, all of which are composed of gait sequences recorded from a sagittal plane. The DAI [42] dataset 1 was acquired with the RGB camera of a Kinect 2 sensor in an 8 meters long hall. It contains 30 sequences of binary silhouettes performed by 5 different subjects, including 15 normal gait sequences and 15 abnormal gait simulations with random impairments. The silhouettes are not centered in the available images, and the segmentation algorithm employed to generate the silhouettes of the walking subjects was not always accurate, resulting in some silhouettes of poor quality.

The DAI 2 dataset [26] is composed of 75 gait sequences of binary silhouettes from 5 different subjects walking for 8 meters. Each one repeats 3 times the simulation of 4 pathological gaits, namely, diplegic, hemiplegic, neuropathic and Parkinsonian gaits, along with their normal gait. As for the previous dataset, these sequences do not include the pre-processing step of centering the silhouettes on the images and are significantly affected by segmentation errors.

The INIT [41] gait dataset 2 has a total of 80 sequences with high-quality binary silhouettes, recorded in a specialized studio with a green background to allow segmentation using chroma keying techniques.

1 http://hdl.handle.net/10045/70567
2 https://www.vision.ujl.es/gaitDB/
Inspired by abnormal gait patterns typical of several neurological diseases, 7 types of gait impairments were simulated by 10 subjects, which also provided with their normal gaits. These impairments include: having one leg take steps with about half of the length of the opposite leg, one arm swinging with roughly half of the amplitude of the other arm, one arm being totally immobile and a gait pattern where the whole body is affected by multiple symptoms.

As the largest of the reported datasets, the GAIT-IST [94] \(^3\) comprises a total of 360 gait sequences captured from 10 subjects. Each subject simulated the same 4 pathologies present in the DAI 2 [26] dataset along with their normal gait, allowing the possibility of cross-database test scenarios. For the pathological gait types, 2 levels of severity were recorded, with a subject performing 4 sequences for each severity, as well as for the normal gait. The acquisition was done using a cellphone camera with a 720p resolution, fixed on a tripod at a height of 1.5 meters, and the setup had a uniform background of a white wall and grey floor. This dataset provides binary silhouette and skeleton images, as well as Gait Energy Image (GEI) and Skeleton Energy Image (SEI) representations for each sequence, all of which are centered on the silhouette or skeleton, and normalized to a size of 224x224 pixels. The silhouettes and skeletons in this dataset all correspond to frames in a complete gait cycle, used to compute the respective GEIs and SEIs.

Table 4.1 summarizes the main characteristics of the mentioned gait datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Number of Gait Types</th>
<th>Number of Subjects</th>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAI [42]</td>
<td>2016</td>
<td>2</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>DAI 2 [26]</td>
<td>2017</td>
<td>5</td>
<td>5</td>
<td>75</td>
</tr>
<tr>
<td>INIT [41]</td>
<td>2018</td>
<td>8</td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>GAIT-IST [94]</td>
<td>2019</td>
<td>5</td>
<td>10</td>
<td>360</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of currently available pathological gait datasets.

### 4.3 Dataset description

The proposed GAIT-IT dataset was recorded in a professional studio, made available for this project for two full days. Details on the studio features can be found at the studio's web site \(^4\). The studio provided with a chroma key green background under artificial lighting that eased the task of image segmentation and allowed the extraction of high-quality binary frames. The acquisition setup consisted of two professional 4K cameras and a cellphone camera. The first camera was placed to capture the side view, at approximately 3 meters from the target, while a second camera was placed to capture the front/rear view, at about a half meter from the starting point of the gait sequence walking position. Both cameras stood on tripods at 1.75 meters from the ground and synchronized video recordings of the gait sequences, from the respective points of view, were made. The cellphone camera, with a resolution of 720x1280 pixels was placed on a tripod at a height of about 1.5 meters, to capture the gait sequences from the side view alongside the professional camera. It provides an alternative source with different

\(^3\)http://www.img.lx.it.pt/GAIT-IST/

\(^4\)https://www.fccn.pt/en/collaboration/studio/
quality that can be used to analyse the impact of video resolution and small changes in the point of view.

In order to further extend the consistency among the DAI 2 and GAIT-IST datasets, it was decided to create a dataset featuring each subject's normal gait along with those same 4 pathological gaits: diplegic, hemiplegic, neuropathic and Parkinsonian. This allows for more cross-database test scenarios while expanding the variety of existing gait data in the number of different subjects, quality of the samples and acquisition scenarios. For each pathology, 2 levels of severity were considered and the subjects were asked to provide with 4 gait sequences per severity level and for their normal gait. These 4 sequences correspond to a subject walking twice from left to right and from right to left, from the perspective of the side view cameras.

The acquisition took place in two different days with the participation of 21 subjects (19 males and 2 females) in the age range of 20-56 years old, with an age distribution as shown in Figure 4.1. Considering that 2 participants repeated the experience on the second day, there is a total amount of 828 gait sequences in the dataset. The repetition from those 2 subjects provides with the possibility to test and analyse the gait differences within the same person, by recording them in two separate days and wearing different clothes.

4.4 Description of Simulated Gait Types

The main attributes of the 4 pathological gaits simulated in the GAIT-IT dataset are briefly reviewed here to highlight a set of movements and impairments that characterize each one - please refer to Section 2.2 for the pathological gait type characterization. The participating subjects were instructed to simulate the gait types based on the following guidelines:

- **Diplegic** (Figure 4.2): Both sides of the body are affected, with flexing of the hips, knees and arms. The knees and ankles are also internally rotated. With a leaning forward posture, walking involves dragging both feet in a circular motion, with each step first going away from the body and then inwards. For the second (highest) severity level, the overall bending is accentuated, the arms go from being held close to the waist to being bent against the chest and the legs make a bigger circular motion slightly crossing in front of each other in what is usually referred as scissors gait.
• **Hemiplegic** (Figure 4.3): Only one side of the body is affected, which in this case it was decided to always be the right side. The right arm remains still and held close to the waist and in the second severity level is flexed against the chest. The affected leg is dragged in a circular motion, with a broader reach for the second severity. The left side remains normal, resulting in a limping walk.

• **Neuropathic** (Figure 4.4): The subjects were asked to simulate foot drop and lift the knees higher than normal, to avoid dragging the toes on the floor. The second severity is characterized by an exaggerated lifting of the leg and swinging it forward to prevent landing each step with the tip of the foot.

• **Parkinsonian** (Figure 4.5): With a stooped posture, the arms are held close to the chest and the lower limbs are flexed and rigid. The subjects were asked to attempt simulating general and erratic body shaking while taking small and relatively fast steps. The second severity level involved an overall exaggeration of these symptoms.

### 4.5 Video Data Processing

The proposed gait dataset contains sequences of binary silhouettes and skeletons, as well as GEI and SEI gait representations computed from those images. To obtain such sequences and representations, the video data from the acquisition goes through a pre-processing stage, which corresponds to the first
module in the general architecture of a gait analysis system illustrated in Figure 3.1. This section describes the main pre-processing steps involved in yielding the binary silhouette and skeleton sequences, GEIs and SEIs from the acquired videos.

4.5.1 Background Subtraction

In order to obtain the binary silhouettes from the video frames, a moving subject is isolated from the rest of the image through background subtraction. This was achieved by taking advantage of the green background featured in the FCCN studio for the application of chroma keying techniques. The distinctive green colour along a uniform background with constant lighting facilitated the detection of foreground and allowed for its accurate segmentation.

First, a video frame containing just the background was taken from both the side and front views. These frames were then converted from RGB to the HSV colour space, which is often used in computer vision for colour based segmentation tasks, as the hue component (H) describes the colour appearance. By making a histogram of the image pixels in that representation, the background is determined by the combination of the resulting hue, saturation (S) and lightness (V) intervals, as shown in Figures 4.6 and 4.7. Given the correlations between the color components in the RGB color space, differences in the illumination can change the values of all three components (R,G,B). With the HSV color space, brightness and chromaticity can be easily separated [95] and background subtraction methods based on this representation can more efficiently account for the presence of shadows and illumination changes.

For each frame in a gait sequence, pixel values in HSV outside this range are detected as foreground and a binary mask is created from this segmentation. These masks isolate the observed subject, resulting in a sequence of binary silhouettes. To account for possible noise being detected as foreground and connected to the silhouette, morphological CLOSE and OPEN operations were applied to the binary
masks. In both operations, the structuring element was square shaped with a size of 5x5 pixels. This shape is often used for small elements relative to high resolution images and was chosen to consider all the neighbours of each pixel in a radius of two pixels in the morphological operations.

It was also necessary to address small clusters of noise caused by shadows in some videos. This typically occurs at the right edge of the image in front view frames and in the top left corner of side view frames. To correct it, small objects were removed from the foreground by defining a size threshold.

The described steps for background subtraction were implemented using the OpenCV library [96] for Python and example results from this process are illustrated in Figures 4.8 and 4.9.

4.5.2 Skeleton Computation

The method used to compute the skeletons was to perform human pose estimation, the task of localizing anatomical parts, on the gait videos. This was achieved with OpenPose [55], a software publicly available
online \(^5\) for free use. It is a system able to detect body, hand, facial and foot keypoints from multiple people simultaneously in real-time, on single images. The system architecture consists of a multi-stage CNN, as illustrated in 4.10. The network predicts a set of 2D confidence maps, corresponding to the location of different body parts, as well as a set of Part Affinity Fields (PAF) able to encode the level of association between parts. The PAF are a feature representation introduced by this work that preserves location and orientation information about the limbs. The confidence maps and PAF are ultimately parsed to output 2D coordinates of the anatomical keypoints.

Using the OpenPose system with the gait videos as inputs, the 2D coordinates of 25 key body parts can be extracted for each frame following a pose output format as illustrated in Figure 4.11, and labeled according to Table 4.2.

With the OpenCV [96] Python library, the skeletons are then obtained by drawing a line between pairs of extracted keypoint coordinates according to these labels. In Figure 4.12, an example is shown where

\(^5\)https://github.com/CMU-Perceptual-Computing-Lab/openpose
Figure 4.11: Pose output format of detected body parts using OpenPose [97].

<table>
<thead>
<tr>
<th>Number</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Nose</td>
</tr>
<tr>
<td>1</td>
<td>Neck</td>
</tr>
<tr>
<td>2</td>
<td>Right Shoulder</td>
</tr>
<tr>
<td>3</td>
<td>Right Elbow</td>
</tr>
<tr>
<td>4</td>
<td>Right Wrist</td>
</tr>
<tr>
<td>5</td>
<td>Left Shoulder</td>
</tr>
<tr>
<td>6</td>
<td>Left Elbow</td>
</tr>
<tr>
<td>7</td>
<td>Left Wrist</td>
</tr>
<tr>
<td>8</td>
<td>Hip (Middle)</td>
</tr>
<tr>
<td>9</td>
<td>Hip (Right)</td>
</tr>
<tr>
<td>10</td>
<td>Right Knee</td>
</tr>
<tr>
<td>11</td>
<td>Right Ankle</td>
</tr>
<tr>
<td>12</td>
<td>Left Hip</td>
</tr>
<tr>
<td>13</td>
<td>Left Knee</td>
</tr>
<tr>
<td>14</td>
<td>Left Ankle</td>
</tr>
<tr>
<td>15</td>
<td>Right Eye</td>
</tr>
<tr>
<td>16</td>
<td>Left Eye</td>
</tr>
<tr>
<td>17</td>
<td>Right Ear</td>
</tr>
<tr>
<td>18</td>
<td>Left Ear</td>
</tr>
<tr>
<td>19</td>
<td>Left Big Toe</td>
</tr>
<tr>
<td>20</td>
<td>Left Small Toe</td>
</tr>
<tr>
<td>21</td>
<td>Left Heel</td>
</tr>
<tr>
<td>22</td>
<td>Right Big Toe</td>
</tr>
<tr>
<td>23</td>
<td>Right Small Toe</td>
</tr>
<tr>
<td>24</td>
<td>Right Heel</td>
</tr>
</tbody>
</table>

Table 4.2: Openpose keypoint output labels

The extracted pose is rendered as a skeleton on the video frame and a binary skeleton is drawn using the outputted keypoint coordinates. One problem was experienced when applying this method to obtain skeletons from video sequences in the front view. When a subject is walking away from the front view camera, the facial keypoints corresponding to the nose, the eyes and often the ears are not visible. Without the coordinates for these points, the head of the subjects can not be drawn in the skeleton, which means that for such sequences the respective skeletons are computed from the neck down.

4.5.3 GEIs and SEIs Computation

The gait videos from the acquisition and the obtained silhouette and skeleton images had a resolution of 1280x720 pixels, which contributes to the quality of the gait features in both representations. However, in current deep learning solutions for feature extraction and classification, high resolution images are not suited to be used as inputs due to their large dimensions. In fact, state-of-the-art CNN models for object classification, such as the VGG19 [69] and ResNet [67] architectures, are designed to handle input images with size 224x224. For this reason, every image in the dataset was resized to 224x224 pixels. Downsampling reduces the amount of information in an image but the final quality and accuracy of the features still benefit from the fact that silhouette and skeleton computation were computed using the high resolution original images.

To resize the obtained silhouettes and skeletons the first step was to crop each image to keep only the relevant information. This significantly reduces the image dimensions without losing any gait information, since only the background around the subject is removed. This was achieved by defining a bounding
box around the detected foreground, as illustrated in Figure 4.13.

The next step was to determine the centroid of the foreground. The width of the cropped image is then padded with zeros to match its height and align it with the obtained centroid. The result is a square image with a centered foreground that can now be normalized to a size of 224x224 pixels, while maintaining the original aspect ratio.
Figure 4.14: Illustration of the method to identify complete cycles from a gait sequence. (a) Representation of the step width measured in a binary frame. (b) Plot of the step width along the frames in a gait sequence, highlighting the maximum distance peaks. Two gait cycles can be determined for the analysed sequence: one cycle between the first and third peaks and a second cycle between the third and fifth peak.

The computation of GEIs and SEIs follows the description provided in Section 3.2.1. The frames corresponding to a subject entering or leaving the camera’s field of view were discarded, as well as the silhouettes and skeletons that did not correspond to a frame included in a complete gait cycle. This means that each gait sequence contains an integer number of cycles. For each gait cycle of each sequence a GEI and a SEI were computed. Another GEI and SEI were also computed for each sequence, considering all the gait cycles it contains. In order to isolate each gait cycle, the step width is measured along a sequence in the side view. This corresponds to tracking the distance from the heel of the back foot to the tip of the front foot as the width of a bounding box around the lower part of the detected silhouette. This can be plotted against the frame numbering as illustrated in Figure 4.14. Considering that a maximum peak in the feet distance plot represents the approximate moment where the feet are furthest apart, one step can be said to occur between two consecutive distance peaks. This means that a complete gait cycle can be determined as the frames between three consecutive peaks. Due to the fact that the videos taken from both views are synchronized for each sequence, the obtained frame numbers that identify a complete gait cycle in the side view are the same for the front view.

Only a lower portion of the binary image is considered in order to isolate the feet when measuring the step width, so that the heel of the back foot and the tip of the front foot can be determined as the leftmost and rightmost pixels of the foreground, respectively. A difficulty in this process was the fact that the size of the considered portion had to be adjusted depending on the pathology simulated in each sequence. The Parkinsonian and diplegic gaits are characterized by walking with bent knees and a stooped posture. For these cases, a smaller portion was defined to avoid capturing the bent knees as the right edge of the foreground. In the case of the simulated neuropathic gait, which involves walking with high steps, a larger portion was necessary to make sure that the feet do not rise above it. An illustration of different image fractions considered to track the step width along a sequence is shown in Figure 4.15.

The GEIs can then be computed from the binary silhouettes according to equation 3.1 from Section 3.2.1 as proposed in [39]. Using the same method, the SEIs can be computed from binary skeletons and an example of both representations obtained from a complete gait cycle is illustrated in Figure 4.16.
Figure 4.15: Examples of different binary silhouette portions considered to measure the step width, represented as a green rectangular box. The blue vertical lines represent the left and right foreground edges. (a) Binary silhouette from a diplegic gait sequence (b) binary silhouette from a neuropathic gait sequence.

Figure 4.16: Example of (a) binary silhouettes in a gait cycle (b) GEI computed from this gait cycle (c) binary skeletons in a gait cycle (d) SEI computed from this gait cycle.
5 Influence of the Dataset on Deep Learning Classifiers

The state-of-the-art for automatic gait pathology classification relies on deep learning vision-based approaches [27, 73, 82], and the development and testing of new solutions requires the availability of suitable pathological gait datasets. This chapter provides pathological gait classification results computed using the publicly available datasets and the proposed GAIT-IT dataset. The purpose is to demonstrate that a deep learning approach trained using GAIT-IT can improve its generalization ability to make predictions over new data, confirming the usefulness of the proposed dataset for gait pathology classification studies.

Section 5.1 starts by describing the baseline solution for pathological gait classification considered to assess the impact of using GAIT-IT to train deep learning classifiers. In Section 5.2, a set of cross-validation tests, using GAIT-IT and GAIT-IST, are conducted to evaluate the performance of the model re-purposed and tested using the same pathological gait dataset. In Section 5.3, cross-dataset tests are conducted with the objective of showing that a larger and better quality dataset can help obtaining improved classification results, especially when the trained algorithms have to operate in conditions different from those used for training.

5.1 Baseline Solution

The baseline solution considered for the following experiments is the VGG-19 [69] Convolutional Neural Network (CNN) model, pre-trained on a subset of the ImageNet dataset [65], and fine-tuned using pathological gait datasets. This choice aims at a direct comparison with state-of-the-art results since it is the model considered in the approaches from [27, 73]. The VGG-19 takes as input images of size $224 \times 224$ with 3 input channels, which go through 5 convolutional blocks for feature extraction, each consisting of consecutive convolutional layers followed by a max-pooling layer, as illustrated in Figure 5.1. With a total of 16 convolutional layers, each is characterized by $3 \times 3$ convolutional filters, while the 5 max-pooling layers have pooling filters of size $2 \times 2$. Classification is then made with a set of 3 fully connected (FC) layers at the end of the architecture. The first two FC layers have 4096 units, corresponding to the size of the flattened feature vector output by the fifth block. The third and last layer has a softmax activation function and was modified to have 5 units corresponding to the 5 gait classes considered here for classification: diplegic, hemiplegic, neuropathic, Parkinsonian and normal gaits. This classification network is not used with its parameters pre-trained on ImageNet, since the objective is re-purpose the VG19 convolutional base for a different classification task.

The gait representations considered as input to the VGG-19 network can be either a GEI or an SEI, with training and testing always assuming the same gait representation as input. As in [27], when using GEIs as inputs, fine-tuning involved re-training the last 3 convolutional blocks of the VGG-19 convolutional base. For the SEI representation, the best results were achieved by retraining all blocks except the first one. Fine-tuning is done using backpropagation, considering categorical cross entropy as the loss function, Stochastic Gradient Descent (SGD) with the Nesterov momentum [98] variation as the optimizer and a learning rate of 0.0002.
5.2 Cross-validation Experiments

The first set of results corresponds to training and testing using the same pathological gait dataset. Classification performance results are reported for each gait representation (GEI and SEI), using 10-fold cross-validation, considering 3 datasets: i) GAIT-IST; ii) GAIT-IT; and iii) combining the GAIT-IST and GAIT-IT datasets.

In each of the above cross-validation tests, the classification accuracy on the training and test sets was recorded over 50 training epochs for each fold. A general performance of the model is then obtained for the given dataset by averaging the accuracies across all folds, which can then be plotted to analyse the progression of the mean classification results along the training epochs. The classification accuracy on the test set at the optimal training epoch was also computed for each fold and the overall accuracy was obtained by averaging over all folds. The overall classification accuracy results are summarized in the first 3 diagonal cells of Table 5.1, with GEI results on the top part of the table and SEI results on the lower part.

i. GAIT-IST

When using the GAIT-IST dataset, 9 subjects were used for training and the remaining subject for testing, for each fold, as done in [27]. In Figure 5.2, the average accuracy and loss plots are shown for both cross-validation scenarios corresponding to each gait representation. In both cases, the performance on the training set quickly converges to near 100% accuracy. It can be observed that the accuracy of the models trained with the SEI representation tends to stabilize in early training and achieve a higher classification accuracy on the test set, while the average performance of the models trained with the GEIs peaks at the 38th epoch after which it stops improving. The mean overall accuracies obtained for each of the 10-folds, computed at the best training epochs, are 94.2% and 98.4%, using the GEIs and SEIs respectively.
ii. GAIT-IT
When using the GAIT-IT dataset, the test set for each fold is defined as $V_k = \{S_i, S_{i+1}, S_{i+2}\}$, where $i = 2 \times k - 1$, $k$ is the fold iteration and $S_i$ represents one of the 21 total subjects, following the numbered labels used for each subject in the dataset. This arrangement had the purpose of using all subjects in the test set at least once and providing with a significant number of folds for the cross-validation. When comparing both representations, it can be observed (see Figure 5.3) that the average performance of the models on the validation set is similar and stabilizes around the 30th training epoch. As before, the accuracy on the training set reaches near 100% in early training. The overall accuracies across all folds are 94.0% and 93.6%, using the GEIs and SEIs respectively.

iii. GAIT-IT and GAIT-IST
When using the combined GAIT-IST and GAIT-IT datasets, in each fold a different subject from GAIT-IST was used in the test set together with 3 subjects from GAIT-IT, making a total of 27 subjects for training and 4 subjects for testing. The test sets are defined as $V_k = \{V_{IST_k}, V_{IT_k}\}$ where $V_{IST_k} = \{S_{IST_k}\}$,
\[ V_{IT_k} = \{S_{IT}, S_{IT+1}, S_{IT+2}\}, \text{ } i = 2 \times k - 1 \text{ and } k \text{ is the fold iteration.} \]

Analysing the accuracy plots in Figure 5.4 shows that the average model performance over all folds stabilizes earlier and reaches slightly better classification results when using the GEIs compared to the SEIs. The overall accuracies are 94.4% and 93.2% for the GEI and SEI representations, respectively.

Figure 5.4: GAIT-IT and GAIT-IST 10-fold average training and validation accuracy history using (a) the GEIs as inputs and (b) the SEIs as inputs.

<table>
<thead>
<tr>
<th>GEI</th>
<th>Trained on</th>
<th>GAIT-IST</th>
<th>GAIT-IT</th>
<th>GAIT-IST+IT</th>
<th>Tested on</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIT-IST</td>
<td>94.2%</td>
<td>72.4%</td>
<td>-</td>
<td>-</td>
<td>55.6%</td>
</tr>
<tr>
<td>GAIT-IT</td>
<td>86.4%</td>
<td>94.0%</td>
<td>-</td>
<td>-</td>
<td>78.0%</td>
</tr>
<tr>
<td>GAIT-IST+IT</td>
<td>-</td>
<td>-</td>
<td>94.4%</td>
<td>81.4%</td>
<td></td>
</tr>
<tr>
<td>GAIT-IT(10)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>65.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEI</th>
<th>GAIT-IST</th>
<th>GAIT-IT</th>
<th>GAIT-IST+IT</th>
<th>Tested on</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIT-IST</td>
<td>98.4%</td>
<td>68.8%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAIT-IT</td>
<td>85.1%</td>
<td>93.6%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAIT-IST+IT</td>
<td>-</td>
<td>-</td>
<td>93.2%</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of all classification accuracies across all datasets using GEIs and SEIs as inputs and re-trained on VGG-19.

5.3 Cross-dataset Tests

Cross-dataset tests were conducted to evaluate the impact that training on a larger dataset can have in the trained model’s generalization capability.

Three scenarios were considered: i) training on the GAIT-IST dataset and testing on the GAIT-IT or the DAI2 datasets; ii) training on GAIT-IT and testing on GAIT-IST or DAI2; iii) training on the combined GAIT-IST and GAIT-IT datasets and testing on DAI2.

Since the silhouettes provided in the DAI2 dataset include some significant segmentation errors, this dataset was not considered for training in the cross-dataset tests, as it might lead the model to learn inaccurate features. In each scenario, test results were recorded at the optimal training epoch and a summary of the correct classification rates is presented in the off-diagonal cells of Table 5.1. The analysis of these results is provided below, but a clear degradation compared to the results of Section 5.2
can be observed. In this case, the best results are obtained with the SEI gait representation, when training on the proposed GAIT-IT dataset and testing on GAIT-IST, for which a cross-dataset average classification accuracy of 85.1% was still achieved.

i. Training with GAIT-IST

When using the GEIs from the complete GAIT-IST dataset to train the VGG-19 network, the model achieved its best results on the GAIT-IT dataset, with 32 training epochs, and an overall accuracy of 72.4%. As expected, the classification accuracy was lower when testing with the DAI2 dataset (55.6%).

When using the SEI representation, the best performance was again achieved when testing on the GAIT-IT dataset, with a classification accuracy of 68.8%, for 23 training epochs.

ii. Training with GAIT-IT

When using the GEIs from the GAIT-IT dataset to train the VGG-19 network, the best overall classification accuracy (86.4% with 24 training epochs) was obtained when testing on the GAIT-IST dataset.

When using the SEI representation, the best results when testing with the GAIT-IST consist of a 85.1% classification accuracy, with 17 training epochs.

The impact of using a larger dataset for training is obvious when testing on the DAI2 dataset using the GEI gait representation (the SEI could not be computed from the DAI2 contents). The overall accuracy increased to 78.0%, with 35 training epochs. This represents an increment exceeding 22% in the model's performance on this cross-dataset classification task, when compared to the results of the model trained with GAIT-IST. This confirms that the larger dataset used for training limited the model overfitting, a clear advantage of the proposed GAIT-IT dataset.

To further evaluate the impact of the high-quality silhouettes provided with the proposed GAIT-IT dataset, the VGG-19 network was also trained using just the first 10 subjects of GAIT-IT (the same number available in the GAIT-IST dataset), and used this truncated version of the dataset, named GAIT-IT(10), for training the model in the same conditions considered when training with GAIT-IST. An improvement of 9.8% in the cross-dataset results was observed when testing with the DAI2 dataset, in comparison to the accuracy value obtained when training with the GAIT-IST dataset (65.4% vs. 55.6%). This improvement seems to confirm the advantage of the proposed GAIT-IT dataset, even when training with the same number of samples. The higher-quality silhouettes allow the extraction of more accurate features and minimize the amount of misleading information resulting from segmentation errors in a preprocessing stage that can affect the classification performance of deep learning approaches.

iii. Training with GAIT-IST and GAIT-IT

Finally, both datasets were combined to train the VGG-19 network, corresponding to a total of 33 subjects. This includes the 10 subjects from GAIT-IST and the 21 subjects (plus the 2 subject repetitions as detailed in Section 4.3) from GAIT-IT. Besides further increasing the amount of available training data, this experiment provides the model with some variety in terms of the conditions in which the gait sequences were acquired.
Since the DAI2 does not contain binary skeleton gait sequences, the SEI representation can not be obtained, which means that only the GEI representation was used in this scenario. The best results were achieved by training the model for 19 epochs with an overall accuracy of 81.4%, an increment of more than 25% and 3% in the classification performance when compared to training the model with just the GAIT-IST and GAIT-IT, respectively.
6 Proposed Pathological Gait Classification System

Most deep learning approaches in gait analysis focus on the extraction of spatial features from an observed subject’s gait cycle representation, using convolutional neural networks. As performed in the experiments described in the previous section (see Section 5), a CNN can be trained (a pre-trained CNN is fine-tuned, in this case) to obtain 2D spatial information relevant for gait classification from single image inputs. Based on gait representations such as the widely used GEI, which provides with a robust and compact representation of a gait cycle, comprising temporal information in a single image, such solutions can achieve state-of-the-art results in pathological gait classification.

However, the condensed representation of a complete gait cycle into one single representative image cannot seize all the temporal features, present in the relationships between the movements captured at each time step. As shown in clinical research in the field of gait analysis, gait patterns can be observed and determined through a sequence of events characterized by their functional purpose [19] (detailed in Section 2.1.1).

The approach proposed in this section combines the spatial and temporal feature extraction abilities of convolutional and recurrent neural networks, respectively, to process gait cycles as a collection of binary frames from individual time steps. The purpose is thus to capture and learn temporal patterns, associated with the inherent order and dependencies among the spatial features obtained at different moments in a gait cycle. The recurrent neural network architecture considered to process temporal information is the LSTM network [75]. As described in Section 3.2.2 regarding RNNs, the LSTM architecture is designed to solve the inherent difficulty of training RNNs, allowing the storage of short and long term dependencies and the regulation of the information that flows through the network.

In Section 6.1, the general architecture of the proposed system is presented, followed by the descriptions of the 4 main modules it comprises. Section 6.2 compares the implementation of different pre-trained CNNs on the proposed system. In the following Section 6.3, two of those pre-trained CNNs are fine-tuned to improve the performance of the system. Section 6.4 proposes a bidirectional RNN to further improve the overall performance. Finally, in Section 6.5, a CNN architecture is proposed, built as a convolutional base specifically for the task of pathological gait classification.

6.1 Proposed Gait Analysis Framework

The proposed gait analysis framework performs pathological gait classification from a given sequence of binary frames that constitute a gait cycle. For the work presented in the following approach, the binary silhouettes from the GAIT-IT pathological gait database (see Section 4.3) are considered as the input data.

The proposed framework consists of 4 main modules, as illustrated with the general system architecture in Figure 6.1:

- In the first module, a fixed amount of 9 key frames is collected from a complete cycle in a gait sequence.
- In a second stage, each of the 9 key frames is processed by a pre-trained CNN to obtain a set of

51
flattened 2D spatial features.

- For the third step, the sequence of 9 sets of spatial features extracted from a complete gait cycle is processed by an LSTM network to obtain spatio-temporal features between frames.
- The last step is to classify the extracted features from the previous step using a dense neural network of fully connected layers.

A description of each of the four main modules is included in the following subsections.

### 6.1.1 Key Frame Selection

The purpose of this module is to select a fixed amount of key frames in each gait cycle that represent the progression of the subject's motion. As mentioned in Section 2.1.1, according to clinical research [19], a complete cycle from an unimpaired gait can be characterized by eight distinct functional phases (see Figure 6.2). By analysing a subject's gait at the relative moments when each phase occurs in normal gait, the differences in body posture of the observed motion can be used to recognize a form of impairment or gait pathology. This served as motivation for the process of frame selection, where 9 key silhouette frames are selected from the total amount of frames in each gait cycle, according to those eight phases. The first and last of the 9 frames mark the beginning and end of the cycle, respectively, while the rest represent the transition between consecutive phases.

In order to obtain these key frames, the first step was to determine and isolate each cycle. This was achieved as in the method to compute the GEIs and SEIs for the GAIT-IT database, detailed in Section 4.5.3. After plotting the feet distance along a sequence, each gait cycle can be determined as the frames between three consecutive maximum distance peaks and two key frames can be directly determined as the first and third of those three peaks, representing the beginning and end of the cycle, respectively. The next step is thus to obtain the other 7 frames between those two. As described in Section 2.1.1, the transition between the stance and swing phases occurs at an estimated 60% of a complete cycle in normal gait. This corresponds to the transition between the fifth and sixth previously mentioned functional phases, and a third key frame is thus determined. With this frame as reference, the
first 60% of the cycle is divided in 5 equal portions that correspond to the five stance phases, yielding four more key frames at the phase transitions. The same is done for the last 40% of the cycle, which is divided in three equal portions representing the three swing phases, yielding the final two key frames. The key frames in relation to the eight functional phases of normal gait are illustrated in figure 6.2. An example of 9 key frames representing a complete cycle from a simulated Parkinsonian gait are shown in the following figure 6.3, portraying the differences in the observed motion at the same relative moments defined for normal gait with respect to the whole cycle.

In figure 6.4, two plots of the feet distance along the frames in a gait sequence are shown highlighting the extracted 9 key frames. The first and second plots correspond to the normal and Parkinsonian gait sequences, respectively, from which the 9 key frames shown in figures 6.2 and 6.3 were obtained.

6.1.2 CNN Feature Extraction

In the first stage of feature extraction, the objective is to obtain a set of compact representations of certain visual aspects of the input image, meaningful to the classification task at hand. To achieve this, the convolutional base of a pre-trained CNN is used to process each of the 9 key silhouette frames and yield a set of 2D spatial features, flattened into a 1D vector, to be processed by the LSTM network at the second stage.

For this task, 4 different CNN architectures among the top performing models on the ImageNet image recognition challenge [57] are considered. These networks, namely, the ResNet50 [67], the Xception [66] and both VGG architectures proposed in [69], the VGG16 and VGG19 networks, all present a different approach to the task of image classification on large datasets using deep convolutional networks.
Figure 6.4: Feet distance plot along the frames highlighting the extracted 9 key frames in a (a) normal gait sequence and a (b) Parkinsonian gait sequence.

Using around 1.3 million training images, 50,000 validation images and 100,000 testing images from a subset of the ImageNet database [65] containing 1000 different classes, these networks achieved the following the top-1 and top-5 classification accuracies on the validation set [99]:

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception</td>
<td>0.790</td>
<td>0.945</td>
<td>22,910,480</td>
</tr>
<tr>
<td>ResNet50</td>
<td>0.749</td>
<td>0.921</td>
<td>25,636,712</td>
</tr>
<tr>
<td>VGG16</td>
<td>0.713</td>
<td>0.901</td>
<td>138,357,544</td>
</tr>
<tr>
<td>VGG19</td>
<td>0.713</td>
<td>0.900</td>
<td>143,667,240</td>
</tr>
</tbody>
</table>

Table 6.1: Classification accuracies on the ImageNet challenge [57] and number of parameters of the 4 considered CNNs.

ResNet50

The work in [67] proposes a solution to ease the training of increasingly deeper neural networks. It describes, through experimentation, how the accuracy during training gets saturated and then rapidly decreases as a network’s depth increases, not due to overfitting but rather to the increasing difficulty for current solvers to find solutions in feasible time and with finite data.

This work addresses the accuracy degradation problem by introducing a deep residual learning framework [67]. Rather than expecting that a set of stacked layers would fit a desired mapping function, denoted as $H(x)$, it proposes the hypothesis that the stacked nonlinear layer should more easily optimize fitting a residual mapping given by $F(x) := H(x) - x$.

In ResNet architectures, residual learning is implemented in every few stacked layers and the result is a framework where a network’s depth can be further increased to achieve a better performance. These remain easy to optimize when compared to networks of the same size that do not apply residual learning. In [67], residual nets with a depth of up to 152 layers are reported as having a lower complexity when compared to other top performing CNNs such as the VGG networks that have no more than 19 layers. An ensemble of these residual networks were able to achieve 3.57% of error rate when evaluated with the ImageNet dataset [65], a result that won the first place at the ImageNet challenge in 2015 [57]. The previously mentioned ResNet50 is a 50 layers deep implementation of a residual network from [67].
Xception

The feature extraction base of the Xception architecture consists of 36 convolutional layers structured into 14 modules, all of which have residual connections around them except for the first and last modules [66]. It is entirely based on depthwise separable convolution layers, which consist in a depthwise convolution followed by a pointwise convolution. A depthwise convolution can be defined as a spatial convolution performed independently over the input channels, while a pointwise convolution stands for a convolution that uses a $1 \times 1$ kernel, mapping the input data into separate spaces, smaller than the input dimensions.

In [66], the proposed Xception is extensively compared to the Inception architecture, in which it was inspired. Not only did Xception present state-of-the-art results in the single-label classification task on the subset of ImageNet [65] with 1000 classes, but it also achieved larger gains regarding the Inception V3 model [68] in the multi-label classification task on the JFT dataset [100], that has over 350 million high-resolution images with labels from a set of 17,000 classes [66]. The performance gains when compared to the Inception V3 network are argued to be due to a more efficient use of model parameters rather than an increased capacity [66].

Figure 6.6: Illustration of the Xception network architecture [66].
VGG16 and VGG19

The work in [69] presented a major contribution to the field of deep learning by presenting a thorough evaluation of increasingly deeper networks, based on the use of very small convolution filters at each layer of the architecture. The conclusions drawn from this research allowed the respective team to secure the first and second places at the ImageNet challenge in 2014 [57].

The VGG16 and VGG19 networks have a total of 16 and 19 weight layers (comprehending convolutional and fully connected layers), respectively. Every convolutional layer in these networks uses a filter with the smallest receptive field of $3 \times 3$ enough to capture the notion of left/right, up/down and center [69]. This is described to represent a significant advantage over former state-of-the-art architectures (such as former ImageNet challenge top performing models) that used kernels with dimensions such as $7 \times 7$ or even $11 \times 11$ in the first layers. Since, for example, a stack of two $3 \times 3$ convolutional layers has the same receptive field of a single one using a $5 \times 5$, both yielding a feature matrix of the same size, multiple layers with smaller filters are reported to provide a greater advantage. Firstly, more non-linear operations can be incorporated with more layers, and secondly, a significant reduction in the number of parameters can be achieved [69].

Not only has the VGG architecture been among the top performing models as it has been shown to generalize well to other datasets and used as a feature extraction base in deep learning approaches to different tasks [27, 72, 73, 101, 102].

6.1.3 LSTM Feature Extraction

Since gait is by definition a cyclic pattern of movements that produce locomotion, the use of a recurrent neural network has the purpose of capturing those movement patterns, as the relationship between gait observations (in the form of binary frames) at specific moments in a gait cycle. In this second stage of feature extraction, the flattened 2D spatial features obtained from the CNN for each of the 9 key frames are grouped as a sequence to be processed by an LSTM network. This network consists of one layer that has 9 LSTM cells, each taking one of the feature vectors from that sequence as input. These cells are all connected by a cell state that facilitates the flow of information through the entire chain. Each cell also has 4 different neural networks, referred as gate units, to regulate the cell state by deciding what information should be discarded and what information should be added to the cell state at the current time step, as well as deciding how the cell state is used to compute the output of the current cell [75] (see Section 3.2.2 for a more detailed description of this architecture). The LSTM network yields a 1D vector of spatiotemporal features that corresponds to the output of the last LSTM cell, as illustrated in Figure 6.1. The dimension of the output vector is set to a size of 256 and it is directly used as the input for the next stage of classification.

6.1.4 Classification

As the final step, classification is performed by a neural network consisting of 2 fully connected layers and a dropout layer between them. Dropout is a regularization method to prevent neural networks from
overfitting on the training data. During the training process, the inputs or outputs of a given layer can be randomly removed from the network, temporarily, along with its incoming and outgoing connections [103]. The result is illustrated in Figure 6.7. When fitting a network to training data, some units may correct the mistakes of other units before allowing the possibility of those units being fixed with further training. This results in complex co-adaptations which in turn lead to overfitting, since the model is adapting to the training set and losing the ability to generalize [103]. In effect, when applying dropout to a given layer, each update to that layer during training is performed with a different perspective of the configured layer [104].

The first fully connected layer has a dimension of 256 units, the same length of the input it expects from the feature extraction module. The following dropout layer must maintain the 256 size since its purpose is to act on the connections from the first to the second layer of the classifier. Finally, the second fully connected layer has 5 units with a softmax activation function to output the probabilities for each class representing a gait type. The LSTM and this classification network are trained using categorical cross entropy as the loss function and the Adam algorithm [105], with the Nesterov momentum [98] variation and a learning rate of 0.0001, as the optimizer. While the loss function evaluates how the network predictions stand against the actual values of the training data, the optimizer is the method used to minimize it at each iteration during the training process. Finally, the learning rate determines how much the model should update its weights in response to the prediction errors estimated by the loss function.

![Figure 6.7: Illustration of the temporary effect of dropout in a neural network [103].](image)

6.2 Performance Comparison Using Different Convolutional Bases

In this section, the proposed CNN-LSTM system is evaluated with the objective of comparing the performance of the system using each of the four pre-trained convolutional bases from the networks described in 6.1.2, namely, the ResNet50, the Xception, VGG16 and VGG19. These experiments are conducted using the binary silhouettes in the proposed GAIT-IT dataset for training and testing. The performance results for pathological gait classification are obtained using 10-fold cross-validation. As in the experiments from Section 5.2 using just GAIT-IT, the test set for each fold is defined as \(V_k = \{S_i, S_{i+1}, S_{i+2}\}\), where \(i = 2 \times k - 1\), \(k\) is the fold iteration and \(S_i\) represents one of the 21 total subjects, following the numbered labels used for each subject in the dataset. The 18 subjects not in the test set are used for training the LSTM and fully connected layers together.

The experiment results are presented in the following Table 6.2 as the average of the classification
accuracies on the test set in each fold. The best performance of the CNN-LSTM system was verified using the pre-trained VGG16 convolutional base, achieving a classification accuracy of 86.8%. The second and third best results were obtained by using the convolutional base of the VGG19 and ResNet50, respectively. In contrary to what could be expected by inspecting Table 6.1, the proposed system had the worst performance when using the Xception convolutional base. While Xception reported the best results on the ImageNet dataset among the 4 considered CNNs, a better generalization capacity of the VGG architectures was observed for this transfer learning application, as the spatial features extracted from the pre-trained VGG16 and VGG19 convolutional bases supported higher classification rates.

<table>
<thead>
<tr>
<th>Convolutional Base</th>
<th>ResNet50</th>
<th>Xception</th>
<th>VGG16</th>
<th>VGG19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Accuracy</td>
<td>83.2%</td>
<td>80.2%</td>
<td><strong>86.8%</strong></td>
<td>84.2%</td>
</tr>
</tbody>
</table>

Table 6.2: 10-fold cross-validation average classification accuracies of the proposed CNN-LST system, using different convolutional bases from top performing pre-trained CNNs.

### 6.3 Fine-tuning the Convolutional Base

The previous CNN-LSTM experiments achieved considerably high classification accuracies. However, the state-of-the-art CNN solution used in Section 5 still was able to get better results for GEI and SEI inputs on the GAIT-IT dataset. While it is also based on the use of a convolutional base from a pre-trained network, the VGG19 model, the main difference is that the convolutional base is fine-tuned with the training data. This means that the convolutional base is re-trained to handle the GEI or SEI inputs for feature extraction, yielding more useful spatial information in the context of the current task.

In the following experiments, the convolutional base of the proposed CNN-LSTM system is fine-tuned with the binary silhouettes frames in the training set. This allows the extraction of spatial features more specialized on the given inputs, providing the system with better visual representations, relevant for the current context of classification. The pre-trained convolutional bases of the VGG16 and VGG19 networks are considered for fine-tuning since it was with these networks that the previous experiments (see Section 6.2) obtained better results. To fine-tune each of the VGG convolutional bases, a dense network is stacked upon the convolutional layers to perform classification. As in Section 5, this network consists of a set of fully connected layers, with two layers of 4096 units and a softmax layer with 5 units, corresponding to the number of classes that represent the 5 different gait types in the GAIT-IT dataset. Only the last 3 convolutional block are fine-tuned from both convolutional bases. In both cases, this corresponds to freezing the first 4 convolutional layers, that comprise the first two convolutional blocks, and allowing the parameters of the following layers to be updated during the training process with silhouette key frames. After fine-tuning, the last two fully connected layers of the classifier are removed, leaving the first one to output a flattened 1D feature vector of size 4096. The resulting convolutional base is then used in the proposed CNN-LSTM system to train the LSTM and classifier.

To compare the performance of the system using each of the VGG convolutional bases, 10-fold cross-validation is implemented as in the previous section (see Section 6.2). The corresponding results are shown in the following Tables 6.3 and 6.4, in the form of a confusion matrix, detailing the classification.
accuracies for each class as the average of the 10 folds.

<table>
<thead>
<tr>
<th>True Class</th>
<th>Diplegic</th>
<th>Hemiplegic</th>
<th>Neuropathic</th>
<th>Normal</th>
<th>Parkinson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplegic</td>
<td>0.91</td>
<td>0.04</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
</tr>
<tr>
<td>Hemiplegic</td>
<td>0.01</td>
<td>0.95</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neuropathic</td>
<td>0</td>
<td>0.02</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>Parkinsonian</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.98</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.3: Confusion matrix containing the classification accuracies of the proposed system on the GAIT-IT dataset, averaged from the 10-fold cross-validation. These results are obtained by fine-tuning the VGG16 convolutional base of the CNN-LSTM system. The overall accuracy was 96.2%.

<table>
<thead>
<tr>
<th>True Class</th>
<th>Diplegic</th>
<th>Hemiplegic</th>
<th>Neuropathic</th>
<th>Normal</th>
<th>Parkinson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplegic</td>
<td>0.92</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Hemiplegic</td>
<td>0.02</td>
<td>0.94</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neuropathic</td>
<td>0</td>
<td>0.02</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>Parkinsonian</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 6.4: Confusion matrix containing the classification accuracies of the proposed system on the GAIT-IT dataset, averaged from the 10-fold cross-validation. These results are obtained by fine-tuning the VGG19 convolutional base of the CNN-LSTM system. The overall accuracy was 96.2%.

The overall accuracies can be computed as the mean values of the matrix diagonals, which consists in taking the average of the correct classification rates for each class. In both cases, an overall classification accuracy of 96.2% was achieved on the GAIT-IT dataset, leading to the conclusion that the increased number of layers, and thus parameters, in the VGG19 did not present an improvement in the performance of the system.

These results show a significant improvement in the performance of the proposed CNN-LSTM system when the pre-trained networks used as its convolutional base are fine-tuned on the training data. By comparison with the results Table 6.2, this improvement translates into an increment of 9.4% and 12% in the overall classification accuracy, regarding the use of the VGG16 and VGG19 convolutional bases, respectively.

The proposed system was able to achieve a higher classification accuracy than the state-of-the-art solution described and implemented in Section 5, based on a fine-tuned VGG19 to perform pathological gait classification using GEI and SEI gait representations as inputs. Using silhouette frames, the CNN-LSTM system was able to outperform the results of that same solution based on either GEI or SEI inputs by 2.2% and 2.6%, respectively. This shows the advantage of analysing gait as a sequence, using silhouette frames obtained from key moments in a gait cycle. By combining the deep learning techniques of convolutional and recurrent neural networks, the proposed system is able to synthesize meaningful spatial features from the input images and extract temporal information from the development of the observed motion.
6.4 Using a Bidirectional LSTM

Bidirectional recurrent neural networks [79] consist of two recurrent networks processing the same input simultaneously, where one handles the given sequence in its natural order and the other in reverse. It is a widely used approach [81, 82, 106, 107] in contexts where the entire input sequence is available at prediction time, designed to provide RNN based systems with more information from that same sequence. Given that the whole gait sequence is used as input for classification in the current approach to pathological gait analysis, a bidirectional LSTM was implemented to extend the LSTM network at the second stage of feature extraction in the proposed system. To achieve this, the LSTM network from the previous experiments (described in Section 6.1.3) is duplicated to process a given gait sequence forwards and backwards. The output of both LSTM networks in the bidirectional ensemble is combined into a spatiotemporal feature vector with size 512. It corresponds the concatenated output vectors of size 256 from the last LSTM cell of each LSTM network, illustrated in Figure 6.8. As stated in Section 6.1.4 regarding the architecture of the classification module, the dimension of the first fully connected layer in the classifier is set to have the same dimension as the feature vector outputted from the LSTM network. Thus, when implementing the bidirectional LSTM, the size of that fully connected layer as well as the dropout layer that follows is changed from 256 to 512.

In this experiment, the convolutional base used is the fine-tuned VGG16 from the previous Section 6.3. It was chosen over the VGG19 counterpart since it has less parameters and therefore requires less computational memory and time to extract spatial features from the input images. As in Section 6.2, 10-fold cross-validation is used to evaluate the performance of the system with the fine-tuned VGG16 convolutional base and bidirectional LSTM network combined for feature extraction (CNN-BiLSTM). The following Table 6.5 presents a confusion matrix with the classification results, obtained as the average of the 10 folds.

![Figure 6.8: Illustration of the bidirectional LSTM implemented in the proposed system.](image)

The overall accuracy achieved was 96.5%, computed as the mean of the correct classification rate for each class. It represents an increase of 0.3% regarding the results obtained with a unidirectional single LSTM network in the previous Section 6.3. While only a slight gain in the system’s performance was
verified, this experiment can lead to the argument that RNN solutions for gait analysis can benefit from bidirectional implementations, as some information regarding the temporal dependencies can be easier to apprehend by considering the input in both directions.

6.5 Building a Convolutional Base For Pathological Gait Classification

So far, in this work, transfer learning (see Section 3.2.2 for a more detailed review of this technique) has been applied in every experiment through the use of pre-trained CNNs for feature extraction. Since the VGG19 approach using GEI and SEI gait representations in Section 5, up to the CNN-BiLSTM system using a fine-tuned VGG16 proposed in the last Section 6.4, these deep learning solutions have been taking advantage of the visual features that top performing CNN models learned using millions of images from the ImageNet database [65]. The main purpose is to tackle the issue of having a relatively small amount of pathological gait image data to train a convolutional base, when compared to the number of samples that such databases offer for image recognition.

Transfer learning can be especially determinant when dealing with inputs such as the GEI and SEI, which are obtained from compact single image representations of a gait cycle. However, since the proposed CNN-LSTM based systems use 9 silhouette frames per gait cycle directly as inputs, the number of samples available to train a convolutional base is now 9 times higher (a total of 34335 images) when compared to the amount of available GEIs or SEIs computed from a single gait cycle. This considerable increase in the amount of training data supported the hypothesis that a CNN could be built specifically to serve as the convolutional base of the proposed system and trained using only binary frames obtained from pathological gait video sequences.

6.5.1 Proposed PGait CNN Architecture

As there are no general rules for the development of artificial neural networks, the best approaches depend on the context of the given task. New solutions can be based on the empirical knowledge that led to the introduction of notable architectures, such as those reviewed in Section 6.2 [66, 67, 69]. Besides the top performing networks in the ImageNet competition [57], state-of-the-art networks [108] designed for the Kaggle MNIST challenge [109] were also considered to determine the parameters of the proposed PGait CNN architecture. The MNIST database [110, 111] is a database containing a large amount of
handwritten digits represented as binary images, just as the silhouette frames in the proposed system. It is composed of a training and testing set with 60,000 and 10,000 images respectively. This database has been widely used to establish benchmarks for CNN architectures and deep learning techniques in image recognition tasks.

The proposed PGait CNN architecture for image processing in pathological gait classification is illustrated in Figure 6.9. Its main characteristics are as follows:

- **Network Depth**: The PGait CNN network has 5 convolutional layers and can be considered a shallow network when compared to those used in Section 6.2. Besides the fact that the current classification task has a much smaller scale than the ImageNet challenge, the main reason why such a reduction can be achieved is that those previous CNNs are designed to handle colored image inputs, as three-dimensional tensors. Since the proposed model is built to deal with binary frames, without the color dimension, the complexity of the learned features is reduced and thus fewer layers are required [108]. Furthermore, unnecessarily increasing network capacity, such as increasing the number of layers, is a main factor for causing overfitting.

- **Convolution Kernel**: In light of the work that introduced the VGG CNN architectures, all the convolutional layers in proposed PGait CNN use very small filters with a receptive field of $3 \times 3$. A stride of $2 \times 2$ is defined, meaning that these filters slide two columns or two rows at a time, as they traverse the input to perform the convolution operation. This stride size results in a 50% dimensionality reduction of the input at each layer. Thus, at the end of the PGait CNN, after 5 convolutional layers, the $224 \times 224$ input images are transformed in $7 \times 7$ feature maps.

- **Feature Maps**: The number of filters applied to the input of each layer determines the number of feature maps that same layer yields as output. Starting with 32 filters at the image input, this number doubles for the last 2 layers. After 5 layers, the proposed PGait CNN yields 64 feature maps of size $7 \times 7$ as the output.

- **Batch Normalization**: Each convolutional layer is followed by a batch normalization layer, that adjusts and scales its outputs to have a mean value close to 0 and a standard deviation close to 1. Bounding the values that pass between layers helps to stabilize and speed up the training process.

With the described network specifications, the proposed PGait CNN architecture has a total of 74,688 trainable parameters involved in outputting 64 spatial features of size $7 \times 7$ from a $224 \times 224$ binary input image.

### 6.5.2 Performance Evaluation

The first step is to train the PGait CNN with silhouette frames, extracted from GAIT-IT dataset, for the given training set using the key frame selection module presented in Section 6.1.1. To achieve this, its output is flattened and a dense network is stacked upon the CNN, to perform the classification of every silhouette frame into the respective gait type. This network consists of two fully connected layers with dropout in between them. The first layer has a size of 512 while the second has 5 units, with a softmax activation to output class probabilities. After training, the presented PGait CNN architecture is used as the convolutional base of the proposed CNN-BiLSTM system.
Figure 6.9: Illustration of the proposed PGait CNN architecture for feature extraction in pathological gait classification.

The performance of this ensemble is evaluated with 10-fold cross-validation, as done in 6.2. In each fold, the training set is used to first train the convolutional base, as previously described, for 50 epochs. At every epoch, its performance is evaluated on the test set and the model is saved at the best one. Shown in Figure 6.10 are the training and test set accuracies averaged across the 10 folds for every epoch. The results of the CNN-BiLSTM system with the proposed convolutional base are presented in Table 6.6, with a confusion matrix containing the average classification accuracies across the 10 folds. The described system was able to achieve state-of-the-art results, with an overall accuracy of 93.8%. These classification rates are very similar to the VGG19 solution based on GEI and SEI inputs described in Section 5, while below (-2.7%) those reached by the CNN-BiLSTM system with a fine-tuned VGG16 convolutional base.

Figure 6.10: Plot of the 10-fold average training and test accuracies along 50 epochs, corresponding to the training process of the proposed PGait CNN for pathological gait classification.

Regarding classification results alone, this solution does not present an improvement in performance
Table 6.6: Confusion matrix with the classification accuracies of the CNN-BiLSTM system on the GAIT-IT dataset using the proposed CNN as the convolutional base. The results are computed as the average accuracies of the 10 folds from the cross-validation and the overall accuracy was 93.8%. It represents a state-of-the-art classification rate achieved with the proposed CNN for gait analysis on $224 \times 224$ binary frame inputs.

Towards pathological gait classification. However, a clear advantage can be verified by considering the trade-off between classification accuracy and overall system complexity. By using a comparatively shallow CNN as the convolutional base, built specifically for the task at hand, the number of trainable parameters is significantly reduced. In addition, having the CNN yield more compact features will also require less parameters to implement an RNN in a second stage of feature extraction and finally a classifier network.

Significant parameter reduction also means less static memory to store the CNN model and less dynamic memory to execute it. This is an important advantage when considering the possible applications for gait analysis systems. An example is the integration of these solutions on mobile devices, which are characterized by a trade-off between computational power and portability.

In Table 6.7, the total number of parameters is compared among the pathological gait classification systems described so far, as well as the file size associated with storing each one. The CNN-BiLSTM system with the proposed convolutional base has around 36 times less parameters than the VGG16-LSTM network from Section 6.3 and 41 times less parameters than the VGG19 network from Section 5. Using a method from the Keras API [112] for Python to save network models, each system is stored in memory in the HDF5 [113] file format. Proportional to the number of parameters, the file size of the CNN-BiLSTM system with the proposed convolutional base is 41 and 36 times smaller than the VGG19 network from Section 5 and the VGG16-LSTM network from Section 6.3, respectively.

Table 6.7: Comparison of the number of parameters among the different systems presented since Section 5. The CNN-BiLSTM system using the proposed convolutional base achieves a reduction of 36 to 41 times in the number of parameters and model size in megabytes.

As a consequence of reducing the number trainable parameters, the training process or fine-tuning of the convolutional base is also significantly faster. This is a bottleneck in terms of training time in the proposed CNN-LSTM and CNN-BiLSTM systems. After the convolutional base is trained or fine-tuned,
training the LSTM and classifier networks is considerably faster (between 10 and 15 times faster) since the number of trainable parameters left is much smaller. The time it takes to execute the network is also greatly reduced. It characterizes the performance of the trained network in outputting predictions for the given inputs.

In the first part of Table 6.8, the time involving training and fine-tuning, as well as executing the different networks is presented as duration per sample, given that the total duration depends on the number of training or testing samples. By comparison, the CNN-BiLSTM system using the proposed convolutional base is between 6 and 8 times faster during the training process and 6 times faster in yielding classification predictions than its counterparts.

<table>
<thead>
<tr>
<th>Pathological Gait Classification System</th>
<th>Input Type</th>
<th>Duration (milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG19 (Section 5)</td>
<td>GEI/SEI</td>
<td>15</td>
</tr>
<tr>
<td>VGG16-LSTM (Section 6.3)</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>VGG16-BiLSTM (Section 6.4)</td>
<td>Binary Silhouettes</td>
<td>13</td>
</tr>
<tr>
<td>PGait CNN-BiLSTM (Section 6.5)</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

Table 6.8: Training and execution duration compared among the different systems presented since Section 5. The CNN-BiLSTM system with the proposed convolutional base is 6 to 8 times faster in training and 6 times faster in yielding predictions than its counterparts.
7 Experimental Results

The gait analysis framework presented in the previous Section 6 was evaluated under different variations (such as fine-tuning a pre-trained CNN, using a bidirectional LSTM and implementing the proposed PGait CNN as the convolutional base), using a cross-validation strategy on the developed GAIT-IT pathological gait dataset. In Section 7.1 a compilation and analysis of those cross-validation results is presented. In Section 7.2, the GAIT-IST pathological gait dataset is used along with GAIT-IT to perform cross-database tests and further evaluate system performance and generalization capability. A final performance assessment is done in Section 7.3 to compare classification results among each type of gait.

7.1 Cross-validation Results on GAIT-IT

The framework for pathological gait classification relying on the combination of convolution and recurrent deep neural networks, generically denoted as CNN-LSTM, introduced in Section 6, was implemented with several variants, to continually attempt an improvement in classification performance. In each of those experiments, the performance was evaluated on GAIT-IT, through 10-fold cross-validation as defined in the second part of Section 5.2. Thus, these results are directly comparable to those obtained with the solution described in Section 5 (shown in Table 5.1), which consisted of a fine-tuned VGG19 network, adapted to perform pathological gait classification and using either GEI or SEI inputs, and which is here considered as a baseline reference for comparisons. The overall classification rates of this baseline VGG19 approach, as well as those achieved with each variation of the proposed CNN-LSTM system after Section 6.2, were compiled in the following Table 7.1.

The second best classification rate on GAIT-IT of 96.2%, achieved by the VGG16-LSTM model from Section 6.3, represents a 2.2% increase in performance compared to the best result obtained by the baseline VGG19 solution from Section 5 based on GEI inputs. While both systems have a fine-tuned VGG network as the convolutional base, the proposed framework supported an improvement in performance by incorporating an LSTM together with the CNN at the feature extraction module. Featuring the implementation of bidirectionality to extend the LSTM network, the best result is obtained by the VGG16-BiLSTM system presented in Section 6.4, with an overall classification accuracy of 96.5%. Using the simpler PGait CNN described in Section 6.5, developed for the convolutional base of the proposed architecture, the CNN-BiLSTM system achieved an overall accuracy of 93.8%. It stands between the 93.6% and 94% accuracies reached by the VGG19 approach using SEI and GEI inputs, respectively. This can be argued to be a positive trade-off between system complexity and performance, given that state-of-the-art results can be achieved and the number of model parameters was substantially reduced.

7.2 Cross-dataset Tests on GAIT-IST

To further evaluate and compare the proposed framework with the state-of-the-art VGG19 baseline solution, based on GEI and SEI gait representations, a set of generalization cross-dataset tests were
Cross-validation Results

<table>
<thead>
<tr>
<th>Pathological Gait Classification System</th>
<th>Input Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG19 (Section 5)</td>
<td>GEI</td>
<td>94.0%</td>
</tr>
<tr>
<td></td>
<td>SEI</td>
<td>93.6%</td>
</tr>
<tr>
<td>VGG16-LSTM (Section 6.3)</td>
<td></td>
<td>96.2%</td>
</tr>
<tr>
<td>VGG16-BiLSTM (Section 6.4)</td>
<td>Binary Silhouettes</td>
<td>96.5%</td>
</tr>
<tr>
<td>Proposed CNN-BiLSTM (Section 6.5)</td>
<td></td>
<td>93.8%</td>
</tr>
</tbody>
</table>

Table 7.1: Summary of the cross-validation results on GAIT-IT, obtained with the best solutions described in Section 5 and in Section 6. The highest classification rate of 96.5% was achieved by the CNN-BiLSTM with a pre-trained and fine-tuned VGG16 convolutional base.

conducted with the GAIT-IST database. As done in the second part of Section 5.3, each model is trained using all 23 subject simulations from GAIT-IT and tested on all 10 subjects from GAIT-IST. For the VGG19 approach, both results obtained using GEI and SEI inputs can be directly extracted from Table 5.1 in Section 5. Regarding the CNN-LSTM based systems from Section 6, classification was performed on binary silhouette image sequences. While GAIT-IST contains data with considerable quality, it still is affected by some segmentation errors in individual binary silhouettes, as illustrated with the examples in Figure 7.1. As described in the state-of-the-art review (see Section 3.2.1), GEI and SEI representations are robust against noise in individual images in a gait sequence. However, when using binary frames directly as inputs to the CNN-LSTM based systems, these errors were not mitigated in any way. Therefore, by conducting these cross-dataset tests on GAIT-IST, the models trained with the high quality binary silhouettes in GAIT-IT are also tested on their capacity to perform on noisy data.

The cross-dataset test results are included in Table 7.2. A significant improvement to the state-of-the-art VGG19 baseline solution was achieved by the CNN-LSTM system with a fine-tuned VGG16 as the convolutional base, described in Section 6.3. Moreover, with the bidirectional LSTM implementation from Section 6.4, the overall performance was further improved and the top classification accuracy of 91.4% was achieved on GAIT-IST. This represents an increase of 5% compared to the best result (86.4%) obtained by the state-of-the-art VGG19 baseline approach.

When evaluated on GAIT-IT with cross-validation, the CNN-LSTM with the proposed convolutional base (see Section 6.5) achieved very similar results to the VGG19 baseline solution, as illustrated in Table 7.1. However, in the cross-dataset tests on GAIT-IST, it clearly outperformed the VGG19 baseline system, based on GEI and SEI inputs, by 2.2% and 3.3%, respectively, suggesting a greater generalization ability towards new data. In addition to the advantage of providing a significant reduction in network complexity,
Cross-dataset Results

<table>
<thead>
<tr>
<th>Pathological Gait Classification System</th>
<th>Input Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG19 (Section 5)</td>
<td>GEI</td>
<td>86.4%</td>
</tr>
<tr>
<td></td>
<td>SET</td>
<td>85.1%</td>
</tr>
<tr>
<td>VGG16-LSTM (Section 6.3)</td>
<td></td>
<td>90.3%</td>
</tr>
<tr>
<td>VGG16-BiLSTM (Section 6.4)</td>
<td>Binary Silhouettes</td>
<td>91.4%</td>
</tr>
<tr>
<td>Proposed CNN-BiLSTM (Section 6.5)</td>
<td></td>
<td>88.4%</td>
</tr>
</tbody>
</table>

Table 7.2: Summary of the cross-dataset results using GAIT-IT as the training set and GAIT-IST as the testing set. The classification accuracies were obtained with the best solutions described in Section 5 and in Section 6. The highest classification rate of 91.4% was achieved by the CNN-BiLSTM with a pre-trained and fine-tuned VGG16 convolutional base.

these results further support the use of the proposed PGait CNN for gait analysis. Furthermore, this reinforces the relevance of the developed pathological gait dataset since the respective network was solely trained using GAIT-IT, as opposed to having pre-trained parameters.

As a final remark, each CNN-LSTM based system in Table 7.2 was able to surpass state-of-the-art results in the cross-dataset tests, despite the presence of segmentation errors in many binary silhouettes from GAIT-IST, as those in Figure 7.1.

7.3 Performance Comparison on Different Gait Types

The current task of pathological gait analysis consists in the classification of an observed gait into 4 different gait related pathologies and normal gait. So far, in the previous sections, different approaches have been evaluated according to their overall classification accuracy, computed as the average of the correct classification rates for each type of gait. To follow up this analysis, it is also important to assess how those approaches perform on each individual gait type. This can provide meaningful information about which gait class is easier to learn and which proves to be more challenging. It may also provide valuable insights about how the learning process for one type of gait can be affected by another.

In Table 7.3, the classification accuracies obtained for each gait type are shown for every system considered in the previous Table 7.2. These accuracies were extracted from the 10-fold cross-validation experiments conducted with the the respective systems.

To further detail system performance according to each gait type, the average accuracies computed from the overall 10-fold cross-validation confusion matrix of each system are shown in Table 7.4. This allows the visualization of most common mislabeled predictions and brings forth valuable insights about the main difficulties in distinguishing gait among the different gait types.

A first observation can be made regarding the classification of normal gait. It represents the highest classification rate among all gait types and the most consistent result with 99% accuracy achieved by each system. Classification of diplegic gait, opposed normal gait, constitutes the lowest accuracy for each solution, achieving an average value of 89%. It can be observed that this gait pathology is mainly mislabeled as hemiplegic or Parkinsonian in roughly 10% of the total predictions, which can be argued to be the most challenging differentiation for all the considered systems. Hemiplegic gait has an average of 92% correct classification rate, primarily mistaken for diplegic and neuropathic pathologies.
The classification of gait as neuropathic has an average accuracy of 97%, close to that obtained for normal gait and the same as with Parkinsonian gait, which is mainly mislabeled for diplegic gait. This goes according to the fact that the most notable similarities between the simulated gait pathologies in the considered datasets are among the diplegic and Parkinsonian gait, both characterized by a stooped posture and relatively small step lengths.

Table 7.3: Classification accuracies according to each gait type obtained by the proposed CNN-LSTM systems and the state-of-the-art VGG19 solution.

<table>
<thead>
<tr>
<th>System</th>
<th>Diplegic</th>
<th>Hemiplegic</th>
<th>Neuropathic</th>
<th>Normal</th>
<th>Parkinsonian</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG19 GEI</td>
<td>0.89</td>
<td>0.90</td>
<td>0.97</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>VGG19 SEI</td>
<td>0.86</td>
<td>0.91</td>
<td>0.98</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>VGG16-LSTM</td>
<td>0.91</td>
<td>0.95</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>VGG16-BiLSTM</td>
<td>0.92</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>CNN-BiLSTM</td>
<td>0.87</td>
<td>0.90</td>
<td>0.96</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>Average</td>
<td>0.89</td>
<td>0.92</td>
<td>0.97</td>
<td>0.99</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 7.4: Confusion matrix containing the overall 10-fold cross-validation classification accuracies averaged across every system considered in Table 7.3.

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicated Class</th>
<th>Diplegic</th>
<th>Hemiplegic</th>
<th>Neuropathic</th>
<th>Normal</th>
<th>Parkinson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplegic</td>
<td></td>
<td>0.89</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Hemiplegic</td>
<td></td>
<td>0.03</td>
<td>0.92</td>
<td>0.03</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>Neuropathic</td>
<td></td>
<td>0</td>
<td>0.02</td>
<td>0.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>Parkinsonian</td>
<td></td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.97</td>
</tr>
</tbody>
</table>
8 Conclusions

This work is focused on the application of deep learning techniques for the task of pathological gait analysis using 2D video data acquired by a single camera. Section 8.1 summarizes the main achievements presented in this work and Section 8.2 introduces the main propositions for future developments.

8.1 Achievements

In this work, the GAIT-IT dataset is proposed with a total of 828 sequences, featuring 21 subjects performing simulations of 4 different gait related pathologies, namely, diplegia, hemiplegia, neuropathy and Parkinson’s disease, besides normal gait. Recorded in a professional studio, the dataset provides high-quality gait representations from which accurate gait features can be extracted and used for pathological gait classification, as exemplified in Section 4.

A framework for pathological gait classification is developed to process video-based gait data. It consists in the combination of convolutional and recurrent deep neural networks, denoted as a CNN-LSTM architecture, for the extraction of spatial and temporal gait features from sequences of binary frames. The VGG16-BiLSTM implementation from Section 6.4 of the proposed CNN-LSTM framework is able to achieve a top classification accuracy of 96.5% in cross-validation tests on GAIT-IT. When evaluated with cross-dataset tests, it was able to achieve a very interesting overall accuracy of 91.4%, using GAIT-IT for training and the GAIT-IST dataset for testing. These cross-dataset tests allowed the assessment of system performance regarding generalization and the ability to perform on noisy data, given the presence of segmentation errors on the binary silhouettes available from GAIT-IST.

A new, simpler, CNN architecture is proposed for gait analysis on binary gait frames, denoted as PGait CNN. While state-of-the-art deep learning solutions are based on transfer learning, using pre-trained and fine-tuned CNNs, the developed PGait CNN is built specifically to process $224 \times 224$ binary images. When compared to state-of-the-art solutions, the PGait CNN achieves a drastic reduction in the number of model parameters of the network, which means a significantly lower amount of required memory, as well as lower training and execution times. The proposed PGait CNN is then used as the convolutional base of a CNN-BiLSTM solution, providing an alternative implementation of the proposed CNN-LSTM framework. Evaluated through cross-validation on GAIT-IT, it achieved state-of-the-art results, indicating a positive trade-off between network complexity and performance.

8.2 Future Work

This work contributed with the development of a new pathological gait dataset, larger than those currently available and with high quality video-based data. As described in Section 4.3, it comprises gait sequences recorded with two synchronized cameras, capturing both sagittal and frontal points of view. Since this work focused on the use of gait sequences from a sagittal view, a future development can be the integration of frontal view analysis. By combining these orthogonal view points for the analysis of pathological gait, a strong hypothesis may state that the amount of meaningful information is consider-
ably increased, leading to an improvement in the performance of gait analysis systems. Furthermore, in
[114], an alternative LSTM architecture that allows the simultaneous processing of multiple synchronous
sequences was developed in the context of facial recognition. The proposed CNN-LSTM framework can
thus be implemented to incorporate this solution and provide the system with the simultaneous frontal
and sagittal view analysis.

The proposed dataset contains an alternative source with lower quality of the acquired video se-
dquences. A cellphone camera was placed alongside the professional camera to record the gait simula-
tion in the sagittal view. This can be explored in a future work to analyse the impact of video resolution
and small view point changes in gait analysis, relating to a more unconstrained environment for practical
purposes.

As seen with the cross-dataset tests conducted on GAIT-IST [94] in Section 7.2, gait analysis frame-
works such as the proposed CNN-LSTM that process gait as a sequence of individual frames often have
to perform in the presence of segmentation errors. This is especially true in more unconstrained environ-
ments as those that characterize the gait sequences from the DAI [42] and DAI2 [26] gait dataset. Thus,
future contributions can be made with the development of automatic methods to improve the quality of
affected binary gait images, e.g. by considering previous and following frames in the respective gait
sequence for comparison.

Finally, a proposition for future developments is the use of a convolutional LSTM (ConvLSTM), detailed
in Section 3.2.2, as an alternative to the LSTM network in the feature extraction module of the proposed
CNN-LSTM framework from Section 6. ConvLSTM extends the LSTM model with convolutional oper-
ations to replace internal matrix multiplications. This allows ConvLSTM cells to maintain the 2D input
dimension as the information flows through the network, which can render this architecture more suited
to capture spatiotemporal correlations [83].
References


[99] Keras Applications. URL https://keras.io/api/applications/.


