



UNIVERSIDAD  
DE MÁLAGA

Escuela de Ingenierías Industriales  
Programa de Doctorado en Ingeniería Mecatrónica

## TESIS DOCTORAL

# Automatic Extraction of Biometric Descriptors Based on Gait

Rubén Delgado Escaño

Julio de 2022

Dirigida por:  
Nicolás Guil Mata,  
Manuel J. Marín Jiménez





AUTOR: Rubén Delgado Escaño

 <https://orcid.org/0000-0002-2365-6593>

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional:

<http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

Cualquier parte de esta obra se puede reproducir sin autorización pero con el reconocimiento y atribución de los autores.

No se puede hacer uso comercial de la obra y no se puede alterar, transformar o hacer obras derivadas.

Esta Tesis Doctoral está depositada en el Repositorio Institucional de la Universidad de Málaga (RIUMA): [riuma.uma.es](http://riuma.uma.es)



Dr. D. Nicolás Guil Mata.

Catedrático del Departamento de Arquitectura de Computadores de la Universidad de Málaga.

Dr. D. Manuel J. Marín Jiménez.

Profesor Titular del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba.

**CERTIFICAN:**

Que la memoria titulada “Automatic Extraction of Biometric Descriptors Based on Gait”, ha sido realizada por D. Rubén Delgado Escaño bajo nuestra dirección en el Departamento de Arquitectura de Computadores de la Universidad de Málaga y constituye la Tesis que presenta para optar al grado de Doctor en Ingeniería Mecatrónica.

Málaga, Julio de 2022

Dr. D. Nicolás Guil Mata.  
Codirector de la tesis.

Dr. D. Manuel J. Marín Jiménez.  
Codirector de la tesis.

UNIVERSIDAD  
DE MÁLAGA





## DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña RUBÉN DELGADO ESCAÑO

Estudiante del programa de doctorado EN INGENIERÍA MECATRÓNICA de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: AUTOMATIC EXTRACTION OF BIOMETRIC DESCRIPTORS BASED ON GAIT

Realizada bajo la tutorización de NICOLÁS GUIL MATA y dirección de NICOLÁS GUIL MATA Y MANUEL JESÚS MARÍN JIMÉNEZ (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 11 de JUNIO de 2022

Fdo.: RUBÉN DELGADO ESCAÑO Doctorando/a	Fdo.: NICOLÁS GUIL MATA Tutor/a
Fdo.: NICOLÁS GUIL MATA, MANUEL JESÚS MARÍN JIMÉNEZ Director/es de tesis	



UNIVERSIDAD  
DE MÁLAGA



Dr. D. Nicolás Guil Mata.

Catedrático del Departamento de Arquitectura de Computadores de la Universidad de Málaga.

Dr. D. Manuel J. Marín Jiménez.

Profesor Titular del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba.

**CERTIFICAN:**

Que autorizan la lectura de la tesis del doctorando D. Rubén Delgado Escaño, titulada “Automatic Extraction of Biometric Descriptors Based on Gait” y que ninguna de las publicaciones que avalan dicha tesis ha sido utilizada en tesis anteriores.

Málaga, Julio de 2022

Dr. D. Nicolás Guil Mata.  
Codirector de la tesis.

Dr. D. Manuel J. Marín Jiménez.  
Codirector de la tesis.

UNIVERSIDAD  
DE MÁLAGA



A los táperes de mi madre



UNIVERSIDAD  
DE MÁLAGA



# Agradecimientos

---

Después de cuatro años de intenso trabajo, finaliza un periodo de aprendizaje académico y personal con la escritura de esta tesis. Realizar el doctorado ha tenido un gran impacto en mí, y es por eso que quiero agradecer a aquellos que me han ayudado, y aguantado, durante este caótico proceso.

Por parte académica, agradecer a mis directores Nicolás y Manuel Jesús, que me han apoyado enormemente cuando lo necesitaba. A Nicolás, me gustaría agradecerle todo el tiempo que me ha dedicado, ya fuera pensando nuevas líneas de investigación, ayudándome con los tediosos procesos administrativos o en los duros días clave para un deadline o una revisión. Su esfuerzo en el desarrollo de esta tesis ha sido vital para su finalización. A Manuel Jesús, agradecerle todo lo que ha aportado de su vasto conocimiento y experiencia, siempre aportando ideas e intentando que todo el trabajo realizado quede lo mejor posible. Os agradezco vuestra cooperación y por darme esta oportunidad tan importante para mí, me habéis brindado todas las herramientas necesarias para completar esta tesis y más.

Agradezco también a Julián por haberme introducido en la investigación, pues confió en mí y me dio mi primera oportunidad en este mundillo a pesar de mi falta de base en un campo tan complejo como es el Machine Learning. Debo agradecer también a Fran, que más que un compañero de laboratorio ha sido un maestro y amigo. Sin él, sus lecciones, y su paciencia, todo este proceso hubiese sido una cuesta mucho más empinada.

I would also like to thank Karteek for the opportunity to make my research internship with him in an internationally renowned group, allowing me to learn a lot and initiate a new research line.

En lo personal, doy las gracias a mi familia por su apoyo constante y su capacidad de aguante. Agradezco que todavía no me hayan echado de casa para montar una despensa en mi cuarto, y agradezco los táperes de mi madre, que me



han acompañado durante mis 11 años de universidad.

Le doy las gracias a los compañeros de laboratorio, y a los del otro, que han coincidido conmigo en estos años en el departamento y con los que he tomado muchos cafés mientras hacíamos de patitos de goma unos de otros.

Finalmente, agradezco a mis amigos, muy pocos pero muy buenos, el cariño incondicional que me han brindado. No era su obligación escucharme o ayudarme a mantenerme en pie, pero ellos han estado ahí haciendo todo lo que estaba en su mano pese a los altibajos a los que me he tenido que enfrentar. Una segunda familia con la que puedo escapar cuando la vida no me da para más.

# Abstract

---

Nowadays, people identification is a topic of interest due to its implications in terms of safety, service automation and sanitary control. Historically, people have been identified by using its face, iris or fingerprints, and these processes have been automated to identify subjects without the intervention of other humans, with more accurate results and in less time. However, those kinds of systems require the collaboration of the subject to be identified, which implies a problem in some scenarios where collaboration is impossible.

Due to this, gait recognition is presented as an alternative in the field of people recognition, since it does not require the cooperation of the subject, or even the knowledge that they are being identified. It can be done at a certain distance and it is a difficult method to deceive or avoid, since a mask, hood or other typical blocking objects would not deceive the recognition system. Also, gait recognition has low hardware requirements in the process of data capture. It allows the use of cheap security cameras that can be located in high places to cover a large visible space.

Gait recognition has been studied more in detail in recent years. Published papers on this subject are becoming more common and, over time, with better results in more complex datasets. However, not only scientific studies are being proposed for gait recognition. Real applications are being used in Japan where gait recognition is considered as a criminal probe to identify the causer, and has been proposed as a possible security measure at airports and government buildings.

However, the study of gait recognition is not exempt from challenges and problems yet to be solved. This thesis focuses on studying and resolving these points that we believe have not been sufficiently addressed. Firstly, we study the viability of soft-biometric classification in gait recognition, that is, the recognition of human characteristics, such as age and gender, by relying on the recognition of identity. For this, we worked with a multi-task and multi-modal model that

was used as input data from inertial sensors placed on the subject, such as those that could be found in a smartphone. This model has obtained results that mark a new state-of-the-art in identity, age and gender for the dataset used. Secondly, we address the problem of missing data in a dataset. For this purpose, we have implemented a cross-dataset model that can jointly use multiple datasets with different subjects, captured with different sensors, and whose data have different lengths and sampling. This is successfully applied to fall detection for elderly people with inertial sensors, a use case where the datasets tend to have a higher number of young subjects, but the implemented model can generalize to subjects not seen in training. Thirdly, we have implemented a framework to solve a common problem in visual gait recognition datasets: the lack of multiple subject scenarios with occlusions, which adds realism to the data and complicates recognition. The framework, called MuPeG, uses existing datasets to create scenarios with multiple subjects that can visually block each other. In addition, tests have been performed on state-of-the-art datasets to test the impact of these new types of scenarios on recognition models. Fourthly, we propose a solution to the missing modality problem. This is a problem that affects multi-modal models, common models in gait recognition, when one or more of the input modalities are missing. Our solution, UGaitNet, uses a mechanism of logic gates, which activate or deactivate the input modes, and a merge function to join the active inputs. We have found that this allows the model to be used with missing modalities without affecting model performance. Finally, we use knowledge distillation to reduce the computational complexity of a model and its input data, by teaching a model with grayscale images to mimic the predictors obtained by a model using optical flow, a type of input that is more computationally complex but provides more information about the motion of the subject. A new approach called GaitCopy, tested on two state-of-the-art models, has been implemented for this purpose and provides similar performance on the master models and their student models.

In summary, this thesis has explored unconventional aspects of gait-based people identification and proposed new approaches for addressing this challenging problem.

# Contents

<b>Agradecimientos</b>	I
<b>Abstract</b>	III
<b>Contents</b>	v
<b>List of Figures</b>	vii
<b>1.- Introduction</b>	1
1.1. Thesis Motivation . . . . .	3
1.2. Objectives and Phases . . . . .	4
1.3. Contributions . . . . .	6
1.4. Thesis Structure . . . . .	7
<b>2.- Background</b>	9
2.1. Artificial Neural Networks . . . . .	9
2.2. Input Modalities . . . . .	11
2.2.1. Visual Modalities . . . . .	12
2.2.1.1 Raw pixels: RGB & Gray . . . . .	12
2.2.1.2 Silhouettes . . . . .	13
2.2.1.3 Optical Flow . . . . .	13

2.2.2. Inertial Modalities . . . . .	13
2.2.2.1 Accelerometer . . . . .	14
2.2.2.2 Gyroscope . . . . .	14
2.3. Data Fusion . . . . .	15
2.3.1. Early Fusion . . . . .	15
2.3.2. Late Fusion . . . . .	16
2.3.3. Intermediate Fusion . . . . .	17
2.4. Classifiers . . . . .	18
2.4.1. $k$ -Nearest Neighbors . . . . .	18
2.4.2. Fully-Connected Layer . . . . .	19
2.5. Multi-Task Learning . . . . .	19
2.6. Knowledge Distillation . . . . .	20
2.7. Datasets . . . . .	21
2.7.1. Inertial Datasets . . . . .	22
2.7.2. Visual Datasets . . . . .	23
2.7.3. Synthetic Datasets . . . . .	25
<b>3.- Related work</b>	<b>27</b>
3.1. Gait Recognition Inertial Approaches . . . . .	27
3.2. Gait Recognition Visual Approaches . . . . .	30
3.3. Multi-Task Approaches . . . . .	31
3.4. Multimodal Approaches . . . . .	32
3.5. Knowledge Distillation Approaches . . . . .	32
<b>4.- Published Work</b>	<b>34</b>
4.1. List of Published Papers . . . . .	34
4.2. Summary of the papers that support this thesis . . . . .	35
4.2.1. Reference [25] ‘An end-to-end multi-task and fusion CNN for inertial-based gait recognition’ . . . . .	35

4.2.2. Reference [28] ‘MuPeG—The Multiple Person Gait Framework’ . . . . .	36
4.2.3. Reference [24] ‘A cross-dataset deep learning-based classifier for people fall detection and identification’ . . . . .	36
4.2.4. Reference [27] ‘GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy’ . . . . .	37
4.3. Copies of the papers that support this thesis . . . . .	37
4.4. Additional papers . . . . .	42
4.4.1. Reference [26] ‘Multimodal Gait Recognition Under Missing Modalities’ . . . . .	42
4.4.2. Reference [75] ‘UGaitNet: Multimodal Gait Recognition With Missing Input Modalities’ . . . . .	42
4.5. Copy of the additional paper . . . . .	43
<b>5.- Conclusions</b>	<b>47</b>
5.1. Conclusions . . . . .	47
5.2. Future Work . . . . .	48
<b>Appendices</b>	<b>51</b>
<b>A.- Resumen en español</b>	<b>51</b>
A.1. Motivaciones de la Tesis . . . . .	52
A.2. Objetivos y Fases . . . . .	53
A.3. Contribuciones . . . . .	55
A.4. Publicaciones . . . . .	57
A.5. Resumen de los artículos que apoyan esta tesis . . . . .	58
A.5.1. Referencia [25] ‘An end-to-end multi-task and fusion CNN for inertial-based gait recognition’ . . . . .	58
A.5.2. Referencia [28] ‘MuPeG—The Multiple Person Gait Framework’ . . . . .	58

A.5.3. Referencia [24] ‘A cross-dataset deep learning-based classifier for people fall detection and identification’ . . . . .	59
A.5.4. Referencia [27] ‘GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy’ . . . . .	60
A.6. Publicaciones adicionales . . . . .	60
A.6.1. Referencia [26] ‘Multimodal Gait Recognition Under Missing Modalities’ . . . . .	61
A.6.2. Referencia [75] ‘UGaitNet: Multimodal Gait Recognition With Missing Input Modalities’ . . . . .	61
A.7. Conclusiones . . . . .	62
A.8. Trabajo Futuro . . . . .	63
<b>Bibliography</b>	<b>65</b>

# List of Figures

1.1. <b>Vision inputs in gait recognition.</b> Extracted from TUM-GAID dataset [47]. . . . .	2
1.2. <b>Inertial sensors in gait recognition.</b> The left part of the image shows a subject walking through a circuit with the sensor system in the waist. The right part of the image depicts the inertial information recorded during the walk. The top plot shows the measurements for the accelerometer and the bottom plot contains the measurements for the gyroscope. Images obtained from [85]. . . . .	3
2.1. <b>Typical CNN architecture.</b> Common architecture of a CNN, consisting of multiple convolutions and pooling layers concatenated together and ending with a classification layer. . . . .	10
2.2. <b>Recurrence in RNN architecture.</b> The left part of the image shows a classical feed-forward network. The right side shows a recurrent network, whose prediction depends on the current input and the last output produced by the network itself. . . . .	11
2.3. <b>Vision inputs modalities.</b> Extracted from CASIA-B dataset [127].	11
2.4. <b>Early fusion.</b> Fusion applied to data before inserting it in the classification model. . . . .	16
2.5. <b>Late fusion.</b> Fusion at the predictor level, using the output of different models. . . . .	17
2.6. <b>Intermediate fusion.</b> Fusion in a hidden layer, where the model learns joint intermediate representations of each modality. . . . .	18
2.7. <b>Response distillation scheme.</b> The student model learns to mimic the master's final output layer. . . . .	21



2.8. <b>Feature distillation scheme.</b> The student model learns from intermediate feature activations coming from the master model. . .	21
2.9. <b>Situations captured in TUM-GAID</b> : Normal walk ( $N$ ), carrying a backpack ( $B$ ) and wearing coating shoes ( $S$ ). Also, it has an elapsed time case, recorded wearing different clothes ( $TN-TB-TS$ ). . . . .	23
2.10. <b>11 viewpoints captured in CASIA-B.</b> From $0^\circ$ to $180^\circ$ in steps of $18^\circ$ . Images obtained from [127]. . . . . . . . . . . . . . . . . . . . . . . . . . . .	24

# 1

# Introduction

---

Biometrics is the measurement of living things or biological processes, which can be used to identify humans, or general traits of humans, from their specific, personal traits. The field of biometrics [22] is undergoing a rapid growth, driven by the need for robust security and surveillance applications. However, its potential as a natural and effortless means of identification has also paved the way for a host of applications that automatically identify the user or specific user characteristics, providing personalized services. The main types of biometric systems today are based on the recognition of factors such as fingerprint [51], face [112], retina [88], voice [82] or signature [34].

*Gait* can be considered as an unequivocal biometric pattern of human locomotion, since each subject has its own biological characteristics, which makes use cases, such as identifying people from their gait, viable. Although gait has traditionally been a field studied from the point of view of medicine, (*e.g.* for the early diagnosis of diseases such as Parkinson's disease [96], Rett's syndrome [52], or cerebral palsy [90]), it has also been studied from the point of view of biometric security.

Among its advantages in the case of biometric security, we can highlight that the step pattern can be obtained non-invasively and without the user actively collaborating with the system, unlike other traditional methods such as iris analysis, facial recognition or the use of fingerprints. This allows its use in environments where other biometric patterns cannot be used, for example, when subjects must wear special clothing, such as NBC suits (nuclear, biological and chemical suits), or when the environment imposes limitations on biometric systems (camera position, face concealment, privacy legislation, etc.). We must also keep in mind that, due to natural dynamics, these patterns are difficult to duplicate.

The problem of gait recognition has traditionally been studied from the com-



puter vision point of view, something that, since it can be done remotely and is not dependent on observing specific parts of the body, does not require the collaboration of the subject and solves the problems mentioned above. In this field, different types of input data are used in the state-of-the-art (see Fig. 1.1): those based on appearance, such as RGB, gray or silhouettes; or those based on movements, such as Gait Energy Image (GEI) [43] or optical flow. Talking about the classification methodology, the dominance of Convolutional Neural Networks (CNNs) in the state-of-the-art is obvious since this type of model is currently the best type of algorithm for automatic image and video processing.

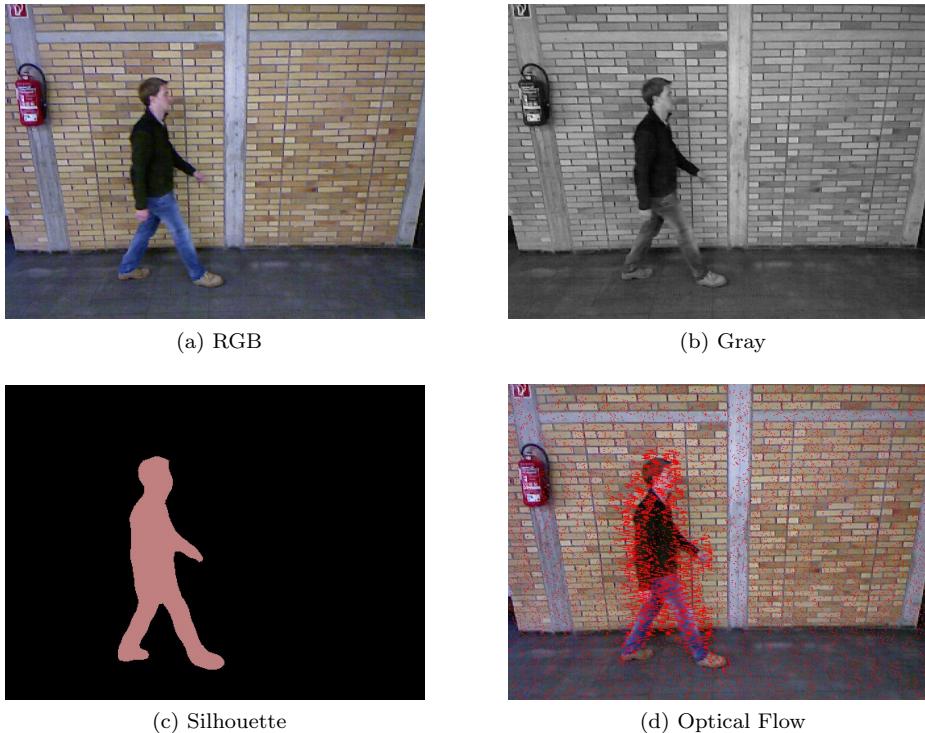
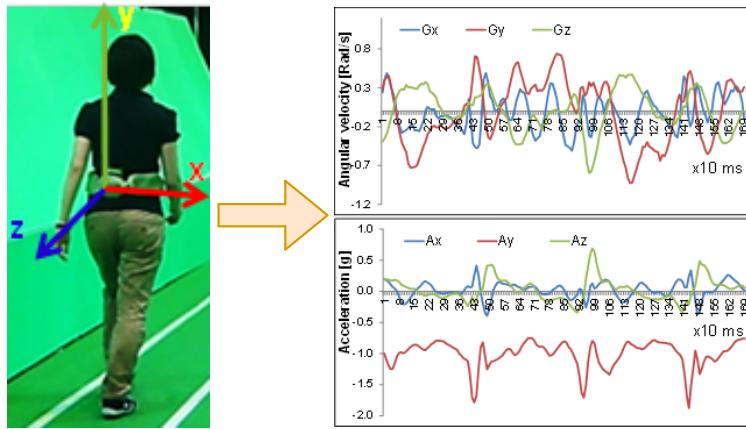


Figure 1.1: **Vision inputs in gait recognition.** Extracted from TUM-GAID dataset [47].

Although there is a lot of gait recognition work based on vision [116], a common problem in the typical datasets for this topic is the assumption of the presence of a single subject in the scene at any given time, limiting its application in a real environment.



**Figure 1.2: Inertial sensors in gait recognition.** The left part of the image shows a subject walking through a circuit with the sensor system in the waist. The right part of the image depicts the inertial information recorded during the walk. The top plot shows the measurements for the accelerometer and the bottom plot contains the measurements for the gyroscope. Images obtained from [85].

Also, gait analysis from inertial sensors has become an active and exploited topic thanks to the cheapening of MEMS (Micro Electro Mechanical Systems) sensors and their integration into smartphones [80] or smartwatches [53], making this step analysis system cheaper and more portable. In this case, information from accelerometers and gyroscopes (see Fig. 1.2) is usually studied and, unlike in vision-based approaches, the use of deep learning for classification does not abound in the state of the art, being more common the use of traditional machine learning systems.

## 1.1. Thesis Motivation

The motivation of this thesis is to investigate on how to solve the problems or turn points that the state-of-the-art on gait recognition is currently facing. For this purpose, the current state-of-the-art has been analyzed from the point of view of datasets, input data, classifiers, the robustness of the proposed systems, and possible optimizations that can be developed.

After this analysis, a series of objectives have been proposed to provide, during the development of this thesis, contributions that will allow to overcome the

current results, solve the problems detected and open, if it is possible, new lines of research.

## 1.2. Objectives and Phases

Next, we will define the objectives that have been set in the development of the thesis:

1. Classification of soft-biometrics features different from the identity and related to gait. For example, age and gender.
2. Implementation of a cross-dataset classifier for gait recognition without the need of fine-tuning models per dataset.
3. Implementation of a video sample generation system with multiple simultaneous subjects, using different computer vision and deep learning techniques. Also, study the simultaneous classification of multiple subjects.
4. Study of missing modalities, i.e. in the loss of information from one or multiple modalities (information sources) in the input of a deep learning-based system, and how to address it.
5. Use of knowledge distillation to optimize gait recognition models.

To achieve these objectives, the following phases have been carried out:

1. We reviewed the state-of-the-art to select the best approaches and the datasets used to evaluate our methods and compare them with other state-of-the-art approaches. We decided to use DFNAPAS [77], SisFall [105], UniMiB-SHAR [78], ASLH [91] and OU-ISIR [85] datasets in inertial approaches, and CASIA-B [127] and TUM-GAID [47] datasets in vision approaches. This step has been necessary to achieve all the proposed objectives.
2. To develop the soft-biometrics classifier [25], different inertial sensors have been studied as input, by type and position, and we have worked on the fusion of different types of sensors. Also, equivalent models in single-task and multi-task configurations have also been compared to observe the impact of simultaneous classification of multiple tasks. This step accomplished objective 1 and has provided information about multimodal and multi-task techniques for the following objectives.

3. Using the knowledge acquired about inertial sensors in step 2, we have implemented a cross-dataset classifier focused on fall detection [24]. Multiple datasets studied in step 1 have been used to test the performance of the cross-dataset approach without the need for additional training. We have also studied the use of a classifier external to the network,  $k$ -Nearest Neighbors classifier, that does not require a training process for the data used, but it has been necessary to adjust its hyperparameter  $k$ . Looking for its application in a realistic use case, high performance in the classification of falls in a group of elderly people has been proven by performing training only with young people, which demonstrates the feasibility of the cross-dataset approach. This has allowed us to fulfill objective 2.
4. We have worked on a framework capable of generating synthetic samples with multiple subjects from real samples with a single subject [28]. For this purpose, multiple segmentation techniques have been tested, selecting at the end a state-of-the-art CNN. An experimental methodology to test the visual quality of the generated images has been studied and proposed. For this purpose, the impact of multiple subjects per scene on the classification models has been explored, differentiating between the accuracy at the subject level and at the group level of subjects in the video. The impact of overlapping subjects is also observed, also differentiating between videos where subjects cross or run parallel. This step has been realized with the two vision datasets selected in step 1 and it completes objective 3, allowing investigators to work with new datasets with more realistic cases.
5. A system based on logic gates has been implemented that allows, at inference time, to disable one or multiple branches, allowing the model to continue working in cases of missing modalities [26, 75]. To combine the information of the different modalities, multiple merge functions have been tested and compared for the different modalities used. Analogous single-task models with late fusion have been tested with our proposed multi-task models. Also, we have experimented with equal models applying or not our approach, to study the effect of our approach on accuracy. This step has completed objective 4, with an architecture able to deal with missing modalities without significant performance degradation.
6. It is proposed to use knowledge distillation to teach models to mimic a computationally expensive input modality, optical flow, by means of a less expensive input modality, gray scale images [27]. Two sets of master-student models are implemented, one of our own and the other based on a state-of-the-art model, in order to study the generalization of our proposal. Ablation

studies have been carried out to verify that the proposed techniques have a real impact on the process of knowledge distillation. This step accomplished objective 5, obtaining a model with a smaller size that in inference requires less computational cost.

### 1.3. Contributions

The contributions of this thesis according to the proposed objectives are:

- For objective 1, soft-biometrics classification, the contributions are:
  1. A new end-to-end approach, based on a CNN architecture, for gait-based recognition and authentication problems that uses raw inertial data as input.
  2. A fusion scheme has also been proposed which takes advantage of data obtained from several inertial sensors, accelerometer and gyroscope, to generate more robust models. The impact of this on classification has also been studied.
  3. A multi-task classification model that takes advantage of the multiple labels contained in the dataset used.
  4. New state-of-the-art results for identity, gender, age and authentication on OU-ISIR [85] dataset.
- For objective 2, cross-dataset classifier, the contributions are:
  1. A new cross-dataset classifier based on a deep architecture and a  $k$ -NN classifier.
  2. A classifier able to detect falls and identify subjects at the same time using a single model, with a multi-task approach.
  3. Improvement of the results obtained by previous proposals of the state of the art in four different public datasets.
- For objective 3, the video generation system with multiple simultaneous subjects, the contributions are:
  1. The first framework to generate augmented gait datasets with multiple persons in the scene (MuPeG) using existing datasets. This framework allows researchers to build a new type of datasets that did not exist before in the state-of-the-art.

2. The new framework opens new challenges for researchers with realistic gait analysis problems.
  3. It has been proposed an experimental methodology that defines the minimum number and type of experiments that must be performed in this kind of datasets.
  4. A first baseline for the proposed experimental methodology is established.
- For objective 4, to mitigate the effect of missing modalities, the contributions are:
    1. We developed UGaitNet, a single network that handles and combines various types of input modalities for gait recognition: pixel gray value, optical flow, depth maps and/or silhouettes.
    2. This network is shown to be fault-tolerant with respect to the absence of one or multiple input modalities.
    3. It obtains state-of-the-art gait descriptors when evaluated on TUM-GAID.
  - For objective 5, knowledge distillation for computation optimization, the contributions are:
    1. It has been introduced a new approach named GaitCopy, based on knowledge distillation, that is able to mimic the behavior of optical-flow-based networks for gait recognition but using gray-level pixel inputs.
    2. It is experimentally shown on CASIA-B and TUM-GAID that, compared to the master networks trained on optical flow data, a similar accuracy can be obtained by the student networks trained on gray inputs.
    3. Networks designed for gray are significantly smaller in the number of parameters than their optical flow analogs.
    4. An optimization in inference time is also obtained because the optical flow does not have to be computed for the models obtained, which is computationally expensive.

## 1.4. Thesis Structure

The rest of this thesis is structured in the following way:

- *Chapter 2* introduces the basics of gait recognition and different elements related to it that have been addressed during the development of this thesis. Also, we present the state-of-the-art datasets, input modalities and classifiers for gait recognition.
- *Chapter 3* reviews the state-of-the-art of gait recognition concepts that we consider most important for the development and understanding of this thesis.
- *Chapter 4* contains a copy of the published works that support this thesis with a brief summary of each one of them.
- *Chapter 5* summarizes the conclusions of this thesis and the contributions it makes to state-of-the-art, and proposes lines of future work.

# 2 Background

---

## 2.1. Artificial Neural Networks

*Artificial Neural Network* [1], or *ANN*, is a group of multiple computer units, or nodes called *neurons*, imitating biological neurons of the brain.

An Artificial Neural Network is capable of learning any nonlinear function. Hence, these networks are popularly known as Universal Function Approximators. One of the main reasons behind the universal approximation is the *activation function*. Activation functions introduce nonlinear properties to the network. This helps the network to learn any complex relationship between inputs and outputs.

In an ANN, neurons are organized in groups, called *layers*. When multiple layers are used together, the term *deep learning* is used to refer to such networks. Normally, a larger number of layers implies a larger number of *trainable parameters*. The number of trainable parameters is an important factor in the design of a neural network: a very small number of parameters means that the model will not be able to generalize enough, producing an effect called *overfitting*; on the other hand, a large number of parameters means that larger amounts of data will be needed to obtain that generalization.

The architecture of an ANN is composed by layers and its *loss function*, a function that calculates the error committed by the network output and the real value that should have been obtained. This will be used during the training process of the network to optimize the trainable parameters, seeking to minimize the value calculated by the loss function.

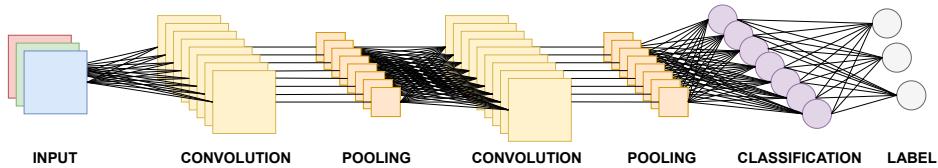


Figure 2.1: **Typical CNN architecture.** Common architecture of a CNN, consisting of multiple convolutions and pooling layers concatenated together and ending with a classification layer.

Although exist many kinds of deep models, this thesis focuses on *Convolutional Neural Networks* and *Recurrent Neuronal Networks* [41], which will be used in the developed approaches.

Convolutional Neural Networks [63] (*CNNs*) are predominant in the state-of-the-art methods to work with images to obtain features. These networks use *convolutional layers*, a spatial structure to simulate traditional kernels applied in convolutional operations. Then, the first layers of the network have a tendency to produce low-level features that focus on corners, edges, patterns, etc., and the later layers will focus on higher-level details with more meaning, such as eyes or faces. Convolutional layers are accompanied by *pooling layers*, whose purpose is to reduce the dimensionality of the feature maps, and a final classification layer (Fig. 2.1).

Recurrent neural network [31] (*RNNs*) is a type of artificial neural network which uses sequential data or time-series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation [121], natural language processing [42], speech recognition [42], and image captioning [126]. They are distinguished by their “memory” as they take information from prior inputs to influence the current input and output (Fig. 2.2). While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depends on the prior elements within the sequence. While future events would also help to determine the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions. This allows it to exhibit temporal dynamic behavior.

To train an ANN, the *hyperparameters* of the model must be fit according to the specific characteristics of the problem. The training process can be split into three steps which are performed iteratively until the model converges. The first training step is the *forward pass* where the input data is passed through

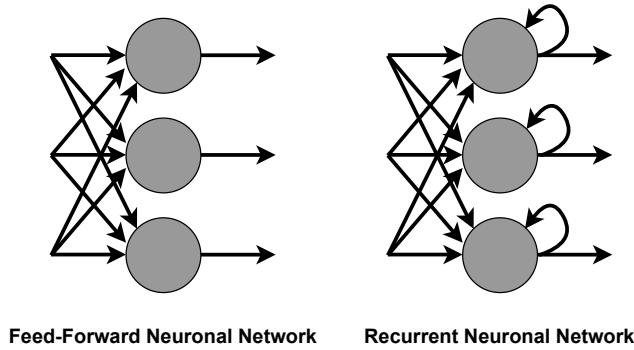


Figure 2.2: **Recurrence in RNN architecture.** The left part of the image shows a classical feed-forward network. The right side shows a recurrent network, whose prediction depends on the current input and the last output produced by the network itself.

the ANN to obtain the activations and loss error. Then, the second training step, called *backpropagation*, computes the partial derivatives of each parameter according to the error obtained from the loss function. These partial derivatives encode the direction and value of the parameter update needed to minimize the loss error. Finally, the *parameter update* minimizes the loss error according to the derivatives obtained in the previous step.

## 2.2. Input Modalities

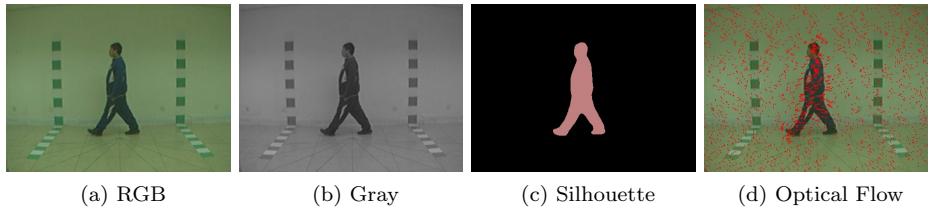


Figure 2.3: **Vision inputs modalities.** Extracted from CASIA-B dataset [127].

In this section, we will discuss the most commonly used data types in gait classification models. We have divided the section between visual data types (related to the representation of the subject in images and videos) and inertial



data types (related to the inertial motion of the subject). We will explain their strengths and weaknesses and give examples of their use in the current state-of-the-art approaches.

### 2.2.1. Visual Modalities

We denote as *visual modalities* those that are related to visual data captures, originally coming from images and videos taken on the subject that performs the gait. Visual modalities are the most common in the state of the art, especially those related to convolutional networks, due to their high performance in the field of computer vision. Fig. 2.3 shows the visual modalities explained in this section.

#### 2.2.1.1 Raw pixels: RGB & Gray

The basic unit of an image is a *pixel*, that is, the smallest element of an image that contains a value for a small square region. Thus, if an image has a shape, for example, of 64x64, it means that it will contain 4096 pixels arranged in rows and columns building a square of 64x64 pixels. Furthermore, these values per pixel are typically in the range of values between 0 and 255.

In an *RGB image*, there are three color channels, i.e. three different values per pixel. These contain a value related to red, a value related to green, and a value related to blue, which makes it possible to obtain any color by combining these three primary colors. This is how color images are encoded in computers [40].

In contrast, *grayscale images* have a single channel, i.e. a single value per pixel. This represents the light, or illumination intensity, in grayscale, where 0 is black and 255 is white. In this way, a visual representation is obtained that is abstracted from color but maintains the shapes and edges of the objects represented in the image, as well as the difference in textures given by the different levels of gray.

These two types of images are common as input in the field of computer vision. RGB images are important if the color and texture of the objects are especially significant for the problem to be solved. This could be, for example, if you want to know the state of a traffic light or to visually check the ripeness of fruit. Grayscale images, on the other hand, are used to force machine learning models to abstract from color in problems where this information can be counterproductive. For example, in the case of gait recognition, color can provide information that

generalizes during training so that when the subject changes clothes, the model is no longer able to correctly identify the subject.

### 2.2.1.2 Silhouettes

A *silhouette* is a representation of an object, animal, or person that contains information about its outermost shape. It is represented by a solid figure of a single color, usually black, with its edges matching the outline of the subject. The interior of a silhouette has no additional features. While an outline represents the edge of an object linearly, the silhouette appears as a solid shape. The silhouette is usually presented on a light background, usually white, or none at all.

In image-based classification algorithms, silhouettes are used because, as opposed to RGB and gray scale images, they hide both the color and the texture of the classification target, allowing the classifier to focus only on the outer contour information and its position.

In the current state of the art, many gait works are based on silhouettes [17, 32, 67]. Then, we can assure that it is the most common visual modality in this field of study.

### 2.2.1.3 Optical Flow

*Optical Flow (OF)* [48] is defined as the motion pattern in a scene caused by the relative motion between an observer and the scene between two instants of time.

The OF is divided into two components or channels, one representing the  $y$ -axis and the other representing the  $x$ -axis, where most of the gait motion flow is concentrated.

Although the optical flow field is an approximate projection of the true motion of the scene, it provides valuable information about the spatial arrangement of the viewed objects and the change rate of the arrangement. OF has shown excellent results in the characterization of gait [10] since it is a motion-centric representation that obviates appearance and focuses on describing a subject by a set of local and subtly varying motions.

### 2.2.2. Inertial Modalities

Inertial modalities are those that contain only inertial information of gait. These can be represented as sequences of data, or signals, on a time axis. These are taken by an *Inertial Measurement Unit (IMU)*, an electronic device that measures and reports a body's specific force, angular rate, and sometimes the orientation of the body. An IMU normally consists of the sensors shown in the following subsections.

Although it is more common to use inertial information for gait in classical Machine Learning algorithms [60, 57], we can also find some works done on CNNs, taking advantage of the fact that the features are automatically obtained by the network during the training process using raw signals as input. There are two main approaches. On the one hand, some papers use raw information coming from the IMU sensors [37]. On the other hand, some approaches transform the inertial information into an image-based representation to feed a CNN, taking advantage of its capabilities to work with images. Thus, in [132, 133], the authors transform the inertial signals in spaced time series, called Gait Dynamics Image (GDI), which are used as CNN input samples.

#### 2.2.2.1 Accelerometer

*Accelerometers* are electromechanical devices that detect static or dynamic acceleration forces. Static forces can include gravity, while dynamic forces can include vibrations or motion.

Generally, accelerometers contain internal capacitive plates, some fixed and some attached to tiny springs that move internally when acceleration forces act on the sensor. As these plates move relative to each other, the capacitance between them changes, thus allowing the acceleration to be determined. Other accelerometers may be centered around piezoelectric materials: these tiny crystal structures output electrical charge when placed under mechanical stress, e.g. accelerations.

Most accelerometers will have a selectable range of forces that they can measure. These ranges can vary from  $\pm 1$  g up to  $\pm 250$  g. Typically, the smaller the range, the more sensitive the accelerometer readings will be. For example, to measure small vibrations on a table, using a small range accelerometer will provide more detailed data than using a range of  $\pm 250$  g, which would be more appropriate for a rocket or something similar.

### 2.2.2.2 Gyroscope

*Gyroscopes* are small sensors that measure angular velocity, which is simply a measure of rotational speed. These are measured in degrees per second ( $^{\circ}/\text{s}$ ) or revolutions per second (RPS).

The gyroscope sensor inside the housing is between 1 and 100 micrometers. When the gyroscope is rotated, a small resonant mass moves as the angular velocity changes. This motion is converted into very low current electrical signals that can be amplified and read out.

These types of sensors can be used to determine orientation and are found in most autonomous navigation systems. For example, if you want to balance a robot you can use a gyroscope to measure the rotation from the balanced position and send corrections to a motor.

There are many specifications to consider when choosing a type of gyroscope to use. The most important are again the range, which is the maximum angular velocity that the gyroscope can read, and the sensitivity, which is measured in mV per degree per second ( $\text{mV}/^{\circ}/\text{s}$ ).

## 2.3. Data Fusion

When several sources of information are available, a method to fuse those sources [95] of data can be used to improve the performance of the global approach. On the one hand, we can combine those data before inserting them into the classification model. This approach is usually known as *early fusion*. A typical example of early fusion is the concatenation of data vectors. On the other hand, we can train independent classifiers for each source of information, and then, define a strategy to fuse the classification or confidence scores. This is known as *late fusion*. Finally, we can define as *intermediate fusion* those occurring in the intermediate stages of the classification network, between layers.

We applied some of these fusion schemes in the works that compose this thesis (Ch. 4), helping the training process and obtaining better results.

### 2.3.1. Early Fusion

Early fusion, or data level fusion, is a traditional way of fusing multiple data before their use in the classification model (Fig. 2.4). For example, [56] proposes

two possible approaches for early fusion techniques: combining data by removing the correlation between two sensors, or fusing data at their lower-dimensional common space.

Early fusion is applicable to raw data or pre-processed data obtained from sensors. When the data sources have different sampling rates between the modalities, data features should be extracted from the data before fusion. Synchronization of data sources is also challenging when one data source is discrete and the others are continuous.

There are two disadvantages to using early fusion. One of the main disadvantages of this method is that a large amount of data will be discarded from the modalities to make a common ground before fusion. On the other hand, this method is synchronizing the timestamp of the different modalities, which forces to collect the data or signals at a common sampling rate.

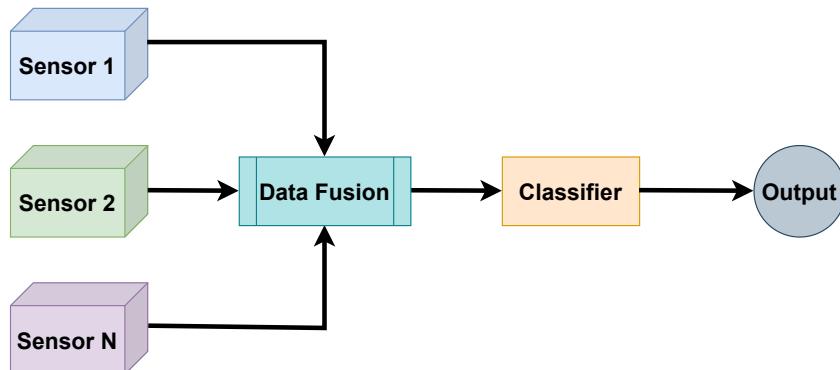


Figure 2.4: **Early fusion.** Fusion applied to data before inserting it in the classification model.

### 2.3.2. Late Fusion

Late fusion, or decision level fusion, uses data sources independently followed by fusion at a decision-making stage, using the predictors obtained by the different classification models (Fig. 2.5).

Late data fusion is inspired by the popularity of ensemble classifiers [7]. This technique is much simpler than the early fusion method. However, in [95] authors argue that there is no conclusive evidence that late fusion performs better than early fusion. Yet, others researchers use late or decision level fusion to analyze

multimodal data problems [100].

Different algorithms exist to determine the optimal way of deciding how to finally combine each of the independently trained models. Bayes rules, max-fusion and average-fusion are some of the commonly late fusion algorithms. When the input data streams are significantly varied in terms of dimension and sampling rate, using late fusion is a simpler and more flexible approach.

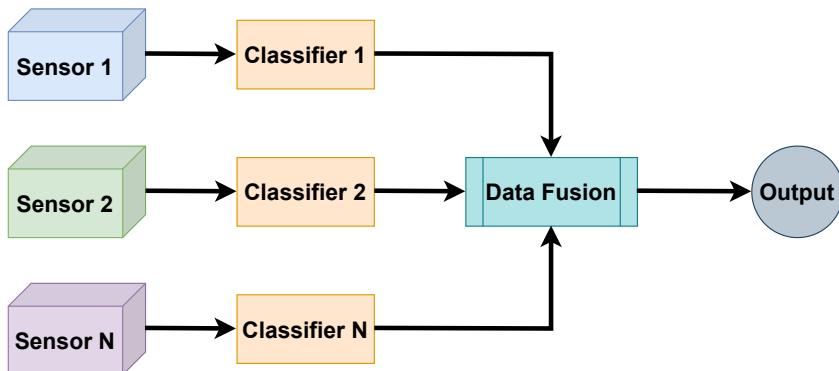


Figure 2.5: **Late fusion.** Fusion at the predictor level, using the output of different models.

### 2.3.3. Intermediate Fusion

The architecture of intermediate fusion is built based on deep neural networks. This method is the most flexible, allowing for data fusion at different stages of model training.

Intermediate fusion in a deep learning multimodal context is a fusion of different modalities representations into a single hidden layer so that the model learns a joint representation of each of the modalities. The layer where the fusion of different modality features has taken place is called a fusion layer or a shared representation layer.

Different modalities can be fused simultaneously into a single shared representation layer or this can be performed gradually using one or multiple modalities at a time (Fig. 2.6). Although it is possible to fuse multiple modality features or weights in a single layer, it may lead to model overfitting, or the network may fail to learn the relationship between each modality. Also, as opposed to early level fusion and late fusion, intermediate fusion offers flexibility to fuse features

at different depths.

[55] uses a neural network where training video stream features are gradually fused across multiple fusion layers. This approach performs better in a large-scale video stream classification problem. [84] shows a gradual fusion method that fused highly correlated input modalities first and less correlated after.

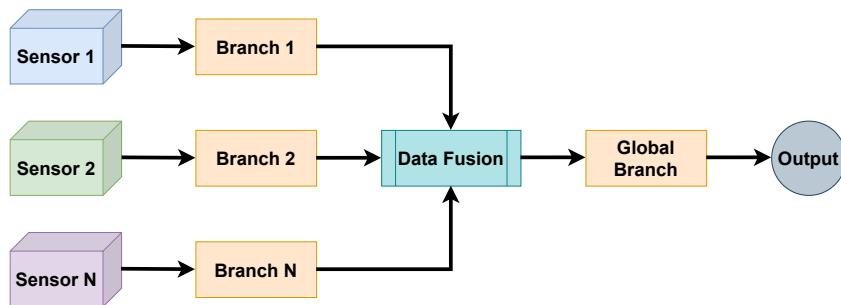


Figure 2.6: **Intermediate fusion.** Fusion in a hidden layer, where the model learns joint intermediate representations of each modality.

## 2.4. Classifiers

A classifier is a type of algorithm used to assign a class label to data input. An example is an image recognition classifier to label an image (e.g., ‘car’, ‘truck’, or ‘person’). Classifier algorithms are trained using labeled data, so it belongs to the category of supervised learning. In image recognition, for example, the classifier receives training data that are labeled. After sufficient training, the classifier can receive unlabeled images as inputs and will output classification labels for each image.

We are going to explain the two types of classifiers that have been used in the development of this thesis.

### 2.4.1. *k*-Nearest Neighbors

In a *k*-Nearest Neighbors classifier [36], a sample is classified by a majority voting strategy of its neighbors. Thus, the sample is assigned to the most voted class among its *k*-Nearest Neighbors. In this case, *k* is the number of neighbors taken into account during the majority voting. To find the neighbors of a new

sample sequence, a similarity or distance metric is computed between such a new sequence and all previously labeled ones. Then, the  $k$  closest sequences to the new one are selected as neighbors.

Therefore, the similarity metric plays an important role during the neighbor selection, as it will strongly influence the closeness relationship. Thus, this classifier does not require a proper training process since there are no parameters to learn. The method only needs to select in advance the  $k$  value and a training set of labeled sequences.

During test time, the similarity metric between the new sequences and the training set is straightforwardly computed to find the neighbors. Then, the most frequent label among the  $k$  nearest neighbors is assigned to the new sequence.

#### 2.4.2. Fully-Connected Layer

Fully Connected layers in neural networks are those layers where all the inputs from one layer are connected to every activation unit of the next layer. This layer is the principal component of the traditional Neural Networks and it is also used in convolutional networks, where the last few layers are fully connected layers that compiles the data extracted by previous layers to form the final output.

This layer concentrates most of the parameters of the network and it is usually used to store the high-level knowledge used by the classifier. Due to its large number of parameters, the amount of these layers included in a network is very limited to prevent overfitting.

This layer can be used as a classifier of the model, with each of its output neurons representing a different class. Including the classifier in the neural network itself makes the model end-to-end trainable.

## 2.5. Multi-Task Learning

Multi-task learning [9] is a training paradigm in which machine learning models are trained with data from multiple tasks simultaneously, learning the common ideas between a collection of related tasks. These shared representations increase data training efficiency and can accelerate the learning process for related tasks, as well as help to reduce the data size requirements and computational cost. However, achieving such effects is difficult and is an active area of research today.

Learning concepts for multiple tasks implies difficulties that are not present

in single-task learning. For example, different tasks could have contradictory learning requirements. In this case, increasing the performance of one model on one task will prejudice performance on a task with different requirements, a phenomenon called negative transfer or destructive interference. Many architectures are designed with specific characteristics to reduce negative transfer, such as task-specific feature spaces and attention mechanisms, but the division of information between tasks is a delicate process.

In multi-task neural network architectures, there are many different factors to consider when creating a shared architecture, such as the portion of the parameters to be shared between tasks. Many of the proposed architectures for the multi-task attempt to balance the degree of information sharing between tasks: too much sharing will result in negative transfer and may lead to worse performance of joint multi-task models than individual models for each task, while too little sharing will cause the model to fail to effectively exploit information between tasks.

## 2.6. Knowledge Distillation

Knowledge distillation is the process of transfer knowledge from a large model to a smaller one. This process must somehow teach the student model without loss of validity. If both models are trained with the same data, the small model may not have sufficient capacity to learn the generalization. However, part of the information about a concise knowledge representation is encoded in the output: when a model correctly predicts a class, it assigns a large value to the output variable corresponding to that class and smaller values to the other output variables. The distribution of values in the layer outputs provides information about how the large model represents knowledge. Thus, one can train only the large model on the data, exploiting its best ability to learn concise knowledge representations, and then distilling that knowledge into the smaller model, which would not be able to learn it on its own, by training it to learn the smooth output of the large model.

We can divide knowledge distillation into two different types: response-based and feature-based.

Response-based distillation focuses on the final output layer of the master model. The student model will learn to mimic the master predictions, according to the hypothesis. This can be done using a loss function known as the distillation loss, which captures the difference between the logits of the student and teacher

models, as shown in Fig. 2.7. The student model will become more accurate in making predictions similar to the teacher as this loss is reduced over time.

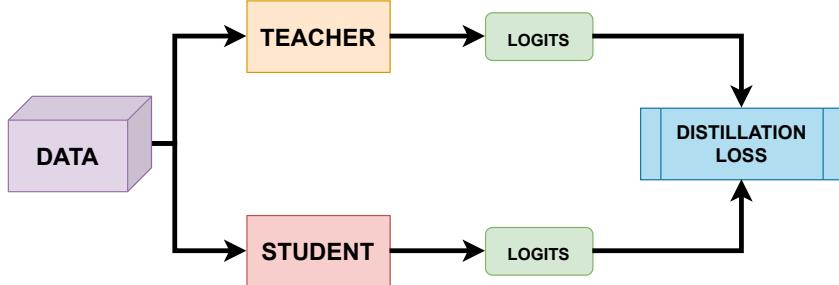


Figure 2.7: **Response distillation scheme.** The student model learns to mimic the master's final output layer.

Moreover, feature-based distillation trains the student model to learn the same intermediate feature activations as the master model (Fig. 2.8). Deep neural networks learn multiple levels of feature representation as features progress through the network. A master model trained also captures the knowledge of the data in its intermediate layers, which is especially important for deep neural networks. The intermediate layers learn to discriminate between specific features, which can then be used to train a learner model.

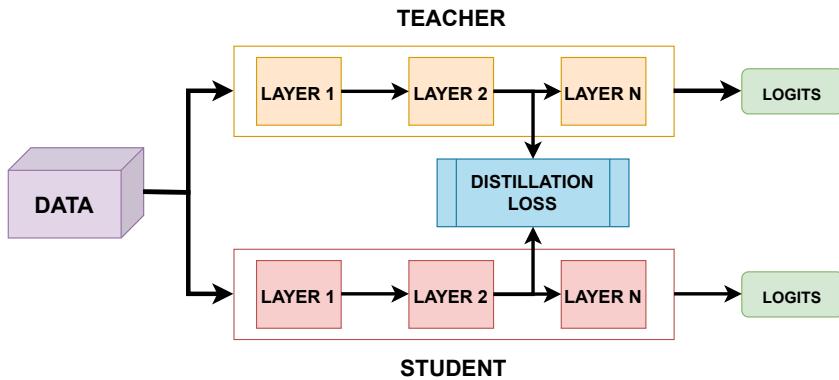


Figure 2.8: **Feature distillation scheme.** The student model learns from intermediate feature activations coming from the master model.

## 2.7. Datasets

In this section, we will discuss the datasets used in the papers that comprise this thesis and their characteristics. We will differentiate between *inertial datasets*, which contain inertial information about the gait taken by inertial sensors, mainly accelerometers and gyroscopes; and *visual datasets*, which contain videos and are thought for computer vision-based models. This section also explains what synthetic datasets are and their utility in the current state-of-the-art.

### 2.7.1. Inertial Datasets

The inertial datasets may contain one or multiple IMU sensors, which may be synchronized in the acquisition or, on the contrary, not synchronized even if they share a sampling rate. The position of the sensors is varied, which makes possible to study the importance of each area of the body with respect to the motion. They are also, in some cases, inside a smartphone, which proposes a realistic scenario with respect to their deployment in a real use case.

- DFNAPAS [77]: Dataset created to identify falls and *activities of daily living (ADL)*. It contains 7816 sequences, collected from sensors placed in the pockets of 10 different subjects, aged between 20 and 42, with a sampling rate of 50 Hz. The recorded sequences have a fixed length of 301 samples, so their duration is approximately 6 seconds. It includes 503 different falls. Activities included are forward falls, backward falls, left and right-lateral falls, syncope, sitting on empty chair, falls using compensation strategies to prevent the impact and falls with contact with an obstacle before hitting the ground.
- SisFall [105]: It is composed of 4505 sequences (2707 ADL sequences and 1798 fall sequences) recorded from three sensors, although during the experimentation presented in the original paper only the waist accelerometer is used. The sequences are taken from 38 subjects, where 15 of them are between 60 and 75 years old, with a sampling rate of 200 Hz and variable length between 1999 and 36000 samples. Its labels include 19 ADLs and 15 different types of falls. It also includes individual information for each subject on age, height, weight and gender.
- UniMiB-SHAR [78]: It consists of 7013 sequences, of which 1699 are falls. Each sequence contains 51 samples recorded from 30 different volunteers with a sampling rate of 50 Hz. The sensor is placed in the subject's pocket.

This dataset includes the traces captured from the movements of subjects between 18 and 60 years old.

- ASLH [91]: It contains 3302 sequences with 1826 falls. 17 subjects, with ages between 20 and 27 years, used a body sensor network encompassing six IMUs, embedding an accelerometer, a gyroscope and a magnetometer. The sequences have a variable length (between 210 and 945 samples) with a sampling rate of 25 Hz. IMUs were located on six different positions of the body: head, chest, waist, wrist, thigh and ankle.
- OU-ISIR Dataset [85]: It is the largest gait dataset based on inertial sensors. The information is collected using one smartphone and three IMU sensors located in the waist of the subjects. The dataset is split into two subsets, the first one (part A), which is composed of 744 subjects (389 males and 355 females) with ages between 2 and 78 years, is recorded only with the central IMU sensor. In this subset, two sequences of data per person have been recorded at a rate of 100Hz. Following the methodology used in similar works ([133, 132, 119, 85, 38, 29, 97, 81]), the first sequence is used for training and the second one for testing. Three labels, namely *identity*, *gender* and *age*, are provided for each subject.

### 2.7.2. Visual Datasets

Vision-based datasets are constructed from variable-length videos containing a single subject walking in a controlled environment, with no obstacles or other subjects in the scene. Typically, these videos are natively recorded in RGB, although they can be recorded additionally in other types of formats. The following datasets have been used in the vision-based work of this thesis:

- TUM-GAID [47]: It contains 305 subjects performing two walking trajectories indoors, captured by a Microsoft Kinect (resolution of  $640 \times 480$  pixels and 30 frames per second). Four situations are captured (Fig 2.9): normal walk (*N*), carrying a backpack (*B*), wearing coating shoes (*S*) and, there is an elapsed time case where 32 subjects were recorded wearing different clothes (*TN-TB-TS*). Traditionally, the train and test splits from [47] are used: 150 subjects for training and 155 subjects for testing. In the training set, all walking conditions are used. In test, for the 155 test subjects, a gallery set is created, containing only normal walk videos. This gallery set is used to extract predictors to classify these unseen subjects. On the other



Figure 2.9: **Situations captured in TUM-GAID** : Normal walk (*N*), carrying a backpack (*B*) and wearing coating shoes (*S*). Also, it has an elapsed time case, recorded wearing different clothes (*TN-TB-TS*).

hand, the probe set used to obtain the results contains all walking conditions in order to measure the robustness of the gait recognition approaches. In addition to RGB videos, the dataset contains depth videos and audio.

- CASIA-B [127] It consists of 124 subjects that walk indoors, captured from 11 viewpoints (Fig. 2.10), from  $0^\circ$  to  $180^\circ$  in steps of  $18^\circ$ , with a resolution of  $320 \times 240$  pixels. Three situations are considered: normal walk (*nm*), wearing a coat (*cl*), and carrying a bag (*bg*). Following [?], the first 74 subjects is used at training and validation, and the last 50 at test. Similar to TUM-GAID, in the test stage two sets are created, i.e., galley and probe, with the same characteristics and purpose that have been explained before.

### 2.7.3. Synthetic Datasets

Synthetic data is the one created in an artificial manner, so researchers can emphasize the necessary features for a given use case. Compared to real-world data, synthetic data generation is faster, more flexible, and more scalable. By adjusting parameters, it can also be an effective way to model and generate data that does not exist in the real world. Moreover, in some use cases, it is an advantage over privacy law restrictions that may affect real data.

As the unbalanced distribution of data classes typically generates a bias towards the majority class due to insufficient training samples from the minority class, synthetic data are mainly employed to correct this unbalance. Thus, for instance, [110] performs a study of data augmentation techniques that are used to solve these scenarios.

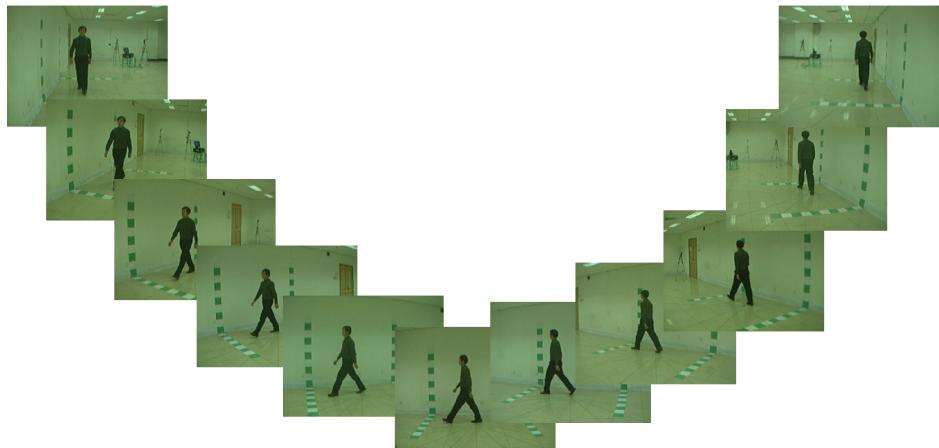


Figure 2.10: **11 viewpoints captured in CASIA-B**. From  $0^\circ$  to  $180^\circ$  in steps of  $18^\circ$ . Images obtained from [127].

For machine learning, the real or synthetic nature of data has a low impact if the synthetic samples present similar characteristics to the real ones. For this reason, the balancing must be done by taking care that the synthetic data are of high quality, in other words, that the synthetic data have a reasonable realism. This depends on the similarity of the distributions or the performance for specific tasks, such as inference or prediction. In this aspect, [5] propose a framework to evaluate the quality of synthetic datasets.

UNIVERSIDAD  
DE MÁLAGA



# 3 Related work

---

This chapter reviews the related work that we consider necessary for the understanding of this thesis. To ease the comprehension of this chapter, it has been divided into sections according to the approach described. Thus, Sections 3.1 and 3.2 reviews inertial and visual approaches respectively. Sections 3.3 and 3.4 focuses on multi-task and multimodal methods. Finally, Section 3.5 describes the knowledge distillation approaches.

## 3.1. Gait Recognition Inertial Approaches

The study of gait using information from inertial sensors attached to the subject is widely applied to many different fields, like human activity recognition [37, 61], fall detection [114], Parkinson’s diagnosis [98] or monitoring of patients with Parkinson’s disease [45, 16]. Another application that is gaining importance in the state-of-the-art is people identification using their way of walking, *i.e.* gait recognition. In general, most of the gait recognition approaches are based on computer vision [3, 104, 73, 14, 74], but there are also previous works that are based on inertial sensors [72, 119], attaching the sensors to the subject in a specific position and orientation. Thus, the data collected by the sensor has the same coordinate system. Note that the sensor position is an important aspect to be taken into account as motion dynamics can vary depending on the sensor location. Some typical positions are hips [115], legs [108], chest, ankles, lower back and wrists, or combinations of the previous positions [23]. More realistic locations, such as in a bag [113] or in a pocket [30], have also been investigated. Other approaches, like [99], focus on the authentication problem using sensors



integrated on smartphones, so the position is not controlled. In [20], the authors review a wide range of approaches applied to the gait recognition problem using different kinds of inputs.

In the field of gait recognition with inertial sensors, many different approaches have appeared during the last years. Dynamic Time Warping (DTW) has been used as a distance measure in [8, 38, 97]. In these works, gait sequences are initially divided into gait cycles and compared, using DTW, with some previously selected reference cycles for each class. A similar approach is presented in [87], where, instead of using DTW as a metric to compare the cycles, a Hidden Markov Models (HMM) is applied. Similarly, in [29], a cyclic rotation metric (CRM) is employed instead of DTW. Classification Trees are used by Watanabe *et al.* [118] as classifiers for gait recognition. They are employed to process inertial data extracted from the mobile phone of the subjects while they are walking. In [19], the authors compare different gait signature metrics to represent the gait information. Finally, these signatures are classified with a  $k$ -Nearest Neighbors algorithm. Other techniques apply radial basis function (RBF) networks to locally approximate the accelerations and angular velocities [119] to identify subjects.

Another important difference between approaches lies in the type of input data, which can be organized and pre-processed in many different ways. Kwapisz *et al.* [60] employs a combination of time-domain features such as average, time between peaks or binned distributions. In [57], the authors proposed a new methodology based on utilizing the fundamental spectral relationship between the movement of different body parts during gait. More complex features are also used, like Time Frequency Representation [23], which is a way to describe a signal simultaneously in a frequency and time space. Higher-Order Statistics [102] are extensions of second-order measures to higher orders, useful to non-Gaussian's real-life signals as gait signals.

With the advent of deep learning and CNNs, instead of developing features manually, the features are automatically obtained by the network during the training process using raw signals as input. In the case of inertial information, there are two main approaches. On the one hand, the models are fed with raw information coming from the inertial sensors [37]. On the other hand, there are approaches that transform the inertial information into an image-based representation to feed a CNN, taking advantage of its capabilities to work with images. Thus, in [132, 133], the authors transform the inertial signals in spaced time series, called Gait Dynamics Image (GDI), which are used as CNN input samples.

Traditionally, the cycle-stationary character of the march has been used to

split the data into small subsequences. This helps to perform a faster and less expensive processing as the amount of input data is smaller than when a full sequence is employed. Thus, the full gait sequence is further subdivided using a cycle-based segmentation [107, 85, 86, 35] or a window-based segmentation [4, 60]. The former explicitly study this previously mentioned cyclic character and creates a precise but complex segmentation. The latter, which is the simplest option, obtains the resulting segmentation following the assumption that one window should contain, at least, one complete gait cycle. Normally, the length of this window is between 1.4 [89] and 10 seconds [60].

In reference to the identification of other tasks using inertial sensors, on one side, it can be also applied to soft biometrics tasks such as *gender* recognition or *age* estimation. Usually, these tasks are independent of the main task [47, 11], but it has been demonstrated, like in applications for face recognition, that they can help to improve the results of the main task [109]. On the other hand, many research efforts have been devoted to the development of efficient and cost-effective Fall Detection Systems (FDS), capable of discriminating falls from ADL (Activities of Daily Living) to automatically issue an alarm to a remote control point as soon as it is suspected that the monitored patient (or user) has fallen unconscious.

About fall detection, authors in [69] and [68] have thoroughly compared the performance of basic thresholding techniques with a wide set of supervised and novelty-based learning solutions (including a CNN). Results show that the best detection ratio is achieved by combining a CNN and a one-class SVM classifier. Similarly, [65] applied a CNN to discriminate falls from ADLs in three public repositories with acceleration measurements. Authors showed that the system efficiency improves when data augmentation is considered to train the network. Taking advantage of the CNN capabilities to learn from images, [125, 44] transform inertial signals into images in order to train an own model or to fine-tune a pretrained AlexNet on ImageNet, respectively. In [18], the authors use signals transformed by dimensionality reduction techniques to train a CNN model. Also, state-of-the-art works use LSTM networks [120, 71], directly feeding raw values obtained from the accelerometer and the gyroscope to check which component of the acceleration or the angular velocity is more determinant to identify the anomalous movements caused by a fall. Differently from these approaches, in [94] the authors use a set of preprocessed windows as input to a LSTM model. Processing even more input data, in [66, 2] the authors build sets of windows that contain manually designed descriptors. Analogously, in other works, such as those by Mauldin *et al.* [76], Musci *et al.* [83] or Theodoridis *et al.* [111], RNNs have been utilized as the decision core of a FDS.

### 3.2. Gait Recognition Visual Approaches

From the point of view of the visual input modalities, when using CNNs for object detection and categorization, the most popular modality is raw pixels [58, 100]. Even so, in gait recognition, color is not as informative as it is for object recognition. For this reason, using only gray [10] intensity will eventually help CNN to focus just on the gait-relevant information.

Alternatively, most of gait recognition works use a stack of binary silhouettes as input data. In this regard, Gait Energy Image (GEI) [43] is the most popular silhouette-based gait descriptor. This descriptor is the result of computing a temporal averaging of the binary silhouette of the subject.

Some authors proposed other approaches derived from GEI. For example, the computation of Histograms of Oriented Gradient (HOG) descriptors from GEI and Chrono-Gait Image (CGI) was proposed in [70]. The main objective of its authors is to preserve temporal information and generate more abundant local shape features. In [33] we can see another example. There, authors try to use the Linear Discriminant Analysis (LDA) to reduce the dimensions of the GEI features. Portillo et al. [92] reduce these dimensions too using the Direct Linear Discriminant Analysis (DLDA). Also, [50] proposes the use of the Enhanced Gabor representation of the GEI, a regularized local tensor discriminant analysis method. [106] divide the silhouette between upper and lower silhouettes, and use its centroids to calculate gait features.

Motion Silhouette Image (MSI) [62] is another silhouette-based approach. It computes a gray-scale image where each pixel contains the temporal history of the motion of that pixel. In this type of representation, the noisy silhouettes have a great impact. [64] propose a new descriptor based on MSI, the Motion Energy Image (MEI). In this new representation the effect of noise in one frame is minimized with the energy of the other frames, assigning to each pixel the mean energy of each silhouette within a fixed size window.

Recent approaches [32, 17] use stacks of random silhouettes to represent the gait information, where each frame is handled independently to extract features that are combined with other frame features to build the final gait signature. However, all these approaches are sensitive to changes in the body shape produced by clothing or camera view-points.

In contrast to descriptors based on silhouettes, [13] proposes a method that uses dense local spatio-temporal features and a Fisher-based representation rearranged as tensors. Another example without silhouettes is the work carried out

by [93], which uses a Kinect camera with an integrated depth sensor for skeleton detection and tracking in real-time.

The advent of Deep Learning architectures [41] has started a new realm of the feature learning field for recognition tasks. In image-based tasks, CNNs approaches have been used with great success [58, 101, 129]. In the last years, the state-of-the-art of action recognition has been marked by deep video architecture, where subsequent stacked frames serve as input for CNNs.

In the gait recognition field, [49] extract gait features from binary silhouettes using Restricted Boltzmann Machines. However, they use a small probe set (*i.e.* only ten subjects) for validating their approach. [124] extract high-level features that are used in a multi-task framework, where the goals are gait, angle view and scene recognition. They use as input data for a CNN the GEI descriptors computed on complete walking cycles. In [122] the authors propose a CNN that accumulates the obtained features, in order to obtain a global representation of the dataset, using a random set of binary silhouettes of a sequence. In [39], authors use raw 2D GEI to train an ensemble of CNNs using as classifier a Multilayer Perceptron (MLP). Similarly, in [3] a multilayer CNN is trained with GEI data. In addition, on [123] the authors developed a new approach based on GEI, where they train a CNN using pairs of gallery-probe samples.

Recently, optical flow maps have been used to represent gait, since they are easy to compute and focus on the subject's motion, independently of body shape. For example, Castro *et al.* [12] use different CNN models to identify subjects from optical flow maps. In [10] the authors use of optical flow as input for training a CNN for gait recognition, obtaining state-of-the-art results.

### 3.3. Multi-Task Approaches

In addition to using gait as an approach to identify people, it can be also applied to other tasks such as gender recognition or age estimation. Usually, these tasks are treated as independent of the main task [47, 11], but [73] shows that there is a relationship between gait, age and gender, and that multi-task training improves the identification of the subject. However, in the actual state-of-the-art, not much attention has been paid to the fact that those tasks are closely related and can benefit one from others. However, in this thesis, we explore in multiple papers the benefits of a novel multi-task CNN-based architecture for gait recognition.

### 3.4. Multimodal Approaches

Although works on gait used to focus on a single input modality, the use of multimodal models where multiples inputs data (*e.g.* optical flow, gray images, depth maps) are used is becoming more common. In such cases, fusion techniques can be used to improve the performance of the processing applied to those data. Two main methods can be employed for data fusion: early fusion and late fusion. On the one hand, early fusion methods, also known as feature fusion methods [117, 134], take data from multiple sensors and produce different features, which are merged at some stage of the pipeline to build a combined descriptor. On the other hand, late fusion methods, also known as decision fusion methods [54], fuse the output of independent classifiers by applying some kind of arithmetic operations. Another option is explored in [23, 15] where early and late fusion are applied together.

Kumar *et al.* [59] use data obtained from multiple inertial sensors to obtain a 3D-skeleton representation together with video images. Also, in [10], it is proposed a CNN model that uses optical flow, depth and gray images at the same time to improve the global gait accuracy of the model. Zhang *et al.* [131] uses RGB inputs to obtain an intermediate representation based on skeletons to disentangle appearance details. Then, global temporal descriptors are obtained with a LSTM from the intermediate skeletons.

### 3.5. Knowledge Distillation Approaches

As examples of knowledge distillation, there are approaches such as those of [21, 103], where a student model learns from RGB information to produce features or signatures similar to optical flow ones obtained from a teacher model. Then, both are combined to boost the final results, without the need for optical flow computation at test time.

Teacher-student approaches also use signatures to compress deep learning models through knowledge distillation between predictions of a big model (*i.e.* the teacher) and a smaller one (*i.e.* the student) [46, 6]. Other recent works have tried to improve the resulting compressed models. Thus, in [130] an ensemble of students learns collaboratively and teaches each other during training. Distillation performance can degrade when differences in size between teacher and student models are big. To alleviate this problem a multi-step knowledge distillation can be applied [79]. In [128] a teacher-free knowledge distillation is

proposed in which the student is able to learn from itself.

# 4 Published Work

---

This chapter collects the set of papers published during the PhD. Specifically, five papers have been published in journals indexed in the Journal of Citation Report (JCR). Moreover, another paper has been published at international conferences. Thus, six papers are the outcome of this thesis.

## 4.1. List of Published Papers

### ■ Journal papers:

- Delgado-Escáñ, R., Castro, F. M., Cázár, J. R., Marín-Jiménez, M. J., Guil, N. (2018). An end-to-end multi-task and fusion CNN for inertial-based gait recognition. *IEEE Access*, 7, 1897-1908.
- Delgado-Escáñ, R., Castro, F. M., R Cázár, J., Marin-Jimenez, M. J., Guil, N. (2020). MuPeG—The Multiple Person Gait Framework. *Sensors*, 20(5), 1358.
- Delgado-Escáñ, R., Castro, F. M., Cázár, J. R., Marín-Jiménez, M. J., Guil, N., Casilar, E. (2020). A cross-dataset deep learning-based classifier for people fall detection and identification. *Computer methods and programs in biomedicine*, 184, 105265.
- Marín-Jiménez, M. J., Castro, F. M., Delgado-Escáñ, R., Kalogeiton, V., Guil, N. (2021). UGaitNet: Multimodal Gait Recognition With Missing Input Modalities. *IEEE Transactions on Information Forensics and Security*, 16, 5452-5462.

- Delgado-Escáño, R., Castro, F. M., Guil, N., Marín-Jiménez, M. J. (2021). GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy. *IEEE Access*, 9, 164339-164347.

- **Conference papers:**

- Delgado-Escáño, R., Castro, F. M., Guil, N., Kalogeiton, V., Marín-Jiménez, M. J. (2021, September). Multimodal gait recognition under missing modalities. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3003-3007).

## 4.2. Summary of the papers that support this thesis

This section presents a summary of each one of the journal papers that support this thesis. For each one of them, we attach a brief summary and a full copy of the published document.

### 4.2.1. Reference [25] ‘An end-to-end multi-task and fusion CNN for inertial-based gait recognition’

In [25], we propose an end-to-end approach based on CNNs which automatically extracts discriminant features from a gait sequence using the raw data acquired directly from the inertial sensors without any pre-processing step. We propose to use some of the techniques to take advantage of the information included in the dataset [85] to be used. This dataset contains information from multiple inertial sensors (*i.e.* accelerometer and gyroscope), thus, we plan to fusion the information from all of them so that the model can build better features and, consequently, improve the global accuracy. In addition, as it also includes three labels per sample with information about *identity*, *age* and *gender* of a subject, our model processes the gait information in a multi-task setup to jointly recognize all of them. Note that, although we focus on that dataset, our approach applies to any dataset that contains one or more sensors and labels for one or more characteristics of the subjects (*i.e.* id, gender, age, etc.). Also, an identity verification (or *authentication*) system has been implemented to decide if two different samples belong to the same subject.

An ablation study is performed to measure the impact on the classification of the two proposed techniques to maximize the utilization of the dataset. Also, ex-

periments on authentication are performed and the trained models are compared with the state-of-the-art.

#### **4.2.2. Reference [28] ‘MuPeG—The Multiple Person Gait Framework’**

There is an important lack of realistic datasets in the gait recognition field, especially concerning the number of subjects in the scene and occlusion of the subjects with respect to the view of the camera. To solve this absence of datasets dealing with real-life situations that allow the development of robust gait recognition approaches, we propose a new framework, called Multiple Person Gait framework (MuPeG). This framework produces augmented datasets with multiple subjects in the scene by taking advantage of existing gait datasets. Thus, MuPeG combines an arbitrary number of subjects from existing datasets and creates new realistic video sequences that contain several subjects. In this way, videos with multiple subjects can be generated without recording a new dataset, which is a tedious and slow process. Moreover, our framework can produce different situations such as people walking beside or crossing each other, which produces multiple kinds of occlusions. Consequently, it can generate gait sequences more similar to real situations, setting up a new kind of benchmark that, to the best of our knowledge, did not exist before.

To measure the suitability and difficulty of the generated dataset, we propose an experimental methodology composed of two types of experiments that we recommend to perform in the synthetic data generated with our framework. The first one focuses on validating the video adequacy and realism, and the second one focuses on measuring the performance of gait recognition approaches in scenes with multiple persons.

#### **4.2.3. Reference [24] ‘A cross-dataset deep learning-based classifier for people fall detection and identification’**

Our objective in [24] is to develop a new approach, based on deep learning, for fall detection and people identification that can be used in different datasets without any fine-tuning of the model parameters. This also allows it to be used with new subjects, which is an important advantage in this use case to be able to use the fall predictor with elderly people without the need for training samples, taking into account that in this situation the sampling can be dangerous for the subject.

As result, we present a dataset-independent deep learning-based model that, by employing a multi-task learning approach, uses raw inertial information as input to simultaneously solve two tasks: fall detection and subject identification.

Experiments show that our cross-dataset classifier is able to detect falls with more than a 98% of accuracy in four datasets recorded under different conditions and with different subjects. Moreover, the number of false positives is very low (less than 1.6%), establishing a new state-of-the-art. Finally, it is proved that training our feature extractor only with young people, the differences between young and elderly people for all metrics are less than 0.4%, showing the robustness of our approach.

#### **4.2.4. Reference [27] ‘GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy’**

Traditionally, gait recognition has been tackled using appearance-related features, as RGB images, or shape-based features such as silhouettes. However, these types of features are sensitive to changes either in the scene or in the subject itself. The typical solution proposed for this in the state-of-the-art is the use of optical flow, which is a motion-based feature of the subject. However, the computation of optical flow is computationally expensive for some applications, for example, in typical embedded systems with low computational capacity and low energy consumption.

Seeking to solve this computational problem, we propose GaitCopy, a deep neural network for gait recognition that generates motion-based features, as with optical flow, but using as input gray-scale images. For this we use knowledge distillation, teaching a model with gray-scale images to imitate the features obtained in another model with optical flow.

We experimentally show, in CASIA-B and TUM-GAID, that a result similar to that of the master networks can be obtained by the gray-based student networks, without the need of explicitly computing optical flow, and even using fewer network parameters (up to  $\times 4.2$  smaller).

### **4.3. Copies of the papers that support this thesis**

This section includes a copy of each of the five published journal papers.

## An End-to-End Multi-Task and Fusion CNN for Inertial-Based Gait Recognition

**Bibliographical Reference:** Delgado-Escano, R., Castro, F. M., Cózar, J. R., Marín-Jiménez, M. J., & Guil, N. (2018). An end-to-end multi-task and fusion CNN for inertial-based gait recognition. *IEEE Access*, 7, 1897-1908.

**Abstract:** People identification using gait information (i.e., the way a person walks) obtained from inertial sensors is a robust approach that can be used in multiple situations where vision-based systems are not applicable. Typically, previous methods use hand-crafted features or deep learning approaches with pre-processed features as input. In contrast, we present a new deep learning-based end-to-end approach that employs raw inertial data as input. By this way, our approach is able to automatically learn the best representations without any constraint introduced by the pre-processed features. Moreover, we study the fusion of information from multiple inertial sensors and multi-task learning from multiple labels per sample. Our proposal is experimentally validated on the challenging dataset OU-ISIR, which is the largest available dataset for gait recognition using inertial information. After conducting an extensive set of experiments to obtain the best hyper-parameters, our approach is able to achieve state-of-the-art results. Specifically, we improve both the identification accuracy (from 83.8% to 94.8%) and the authentication equal-error-rate (from 5.6 to 1.1). Our experimental results suggest that: 1) the use of hand-crafted features is not necessary for this task as deep learning approaches using raw data achieve better results; 2) the fusion of information from multiple sensors allows to improve the results; and, 3) multi-task learning is able to produce a single model that obtains similar or even better results in multiple tasks than the corresponding models trained for a single task.

DOI: [10.1109/ACCESS.2018.2886899](https://doi.org/10.1109/ACCESS.2018.2886899)

### MuPeG—The Multiple Person Gait Framework

**Bibliographical Reference:** Delgado-Escano, R., Castro, F. M., R. Cózar, J., Marin-Jimenez, M. J., & Guil, N. (2020). Mupeg—the multiple person gait framework. Sensors, 20(5), 1358.

**Abstract:** Gait recognition is being employed as an effective approach to identify people without requiring subject collaboration. Nowadays, developed techniques for this task are obtaining high performance on current datasets (usually more than 90% of accuracy). However, those datasets are simple as they only contain one subject in the scene at the same time. This fact limits the extrapolation of the results to real world conditions where, usually, multiple subjects are simultaneously present at the scene, generating different types of occlusions and requiring better tracking methods and models trained to deal with those situations. Thus, with the aim of evaluating more realistic and challenging situations appearing in scenarios with multiple subjects, we release a new framework (MuPeG) that generates augmented datasets with multiple subjects using existing datasets as input. By this way, it is not necessary to record and label new videos, since it is automatically done by our framework. In addition, based on the use of datasets generated by our framework, we propose an experimental methodology that describes how to use datasets with multiple subjects and the recommended experiments that are necessary to perform. Moreover, we release the first experimental results using datasets with multiple subjects. In our case, we use an augmented version of TUM-GAID and CASIA-B datasets obtained with our framework. In these augmented datasets the obtained accuracies are 54.8% and 42.3% whereas in the original datasets (single subject), the same model achieved 99.7% and 98.0% for TUM-GAID and CASIA-B, respectively. The performance drop shows clearly that the difficulty of datasets with multiple subjects in the scene is much higher than the ones reported in the literature for a single subject. Thus, our proposed framework is able to generate useful datasets with multiple subjects which are more similar to real life situations.

**DOI:** [10.3390/s20051358](https://doi.org/10.3390/s20051358)

## A cross-dataset deep learning-based classifier for people fall detection and identification

**Bibliographical Reference:** Delgado-Escano, R., Castro, F. M., Cozar, J. R., Marin-Jimenez, M. J., Guil, N., & Casilar, E. (2020). A cross-dataset deep learning-based classifier for people fall detection and identification. *Computer methods and programs in biomedicine*, 184, 105265.

**Abstract:** **Background and Objective:** Fall detection is an important problem for vulnerable sectors of the population such as elderly people, who frequently live alone. Note that a fall can be very dangerous for them if they cannot ask for help. Hence, in those situations, an automatic system that detected and informed to emergency services about the fall and subject identity could help to save lives. This way, they would know not only when but also who to help. Thus, our objective is to develop a new approach, based on deep learning, for fall detection and people identification that can be used in different datasets without any fine-tuning of the model parameters.

**Methods:** We present a dataset-independent deep learning-based model that, by employing a multi-task learning approach, uses raw inertial information as input to solve simultaneously two tasks: fall detection and subject identification. By this way, our approach is able to automatically learn the best representations without any constraint introduced by the pre-processed features.

**Results:** Our cross-dataset classifier is able to detect falls with more than a 98% of accuracy in four datasets recorded under different conditions (i.e. accelerometer device, sampling rate, sequence length, age of the subjects, etc.). Moreover, the number of false positives is very low – on average less than 1.6% – establishing a new state-of-the-art. Finally, our classifier is also capable of correctly identifying people with an average accuracy of 79.6%.

**Conclusions:** The presented approach performs both tasks (fall detection and people identification) by using a single model and achieving real-time execution. The obtained results allow us to assert that a single model can be used for both fall detection and people identification under different conditions, easing its real implementation, as it is not necessary to train the model for new subjects.

DOI: [10.1016/j.cmpb.2019.105265](https://doi.org/10.1016/j.cmpb.2019.105265)

### GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy

**Bibliographical Reference:** Delgado-Escáño, R., Castro, F. M., Guil, N., & Marín-Jiménez, M. J. (2021). GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy. *IEEE Access*, 9, 164339-164347.

**Abstract:** This paper addresses the problem of gait-based people identification by copying optical flow-based signatures. The proposed model, coined as GaitCopy, receives as input a stack of gray images and returns the gait signature of the represented subject. The novel property of this network is that it is not trained to only generate discriminative signatures, but to copy signatures generated by a Master network trained on optical flow inputs. Then, GaitCopy is enforced to extract signatures based on motion and not based on appearance, despite having been trained with pixel inputs. We implement two different versions of GaitCopy, one mainly composed of 3D convolutional layers to capture local temporal information; and a second one based on GaitSet which uses 2D convolutional layers under a temporal setup. We evaluate our approach on two public gait datasets: CASIA-B and TUM-GAID. We observe that compact networks, up to  $\times 4.2$  smaller for TUM-GAID, can be obtained by using our approach, while keeping a competitive recognition accuracy with respect to the state of the art, and without the need of explicit optical flow computation. Even with such network compression, the results obtained in TUM-GAID are comparable to those of the state of the art, with an average accuracy of 97% on the test set.

**DOI:** [10.1109/ACCESS.2021.3134705](https://doi.org/10.1109/ACCESS.2021.3134705)

## 4.4. Additional papers

This section includes two papers that are not part of the thesis but we consider important for an overall understanding of the work carried out. We attach a brief summary and a full copy of the published documents.

### 4.4.1. Reference [26] ‘Multimodal Gait Recognition Under Missing Modalities’

Typical approaches of gait recognition use a single modality, there is a single source of input information that provides only one kind of information. However, multimodal systems are currently being widely used in the field of gait recognition thanks in part to cheaper devices that can capture different information simultaneously. Also, this provides more information to the models and this results in more robust and accurate prediction systems. In these systems multiple inputs are received simultaneously, which can be different transformations of the same information source or even different information sources.

Though, these types of systems are susceptible to loss of performance due to the lack of any of the input modalities, doing the system not robust to missing modalities. Based on this, We propose a single and flexible framework that uses a variable number of input modalities. For each modality, it consists of a branch and a binary unit indicating whether the modality is available; these are gated and merged together by ‘max’ operation. As a result, it generates a single and compact ‘multimodal’ gait signature that encodes biometric information of the input. Our framework outperforms the state of the art on TUM-GAID and extensive experiments reveal its effectiveness for handling missing modalities even in the multiview setup of CASIA-B.

### 4.4.2. Reference [75] ‘UGaitNet: Multimodal Gait Recognition With Missing Input Modalities’

In this work, we extend the conference version paper [26]. The proposed framework for solving the problem of missing modalities in multimodal systems is extended with a novel ‘sign-max’ merge operation, that uses both positive and negative values. Also, in the triplet loss, the Semi-Hard triplet selection is changed by BatchAll, causing that more triplets contribute to the training gradients.

At experimentation level, silhouette modality is included in the three previously modalities studied in the conference paper (optical flow, gray and depth). It is also shown that the framework is branch-agnostic, so branches can be built with different architectures. Additionally, experiments realized in the conference paper are expanded, particularly in CASIA-B, where we included a cross-view experimentation.

## 4.5. Copy of the additional paper

## Multimodal Gait Recognition Under Missing Modalities

**Bibliographical Reference:** Delgado-Escano, R., Castro, F. M., Guil, N., Kalogeiton, V., & Marín-Jiménez, M. J. (2021, September). Multimodal gait recognition under missing modalities. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 3003-3007). IEEE.

**Abstract:** Multimodal systems for gait recognition have gained a lot of attention. However, there is a clear gap in the study of missing modalities, which represents real-life scenarios where sensors fail or data get corrupted. Here, we investigate how to handle missing modalities for gait recognition. We propose a single and flexible framework that uses a variable number of input modalities. For each modality, it consists of a branch and a binary unit indicating whether the modality is available; these are gated and merged together. Finally, it generates a single and compact ‘multimodal’ gait signature that encodes biometric information of the input. Our framework outperforms the state of the art on TUM-GAIT and extensive experiments reveal its effectiveness for handling missing modalities even in the multiview setup of CASIA-B. The code is available online: <https://github.com/avagait/gaitmiss>.

DOI: [10.1109/ICIP42928.2021.9506162](https://doi.org/10.1109/ICIP42928.2021.9506162)

## UGaitNet: Multimodal Gait Recognition With Missing Input Modalities

**Bibliographical Reference:** Marín-Jiménez, M. J., Castro, F. M., Delgado-Escáño, R., Kalogeiton, V., & Guil, N. (2021). UGaitNet: Multimodal Gait Recognition With Missing Input Modalities. *IEEE Transactions on Information Forensics and Security*, 16, 5452-5462.

**Abstract:** Gait recognition systems typically rely solely on silhouettes for extracting gait signatures. Nevertheless, these approaches struggle with changes in body shape and dynamic backgrounds; a problem that can be alleviated by learning from multiple modalities. However, in many real-life systems some modalities can be missing, and therefore most existing multimodal frameworks fail to cope with missing modalities. To tackle this problem, in this work, we propose UGaitNet, a unifying framework for gait recognition, robust to missing modalities. UGaitNet handles and mingles various types and combinations of input modalities, i.e. pixel gray value, optical flow, depth maps, and silhouettes, while being camera agnostic. We evaluate UGaitNet on two public datasets for gait recognition: CASIA-B and TUM-GAID, and show that it obtains compact and state-of-the-art gait descriptors when leveraging multiple or missing modalities. Finally, we show that UGaitNet with optical flow and grayscale inputs achieves almost perfect (98.9%) recognition accuracy on CASIA-B (same-view “normal”) and 100% on TUM-GAID (“ellapsed time”). Code will be available at <https://github.com/avagait/ugaitnet>

**DOI:** [10.1109/TIFS.2021.3132579](https://doi.org/10.1109/TIFS.2021.3132579)

UNIVERSIDAD  
DE MÁLAGA



# 5

# Conclusions

---

This chapter presents the final thoughts of this thesis. Section 5.1 summarizes the conclusions from the previously described works, and Section 5.2 describes the future lines of work that we plan to explore.

## 5.1. Conclusions

Gait recognition is a well-studied problem, with an extensive state-of-the-art and excellent results in the usual datasets. However, in our review of the state-of-the-art, we found what we believe to be problems or weaknesses that have been little studied, which we decided to focus on with the aim of helping future research to address them. Thus, the thesis has focused on the development and application of techniques that have allowed us to improve the state-of-the-art from the perspective of datasets, input data, classification or generalization of the proposed systems.

First, in [25], we implemented a new end-to-end approach, based on CNN architectures, for the gait-based recognition and authentication problems that uses raw inertial data as input. A fusion scheme has also been proposed, which takes advantage of data obtained from several inertial sensors. This improves the prediction of the system with respect to the use of this information separately. In addition, we have developed a multi-task learning model that works with the multiple labels of the dataset. This shows that there is a relationship between soft-biometrics tasks and gait, and their joint training improves subject identification. Extensive cross-validation has been employed to establish the best

hyper-parameter values of the models, such as the layer to fuse, the weight of each loss function, etc.

Continuing with inertial sensors, in [24], we have presented a new cross-dataset classifier based on a deep architecture and a  $k$ -NN classifier for fall detection and people identification. Our method is able to detect falls and identify subjects at the same time using a single model, thanks to the multi-task learning approach used, without requiring an additional training process per dataset.

Focusing on vision-based datasets, MuPeG [28] is a framework to generate augmented gait datasets with multiple persons in the scene using existing datasets. This framework allows researchers to build a new type of datasets that did not exist before, which permits to deal with realistic gait analysis problems. In addition, an experimental study has been carried out to demonstrate how these more realistic cases present a new challenge to be addressed.

Returning to multimodal approaches, we have faced the problem of missing modalities, which usually causes a high decrease in the performance of the models or even the impossibility of their use. On gait recognition, we proposed a multimodal framework [26, 75], adaptable to any network. It uses a logic gates mechanics and a merge function that let to deactivate inputs which no information available. This allows it to be used with a minimum loss of accuracy even keeping the results close to the state-of-the-art in these cases.

Finally, in [27] we propose the use of knowledge distillation to teach a network with gray images as input to mimic the features of a network with optical flow as input. This implies an optimization for the network both in the number of parameters and previous computation time, since the optical flow is computationally expensive. We demonstrate experimentally that student networks, trained with gray inputs, can obtain similar accuracy to master networks, trained with optical flow data. And, it is visually proven that the signatures of the student networks are more similar to those trained with optical flow than to those trained with gray from scratch.

## 5.2. Future Work

Multiple lines of future work are proposed, considered after the completion of the different works that make up this thesis:

- In [25], we plan to study in detail the multi-task setup to improve its performance. Additionally, we plan to study how gait is affected by illness or

fatigue in terms of recognition and soft-biometrics classification accuracy.

- We plan in [24] to improve the performance of our approach to solve the two cases where we obtain lower results: MAA in UniMiB-SHAR dataset and Accuracy in ASLH dataset. Also, our intuition indicates that including information from more sensors may improve the training process and the results obtained by our approach. In addition, we intend to study the viability of a real deployment of our architecture in hardware with low computational capacity and low power consumption.
- Taking advantage of the framework MuPeG [28], we are working to publish a new dataset, derived from CASIA-B, that presents more realistic situations to work with. With this, we plan to develop both new tracking approaches and better CNN models able to deal with multiple persons in a scene. Moreover, we will try to define more experiments to take advantage of the multiple person gait datasets generated with our framework.
- Another line of research, based on [26, 75], would be the addition of new modalities, such as body pose, to already trained UGaitNet models.
- Also, in [27], we plan to investigate the performance of the model on other combinations of master-student modalities, such as depth maps or silhouettes.

UNIVERSIDAD  
DE MÁLAGA



# **Apéndice A**

## **Resumen en español**

---

La biometría es la medición de los seres vivos o de los procesos biológicos, que puede utilizarse para identificar a los seres humanos, o sus rasgos generales, a partir de sus rasgos específicos y personales. El campo de la biometría [22] está experimentando un rápido crecimiento, impulsado por la necesidad de aplicaciones robustas de seguridad y vigilancia. Sin embargo, su potencial como medio de identificación natural y sin esfuerzo también ha allanado el camino para una gran cantidad de aplicaciones que identifican automáticamente al usuario o características específicas del mismo, proporcionando servicios personalizados. Los principales tipos de sistemas biométricos actuales se basan en el reconocimiento de factores como la huella dactilar [51], el rostro [112], la retina [88], la voz [82] o la firma [34].

El paso puede considerarse como un patrón biométrico inequívoco de la locomoción humana, ya que cada sujeto tiene sus propias características biológicas, lo que hace viable casos de uso como la identificación de personas a partir de la forma de caminar. Aunque el paso ha sido tradicionalmente un campo estudiado desde el punto de vista de la medicina, (por ejemplo, para el diagnóstico precoz de enfermedades como la enfermedad de Parkinson [96], el síndrome de Rett [52], o la parálisis cerebral [90]), también se ha estudiado desde el punto de vista de la seguridad biométrica.

Entre sus ventajas en el caso de la seguridad biométrica, destaca que el patrón de los pasos puede obtenerse de forma no invasiva y sin que el usuario colabore activamente con el sistema, a diferencia de otros métodos tradicionales como el análisis del iris, el reconocimiento facial o el uso de huellas dactilares. Esto permite su uso en entornos en los que no se pueden utilizar otros patrones biométricos, por ejemplo, cuando los sujetos deben llevar una vestimenta especial, como los

trajes NBQ (trajes nucleares, biológicos y químicos), o cuando el entorno impone limitaciones a los sistemas biométricos (posición de la cámara, ocultación del rostro, legislación sobre privacidad, etc.). También hay que tener en cuenta que, debido a su dinámica de origen natural, estos patrones son difíciles de duplicar.

El problema del reconocimiento de la forma de caminar se ha estudiado tradicionalmente desde el punto de vista de la visión por ordenador, algo que, al poder realizarse a distancia y no depender de la observación de partes concretas del cuerpo, no requiere la colaboración del sujeto y resuelve los problemas antes mencionados. En este campo, en el estado del arte se utilizan diferentes tipos de datos de entrada: los basados en la apariencia, como RGB, escala de grises o siluetas; o los basados en el movimiento, como el Gait Energy Image (GEI) [43] o el flujo óptico. En cuanto a la metodología de clasificación, es evidente el dominio de las redes neuronales convolucionales (CNN) en el estado del arte, ya que este tipo de modelo es actualmente el mejor tipo de algoritmo para el procesamiento automático de imágenes y vídeos.

Aunque hay muchos trabajos de reconocimiento del paso a basados en la visión [116], un problema común en los conjuntos de datos típicos para este caso es la suposición de la presencia de un solo sujeto en la escena en un momento dado, lo que limita su aplicación en un entorno real.

Además, el análisis de la marcha a partir de sensores iniciales se ha convertido en un tema activo y explotado gracias al abaratamiento de los sensores MEMS (Micro Electro Mechanical Systems) y su integración en smartphones [80] o smartwatches [53], lo que abarata y hace más portátil este sistema de análisis de la marcha. En este caso, se suele estudiar la información procedente de acelerómetros y giroscopios y, a diferencia de lo que ocurre en los enfoques basados en la visión, no abunda en el estado del arte el uso del aprendizaje profundo para la clasificación, siendo más común el uso de sistemas tradicionales de aprendizaje automático.

## A.1. Motivaciones de la Tesis

La motivación de esta tesis es investigar sobre cómo resolver los problemas o puntos de inflexión a los que se enfrenta actualmente el estado del arte del reconocimiento de la forma de caminar. Para ello, se ha analizado el estado del arte actual desde el punto de vista de los conjuntos de datos, los datos de entrada, los clasificadores, la robustez de los sistemas propuestos y las posibles optimizaciones que se pueden desarrollar.

Tras este análisis, se han propuesto una serie de objetivos para proporcionar, durante el desarrollo de esta tesis, aportaciones que permitan superar los resultados actuales, resolver los problemas detectados y abrir, si es posible, nuevas líneas de investigación.

## A.2. Objetivos y Fases

A continuación, definiremos los objetivos que se han planteado en el desarrollo de la tesis:

1. Clasificación de rasgos biométricos suaves distintos a la identidad y relacionados con la marcha. Por ejemplo, la edad y el sexo.
2. Implementación de un clasificador cross-dataset para el reconocimiento de la marcha sin necesidad de ajustar los modelos a cada conjunto de datos distinto.
3. Implementación de un sistema de generación de muestras de vídeo con múltiples sujetos simultáneos, utilizando diferentes técnicas de visión por computador y aprendizaje profundo. Asimismo, se estudia la clasificación simultánea de múltiples sujetos.
4. Estudio de las modalidades ausentes, es decir, la pérdida de información de una o varias modalidades (o fuentes de información) en la entrada de un sistema basado en el aprendizaje profundo, y cómo abordarlo.
5. Uso de técnicas de destilación de conocimientos para optimizar los modelos de reconocimiento del paso.

Para alcanzar estos objetivos, se han llevado a cabo las siguientes fases:

1. Revisamos el estado del arte para seleccionar los mejores enfoques y los conjuntos de datos utilizados para evaluar nuestros métodos y compararlos con otros enfoques del estado del arte. Decidimos utilizar DFNAPAS [77], SisFall [105], UniMiB-SHAR [78] ASLH [91] y OU-ISIR [85] en enfoques basados en sensores iniciales, y CASIA-B [127] y TUM-GAID [47] en enfoques basados en visión por computador. Este paso ha sido necesario para alcanzar todos los objetivos propuestos.

2. Para el desarrollo del clasificador de biometría suave [25], se han estudiado diferentes sensores inerciales como entrada, estudiados por tipo y posición, y se ha trabajado en la fusión de diferentes tipos de sensores. Además, se han comparado modelos equivalentes en configuraciones de una y múltiples tareas para observar el impacto de la clasificación simultánea de estas. Con este paso se ha cumplido el objetivo 1 y se ha proporcionado información sobre las técnicas multimodales y multitarea para los siguientes objetivos.
3. Utilizando los conocimientos adquiridos sobre sensores inerciales en el paso 2, hemos implementado un clasificador cross-dataset centrado en la detección de caídas [24]. Se han utilizado múltiples conjuntos de datos estudiados en el paso 1 para probar el rendimiento del enfoque cross-dataset sin necesidad de entrenamiento adicional. También se ha estudiado el uso de un clasificador externo a la red, el clasificador  $k$ -Nearest Neighbors, que no requiere un proceso de entrenamiento para los datos utilizados, pero ha sido necesario ajustar su hiperparámetro  $k$ . Buscando su aplicación en un caso de uso realista, se ha comprobado un alto rendimiento en la clasificación de caídas en un grupo de personas mayores realizando un entrenamiento sólo con personas jóvenes, lo que demuestra la viabilidad del enfoque cross-dataset. Esto nos ha permitido cumplir el objetivo 2.
4. Se ha trabajado en un framework capaz de generar muestras sintéticas con múltiples sujetos a partir de muestras reales con un único sujeto [28]. Para ello se han probado múltiples técnicas de segmentación, seleccionando al final una CNN del estado del arte. Se ha estudiado y propuesto una metodología experimental para comprobar la calidad visual de las imágenes generadas. Para ello, se ha explorado el impacto de múltiples sujetos por escena en los modelos de clasificación, diferenciando entre la precisión a nivel de sujeto y a nivel de grupo de sujetos en el vídeo. También se ha observado el impacto del solapamiento de sujetos, diferenciando también entre los vídeos en los que los sujetos se cruzan o caminan en paralelo. Este paso se ha realizado con los dos conjuntos de datos de visión seleccionados en el paso 1 y completa el objetivo 3, permitiendo a los investigadores trabajar con nuevos conjuntos de datos con casos más realistas.
5. Se ha implementado un sistema basado en puertas lógicas que permite, en el momento de la inferencia, desactivar una o varias ramas de entrada, permitiendo que el modelo siga funcionando en los casos de falta de modalidades [26, 75]. Para combinar la información de las diferentes modalidades, se han probado y comparado múltiples funciones de fusión para las diferentes modalidades utilizadas. Se han probado modelos análogos de una

sola tarea con fusión tardía con nuestros modelos multitarea propuestos. Además, hemos experimentado con modelos iguales aplicando o no nuestro enfoque, para estudiar el efecto de nuestro enfoque en la precisión. Este paso ha completado el objetivo 4, con una arquitectura capaz de tratar con las modalidades que faltan sin una degradación significativa del rendimiento.

6. Se propone utilizar destilación de conocimiento para enseñar a los modelos a imitar una modalidad de entrada computacionalmente costosa, el flujo óptico, mediante una modalidad de entrada menos costosa, imágenes en escala de grises [27]. Se implementan dos conjuntos de modelos maestros, uno propio y otro basado en un modelo del estado del arte, para estudiar la generalización de nuestra propuesta. Se han realizado estudios de ablación para comprobar que las técnicas propuestas tienen un impacto real en el proceso de destilación del conocimiento. Con este paso se ha cumplido el objetivo 5, obteniendo un modelo de menor tamaño que en la inferencia requiere un menor coste computacional.

## A.3. Contribuciones

Las aportaciones de esta tesis según los objetivos propuestos son:

- Para el objetivo 1, clasificación de rasgos biométricos suaves, las contribuciones son:
  1. Un nuevo enfoque integral, basado en una arquitectura CNN, para problemas de reconocimiento y autentificación basados en la forma de caminar que utiliza datos inerciales en bruto como entrada.
  2. Se ha propuesto también un esquema de fusión que aprovecha los datos obtenidos de distintos tipos de sensores inerciales, acelerómetro y giroscopio, para generar modelos más robustos. Se ha estudiado el impacto de esto en la clasificación.
  3. Un modelo de clasificación multitarea que aprovecha las múltiples etiquetas contenidas en el conjunto de datos utilizado.
  4. Nuevos resultados del estado del arte para la identidad, el género, la edad y la autenticación en el conjunto de datos OU-ISIR [85].
- Para el objetivo número 2, clasificador cross-dataset, las contribuciones son:
  1. Un nuevo clasificador cross-dataset basado en una arquitectura de aprendizaje profundo y un clasificador  $k$ -NN.

2. Un clasificador capaz de detectar las caídas e identificar a los sujetos al mismo tiempo utilizando un único modelo, con un enfoque multitarea.
  3. Mejora de los resultados obtenidos por propuestas anteriores del estado del arte en cuatro conjuntos de datos públicos diferentes.
- Para el objetivo 3, el sistema de generación de vídeo con múltiples sujetos simultáneos, las contribuciones son:
1. El primer framework para generar nuevos conjuntos de datos para el reconocimiento del paso con múltiples personas en la escena (MuPeG) utilizando conjuntos de datos existentes. Este marco permite a los investigadores construir un nuevo tipo de conjuntos de datos que no existía antes en el estado del arte.
  2. El nuevo framework abre nuevos retos para los investigadores con problemas realistas de análisis de la marcha.
  3. Se ha propuesto una metodología experimental que define el número y tipo de experimentos mínimos que deben realizarse en este tipo de conjuntos de datos.
  4. Se establecen unos resultados de referencia para la metodología experimental propuesta.
- En el objetivo 4, mitigar el efecto de las modalidades no disponibles, las contribuciones son:
1. Se ha desarrollado UGaitNet, una red que maneja y combina varios tipos de modalidades de entrada para el reconocimiento del paso: imágenes en escala de grises, flujo óptico, mapas de profundidad y/o siluetas.
  2. Se demuestra que esta red es tolerante a los fallos con respecto a la ausencia de una o varias modalidades de entrada.
  3. Se obtienen descriptores del paso comparables con el estado del arte cuando se evalúa en el dataset TUM-GAIT.
- Para el objetivo 5, destilación de conocimientos para la optimización de computo, las contribuciones son:
1. Se ha introducido un nuevo enfoque denominado GaitCopy, basado en la destilación de conocimientos, que es capaz de imitar el comportamiento de las redes basadas en flujo óptico para el reconocimiento de la marcha, pero utilizando como entrada imágenes en escala de grises.

2. Se demuestra experimentalmente en CASIA-B y TUM-GAID que, comparando con las redes maestras entrenadas con flujo óptico, se puede obtener una precisión similar en las redes alumnos entrenadas con escala de grises.
3. Las redes diseñadas para escala de grises son significativamente más pequeñas en número de parámetros que sus análogas basadas en flujo óptico.
4. Se optimiza también el tiempo de inferencia al evitar calcular el flujo óptico en los nuevos modelos, lo cual es computacionalmente costoso.

## A.4. Publicaciones

### ■ Artículos de revista:

- Delgado-Escáño, R., Castro, F. M., Cázar, J. R., Marín-Jiménez, M. J., Guil, N. (2018). An end-to-end multi-task and fusion CNN for inertial-based gait recognition. *IEEE Access*, 7, 1897-1908.
- Delgado-Escáño, R., Castro, F. M., R. Cázar, J., Marin-Jimenez, M. J., Guil, N. (2020). MuPeG—The Multiple Person Gait Framework. *Sensors*, 20(5), 1358.
- Delgado-Escáño, R., Castro, F. M., Cázar, J. R., Marín-Jiménez, M. J., Guil, N., Casilar, E. (2020). A cross-dataset deep learning-based classifier for people fall detection and identification. *Computer methods and programs in biomedicine*, 184, 105265.
- Marín-Jiménez, M. J., Castro, F. M., Delgado-Escáño, R., Kalogeiton, V., Guil, N. (2021). UGaitNet: Multimodal Gait Recognition With Missing Input Modalities. *IEEE Transactions on Information Forensics and Security*, 16, 5452-5462.
- Delgado-Escáño, R., Castro, F. M., Guil, N., Marín-Jiménez, M. J. (2021). GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy. *IEEE Access*, 9, 164339-164347.

### ■ Artículos de conferencias internacionales:

- Delgado-Escáño, R., Castro, F. M., Guil, N., Kalogeiton, V., Marín-Jiménez, M. J. (2021, September). Multimodal gait recognition under missing modalities. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 3003-3007).

## A.5. Resumen de los artículos que apoyan esta tesis

En esta sección se presenta un resumen de cada uno de los artículos publicados en revistas que apoyan esta tesis.

### A.5.1. Referencia [25] ‘An end-to-end multi-task and fusion CNN for inertial-based gait recognition’

En [25], proponemos un enfoque integral basado en CNNs que extrae automáticamente características discriminantes de una secuencia de marcha utilizando los datos en bruto adquiridos directamente de los sensores iniciales sin ningún preprocesamiento. También proponemos el uso de técnicas para aprovechar la información adicional incluida en el conjunto de datos [85]. Este conjunto de datos contiene información de múltiples sensores iniciales (acelerómetro y giroscopio), por lo que planeamos fusionar la información de todos ellos para que el modelo pueda construir mejores características y, en consecuencia, mejorar la precisión global. Además, al incluir también tres etiquetas por muestra con información sobre la *identidad*, la *edad* y el *género* de un sujeto, nuestro modelo procesa la información del paso en una configuración multitarea para reconocer conjuntamente todas ellas. Obsérvese que, aunque nos centramos en ese conjunto de datos, nuestro planteamiento es aplicable a cualquier conjunto de datos que contenga uno o más sensores y etiquetas para una o más características de los sujetos (por ejemplo, identificación, sexo, edad, etc.). Además, se ha implementado un sistema de verificación de identidad (o *autenticación*) para decidir si dos muestras diferentes pertenecen al mismo sujeto.

Se realiza un estudio de ablación para medir el impacto en la clasificación de las dos técnicas propuestas para maximizar la utilización del conjunto de datos. También se realizan experimentos de autenticación y se comparan los modelos entrenados con el estado del arte.

### A.5.2. Referencia [28] ‘MuPeG—The Multiple Person Gait Framework’

Existe una importante falta de datos realistas en el campo del reconocimiento del paso, especialmente en lo que respecta al número de sujetos en escena y a la ocultación de los sujetos con respecto a la vista de la cámara. Para resolver esta

ausencia en los conjuntos de datos que traten situaciones de la vida real y que permitan el desarrollo de enfoques robustos del paso, proponemos un nuevo framework, denominado Multiple Person Gait framework (MuPeG). Este framework produce conjuntos de datos aumentados artificialmente con múltiples sujetos en la misma escena aprovechando los datasets ya existentes. Así, MuPeG combina un número arbitrario de sujetos y crea nuevas secuencias de vídeo realistas que contienen varios sujetos. De este modo, se pueden generar vídeos con múltiples sujetos sin necesidad de grabar un nuevo conjunto de datos, lo que supone un proceso tedioso y lento. Además, nuestro framework puede producir diferentes situaciones, como personas que caminan unas al lado de otras o personas que se cruzan entre si, lo que produce distintos tipos de occlusiones. En consecuencia, puede generar secuencias del paso más parecidas a situaciones reales, estableciendo un nuevo tipo de referencia que, hasta donde sabemos, no existía antes.

Para medir la idoneidad y dificultad del conjunto de datos generado, proponemos una metodología experimental compuesta por dos tipos de experimentos que recomendamos realizar en los datos generados con nuestro framework. El primero se centra en validar el realismo y la correlación del vídeo generado con respecto a los originales, y el segundo se centra en medir el rendimiento de los enfoques de reconocimiento de la marcha en escenas con múltiples personas.

### A.5.3. Referencia [24] ‘A cross-dataset deep learning-based classifier for people fall detection and identification’

Nuestro objetivo en [24] es desarrollar un nuevo enfoque, basado en deep learning, para la detección de caídas e identificación de personas que pueda ser utilizado en diferentes datasets sin necesidad de afinar los parámetros del modelo para cada conjunto de datos por separado. Esto también permite su uso con nuevos sujetos, lo que supone una ventaja importante en este caso de uso para poder utilizar el predictor de caídas con personas mayores sin necesidad de muestras de entrenamiento, teniendo en cuenta que en esta situación el muestreo puede ser peligroso para el sujeto.

Como resultado, presentamos un modelo basado en deep learning independiente del conjunto de datos que, empleando un enfoque de aprendizaje multitarifa, utiliza la información inercial bruta como entrada para resolver simultáneamente dos tareas: la detección de caídas y la identificación del sujeto.

Los experimentos muestran que nuestro clasificador cross-dataset es capaz de detectar las caídas con más de un 98 % de precisión en cuatro conjuntos de datos registrados en diferentes condiciones y con diferentes sujetos. Además, el



número de falsos positivos es muy bajo (menor al 1,6 %), estableciendo un nuevo estado del arte. Por último, se demuestra que entrenando nuestro extractor de características sólo con personas jóvenes, la diferencia entre personas jóvenes y personas ancianas para todas las métricas es inferior al 0,4 %, lo que demuestra la robustez de nuestro enfoque.

#### A.5.4. Referencia [27] ‘GaitCopy: Disentangling Appearance for Gait Recognition by Signature Copy’

Tradicionalmente, el reconocimiento del paso se ha abordado o utilizando características relacionadas con la apariencia, como las imágenes RGB, o características basadas en la forma, como las siluetas. Sin embargo, estos tipos de características son sensibles a los cambios en la escena o en el propio sujeto. La solución tradicionalmente propuesta para esto en el estado del arte es el uso del flujo óptico, que es una característica basada en el movimiento del sujeto. Sin embargo, el cálculo del flujo óptico es costoso desde el punto de vista computacional para algunas aplicaciones, por ejemplo, en los típicos sistemas embebidos de baja capacidad computacional y bajo consumo de energía.

Buscando resolver este problema computacional, proponemos GaitCopy, una red neuronal profunda para el reconocimiento del paso que genera características basadas en el movimiento, como con el flujo óptico, pero utilizando como entrada imágenes en escala de grises. Para ello utilizamos la destilación de conocimiento, enseñando a un modelo con imágenes en escala de grises a imitar las características obtenidas en otro modelo con flujo óptico.

Demostramos experimentalmente en CASIA-B y TUM-GAID que se puede obtener en las redes estudiantes entrenadas en escala de grises un resultado similar al de las redes maestras, sin necesidad de calcular explícitamente el flujo óptico e incluso utilizando menos parámetros en la red (hasta  $\times 4.2$  veces más pequeños).

### A.6. Publicaciones adicionales

Esta sección incluye dos artículos que no forman parte de la tesis pero que consideramos importantes para la comprensión global del trabajo realizado.

### A.6.1. Referencia [26] ‘Multimodal Gait Recognition Under Missing Modalities’

Los enfoques típicos de reconocimiento de la marcha utilizan una sola modalidad, hay una sola fuente de información de entrada que proporciona un solo tipo de información. Sin embargo, los sistemas multimodales se utilizan actualmente de forma generalizada en el campo del reconocimiento de la marcha gracias, en parte, a dispositivos más baratos que pueden capturar diferentes informaciones simultáneamente. Además, esto proporciona más información a los modelos y esto da lugar a sistemas de predicción más robustos y precisos. En estos sistemas se reciben múltiples entradas simultáneamente, que pueden ser diferentes transformaciones de la misma fuente de información o incluso diferentes fuentes de información.

Sin embargo, este tipo de sistemas son susceptibles de perder rendimiento debido a la falta de alguna de las modalidades de entrada, haciendo que el sistema no sea robusto a las modalidades que faltan. Basándonos en esto, proponemos un marco único y flexible que utiliza un número variable de modalidades de entrada. Para cada modalidad, consta de una rama y una unidad binaria que indica si la modalidad está disponible; éstas se cierran y se fusionan mediante la operación “max”. Como resultado, genera una única y compacta firma de la marcha “multimodal” que codifica la información biométrica de la entrada. Nuestro marco supera el estado de la técnica en TUM-GAIT y amplios experimentos revelan su eficacia para manejar las modalidades que faltan, incluso en la configuración multivista de CASIA-B.

### A.6.2. Referencia [75] ‘UGaitNet: Multimodal Gait Recognition With Missing Input Modalities’

En este trabajo ampliamos el artículo [26], publicado en una conferencia internacional. El framework propuesto para resolver el problema de las modalidades ausentes en los sistemas multimodales se amplía con una novedosa operación de fusión ‘sign-max’, que utiliza tanto valores positivos como negativos. Además, en el triplet loss, la selección de triplets Semi-Hard se cambia por BatchAll, causando que más triplets contribuyan a los gradientes de entrenamiento.

A nivel de experimentación, se incluye la modalidad de silueta a las tres modalidades previamente estudiadas en el artículo publicado anteriormente en conferencia (flujo óptico, escala de grises y profundidad). También se muestra que el framework es agnóstico a las ramas, por lo que se pueden construir ramas

con diferentes arquitecturas. Además, se amplían los experimentos realizados en parala conferencia, especialmente en CASIA-B, donde se incluyen experimentos enfocados en las vistas cruzadas.

## A.7. Conclusiones

El reconocimiento del paso es un problema bien estudiado, con un extenso estado del arte y excelentes resultados en los conjuntos de datos habituales. Sin embargo, en nuestra revisión del estado del arte, hemos encontrado lo que creemos que son problemas o debilidades poco estudiadas, en los que hemos decidido centrarnos con el objetivo de ayudar a futuras investigaciones a abordarlos. Así, esta tesis se ha centrado en el desarrollo y aplicación de técnicas que nos han permitido mejorar el estado del arte desde la perspectiva de los datasets, los datos de entrada, la clasificación o la generalización de los sistemas propuestos.

En primer lugar, en [25], implementamos un nuevo modelo basado en arquitecturas CNN para los problemas de reconocimiento y autenticación basados en la forma de caminar que utiliza datos iniciales en bruto como entrada. También se ha propuesto un esquema de fusión que aprovecha los datos obtenidos de varios sensores iniciales. Esto mejora la predicción del sistema con respecto al uso de esta información por separado. Además, el modelo desarrollado realiza aprendizaje multitarea al trabajar con las distintas etiquetas del conjunto de datos. Esto demuestra que existe una relación entre las tareas relacionadas con la biometría suave y el paso, y que su entrenamiento conjunto mejora la identificación del sujeto. Se ha empleado una extensa validación cruzada para establecer los mejores valores de los hiperparámetros de los modelos, como la capa a fusionar, el peso de cada función de perdida, etc.

Siguiendo con los sensores iniciales, en [24] hemos presentado un nuevo clasificador cross-dataset basado en una arquitectura profunda y un clasificador  $k$ -NN para la detección de caídas e identificación de personas. Nuestro método es capaz de detectar caídas e identificar sujetos al mismo tiempo utilizando un único modelo, gracias al enfoque de aprendizaje multitarea utilizado, sin requerir un proceso de entrenamiento adicionalpara cada conjunto de datos distinto.

Centrándonos en los conjuntos de datos basados en visión, MuPeG [28] es un framework para generar datasets artificiales con múltiples sujetos en escena utilizando datos ya existentes. Este framework permite a los investigadores construir escenarios de entrenamiento que no existían antes, lo que permite abordar problemas de análisis del paso más realistas. Además, se ha llevado a cabo un

estudio experimental para demostrar que estos escenarios plantean un nuevo reto.

Volviendo a los enfoques multimodales, nos hemos enfrentado al problema de las modalidades ausentes, que suele provocar una elevada disminución del rendimiento de los modelos o incluso imposibilitar su uso. Hemos propuesto un framework multimodal [26, 75] para reconocimiento del paso adaptable a cualquier red. Utilizando un mecanismo basado en puertas lógicas y una función de fusión puede desactivar las entradas de las que no dispone de información, lo que permite seguir utilizando el modelo con una pérdida mínima de precisión, manteniendo incluso resultados cercanos al estado del arte en esos casos.

Finalmente, en [27] proponemos el uso de la destilación de conocimiento para enseñar a una red con imágenes en escala de grises como entrada a imitar las características de salida de una red que usa como entrada flujo óptico. Esto implica una optimización para la red tanto en número de parámetros como en tiempo de cómputo en el pre-procesamiento debido al alto coste computacional del flujo óptico. Demostramos experimentalmente que las redes estudiantes, entrenada con imágenes en escala de grises, pueden obtener una precisión similar a la redes maestras, entrenada con flujo óptico. Además, se demuestra visualmente que las características obtenidas por la redes alumnos se parece más a las producidas por las redes entrenadas con flujo óptico que a las entrenadas con imágenes en escala de grises desde el principio.

## A.8. Trabajo Futuro

Se proponen múltiples líneas de trabajo futuro, consideradas tras la realización de los diferentes trabajos que componen esta tesis:

- En [25], se planea estudiar en detalle la configuración multitarea para mejorar su rendimiento. Además, tenemos previsto estudiar cómo la marcha se ve afectada por condiciones como la enfermedad o la fatiga en términos de precisión de reconocimiento y en la clasificación de rasgos biometría suave.
- Planeamos en [24] mejorar el rendimiento de nuestro procedimiento para resolver los dos casos en los que obtenemos los resultados más bajos: MAA en el dataset UniMiB-SHAR y la precisión de la clasificación en el dataset ASLH. Además, nuestra intuición nos indica que incluir información de más sensores puede mejorar el proceso de entrenamiento y los resultados obtenidos por nuestra aproximación. Además, pretendemos estudiar la viabilidad de un despliegue real de nuestra arquitectura en un hardware de

baja capacidad computacional y bajo consumo energético.

- Aprovechando el framework MuPeG [28], estamos trabajando para publicar un nuevo conjunto de datos, derivado del dataset CASIA-B, que presenta situaciones más realistas con las que trabajar. Con esto, planeamos desarrollar tanto nuevos métodos de seguimiento en escena de los sujetos como mejores modelos basados en CNNs capaces de tratar con múltiples personas de forma simultanea en una escena. Además, se definirán más experimentos sobre los nuevos tipos de conjuntos de datos generados con nuestro framework.
- Otra línea de investigación, basada en [26, 75], sería la adición de nuevas modalidades, como la pose del cuerpo, a los modelos basados en UGaitNet ya entrenados.
- También, sobre [27], tenemos previsto investigar el rendimiento del modelo en otras combinaciones de modalidades maestro-alumno, como los mapas de profundidad o las siluetas.

# Bibliography

- [1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018. (Cited on page 9)
- [2] Dharmitha Ajerla, Sazia Mahfuz, and Farhana Zulkernine. A Real-Time Patient Monitoring Framework for Fall Detection. *Wireless Communications and Mobile Computing*, 2019:1–13, 09 2019. (Cited on page 29)
- [3] M. Alotaibi and A. Mahmood. Improved gait recognition based on specialized deep convolutional neural networks. In *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7, 2015. (Cited on pages 27 and 31)
- [4] D Anguita, A Ghio, L Oneto, X Parra, and J Reyes-Ortiz. *Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine*, pages 216–223. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. (Cited on page 29)
- [5] Christian Arnold and Marcel Neunhoeffer. Really useful synthetic data—a framework to evaluate the quality of differentially private synthetic data. *arXiv preprint arXiv:2004.07740*, 2020. (Cited on page 25)
- [6] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *NeurIPS*, volume 27, 2014. (Cited on page 32)
- [7] Subhash C Bagui. Combining pattern classifiers: methods and algorithms, 2005. (Cited on page 16)



- [8] N. V. Boulgouris, K. N. Plataniotis, and D. Hatzinakos. Gait recognition using dynamic time warping. In *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, 2004. (Cited on page 28)
- [9] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. (Cited on page 19)
- [10] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca. Multimodal feature fusion for CNN-based gait recognition: an empirical comparison. *Neural Computing and Applications*, 2020. (Cited on pages 13, 30, 31 and 32)
- [11] Francisco M Castro, Manuel J Marín-Jiménez, and Nicolás Guil. Multi-modal features fusion for gait, gender and shoes recognition. *Machine Vision and Applications*, 27(8):1213–1228, 2016. (Cited on pages 29 and 31)
- [12] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Santiago López-Tapia, and Nicolás Pérez de la Blanca. Evaluation of CNN architectures for gait recognition based on optical flow maps. In *BIOSIG*, pages 251–258, 2017. (Cited on page 31)
- [13] Francisco M Castro, Manuel J Marín-Jimenez, and Rafael Medina-Carnicer. Pyramidal fisher motion for multiview gait recognition. In *2014 22nd International Conference on Pattern Recognition*, pages 1692–1697. IEEE, 2014. (Cited on page 30)
- [14] Francisco M. Castro, M.J. Marín-Jiménez, N. Guil Mata, and R. Muñoz Salinas. Fisher motion descriptor for multiview gait recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(1), 2017. (Cited on page 27)
- [15] Yanmei Chai, Jie Ren, Huimin Zhao, Yang Li, Jinchang Ren, and Paul Murray. Hierarchical and multi-featured fusion for effective gait recognition under variable scenarios. *Pattern Analysis and Applications*, 19:905–917, 2015. (Cited on page 32)
- [16] H. Chang, Y. Hsu, S. Yang, J. Lin, and Z. Wu. A wearable inertial measurement system with complementary filter for gait analysis of patients with stroke or parkinson’s disease. *IEEE Access*, 4:8442–8453, 2016. (Cited on page 27)
- [17] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Cross-view Gait Recognition through Utilizing Gait as a Deep

- Set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. (Cited on pages 13 and 30)
- [18] H Cho and SM Yoon. Applying singular value decomposition on accelerometer data for 1D convolutional neural network based fall detection. *Electronics Letters*, 55(6):320–322, 2019. (Cited on page 29)
- [19] Sangil Choi, Ik-Hyun Youn, R. LeMay, S. Burns, and Jong-Hoon Youn. Biometric gait recognition based on wireless acceleration sensor using k-nearest neighbor classification. In *2014 International Conference on Computing, Networking and Communications (ICNC)*, pages 1091–1095, 2014. (Cited on page 28)
- [20] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1 – 27, 2018. (Cited on page 28)
- [21] Nieves Crasto, Philippe Weinzaepfel, Kartek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. (Cited on page 32)
- [22] Shaveta Dargan and Munish Kumar. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, 143:113114, 2020. (Cited on pages 1 and 51)
- [23] Omid Dehzangi, Mojtaba Taherisadr, and Raghvendar ChangalVala. Imu-based gait recognition using convolutional neural networks and multi-sensor fusion. In *Sensors*, 2017. (Cited on pages 27, 28 and 32)
- [24] Rubén Delgado-Escáño, Francisco M Castro, Julián R Cózar, Manuel J Marín-Jiménez, Nicolás Guil, and Eduardo Casilar. A cross-dataset deep learning-based classifier for people fall detection and identification. *Computer methods and programs in biomedicine*, 184:105265, 2020. (Cited on pages 5, 36, 48, 49, 54, 59, 62 and 63)
- [25] Ruben Delgado-Escano, Francisco M Castro, Julián Ramos Cózar, Manuel J Marín-Jiménez, and Nicolas Guil. An end-to-end multi-task and fusion cnn for inertial-based gait recognition. *IEEE Access*, 7:1897–1908, 2018. (Cited on pages 4, 35, 47, 48, 54, 58, 62 and 63)
- [26] Rubén Delgado-Escáño, Francisco M Castro, Nicolás Guil, Vicky Kalogeiton, and Manuel J Marín-Jiménez. Multimodal gait recognition under

- missing modalities. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2021. (Cited on pages 5, 42, 48, 49, 54, 61, 63 and 64)
- [27] Ruben Delgado-Escáño, Francisco M Castro, Nicolas Guil, and Manuel J Marín-Jiménez. Gaitcopy: Disentangling appearance for gait recognition by signature copy. *IEEE Access*, 9:164339–164347, 2021. (Cited on pages 5, 37, 48, 49, 55, 60, 63 and 64)
- [28] Ruben Delgado-Escano, Francisco M Castro, Julián R Cózar, Manuel J Marin-Jimenez, and Nicolas Guil. Mupeg—the multiple person gait framework. *Sensors*, 20(5):1358, 2020. (Cited on pages 5, 36, 48, 49, 54, 58, 62 and 64)
- [29] M. O. Derawi, P. Bours, and K. Holien. Improved cycle detection for accelerometer based gait authentication. In *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 312–317, 2010. (Cited on pages 23 and 28)
- [30] Mohammad Derawi and Patrick Bours. Gait and activity recognition using commercial phones. *computers & security*, 39:137–144, 2013. (Cited on page 27)
- [31] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. (Cited on page 10)
- [32] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Sahui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020. (Cited on pages 13 and 30)
- [33] Zheyi Fan, Jiao Jiang, Shuqin Weng, Zhonghang He, and Zhiwen Liu. Human gait recognition based on discrete cosine transform and linear discriminant analysis. In *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–6. IEEE, 2016. (Cited on page 30)
- [34] Marcos Faundez-Zanuy. Signature recognition state-of-the-art. *IEEE aerospace and electronic systems magazine*, 20(7):28–32, 2005. (Cited on pages 1 and 51)
- [35] P. Fernandez-Lopez, J. Sanchez-Casanova, P. Tirado-Martín, and J. Liu-Jimenez. Optimizing resources on smartphone gait recognition. In *2017*

- IEEE International Joint Conference on Biometrics (IJCB)*, pages 31–36, Oct 2017. (Cited on page 29)
- [36] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989. (Cited on page 18)
- [37] Matteo Gadaleta and Michele Rossi. Idnet: Smartphone-based gait recognition with convolutional neural networks. *Pattern Recognition*, 74:25–37, 2018. (Cited on pages 14, 27 and 28)
- [38] Davrondzhon Gafurov, Einar Snekkenes, and Patrick Bours. Improved gait recognition performance using cycle matching. In *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, pages 836–841. IEEE, 2010. (Cited on pages 23 and 28)
- [39] Bence Gálai and Csaba Benedek. Feature selection for lidar-based gait recognition. In *Computational Intelligence for Multimedia Understanding (IWCIM), 2015 International Workshop on*, pages 1–5, 2015. (Cited on page 31)
- [40] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Prentice Hall, 2008. (Cited on page 12)
- [41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>. (Cited on pages 10 and 31)
- [42] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. (Cited on page 10)
- [43] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006. (Cited on pages 2, 30 and 52)
- [44] J. He, Z. Zhang, X. Wang, and S. Yang. A Low Power Fall Sensing Technology Based on FD-CNN. *IEEE Sensors Journal*, 19(13):5110–5118, July 2019. (Cited on page 29)
- [45] Simon Herrlich, Sven Spieth, Rachid Nouna, Roland Zengerle, Libero I. Giannola, Diego Esteban Pardo-Ayala, Eugenio Federico, and Pierangelo

- Garino. *Ambulatory Treatment and Telemonitoring of Patients with Parkinson's Disease*, pages 295–305. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. (Cited on page 27)
- [46] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS DLRL Workshop*, 2015. (Cited on page 32)
- [47] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1):195 – 206, 2014. (Cited on pages 2, 4, 23, 24, 29, 31 and 53)
- [48] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. (Cited on page 13)
- [49] Emdad Hossain and Girija Chetty. Multimodal feature learning for gait biometric based human identity recognition. In *Neural Information Processing*, pages 721–728, 2013. (Cited on page 31)
- [50] Haifeng Hu. Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(7):1274–1286, July 2013. (Cited on page 30)
- [51] Tsai-Yang Jea and Venu Govindaraju. A minutia-based partial fingerprint recognition system. *Pattern recognition*, 38(10):1672–1684, 2005. (Cited on pages 1 and 51)
- [52] K. Jellinger, D. Armstrong, H. Y. Zoghbi, and A. K. Percy. Neuropathology of rett syndrome. *Acta Neuropathologica*, 76(2):142–158, Mar 1988. (Cited on pages 1 and 51)
- [53] A. H. Johnston and G. M. Weiss. Smartwatch-based biometric gait recognition. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015. (Cited on pages 3 and 52)
- [54] L. Kaliciak, H. Myrhaug, A. Goker, and D. Song. On the duality of specific early and late fusion strategies. In *17th International Conference on Information Fusion (FUSION)*, pages 1–8, 2014. (Cited on page 32)

- [55] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. (Cited on page 17)
- [56] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013. (Cited on page 15)
- [57] Siddhartha Khandelwal and Nicholas Wickström. Novel methodology for estimating initial contact events from accelerometers positioned at different body locations. *Gait & posture*, 59:278–285, 2018. (Cited on pages 14 and 28)
- [58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. (Cited on pages 30 and 31)
- [59] Pradeep Kumar, Subham Mukherjee, Rajkumar Saini, Pallavi Kaushik, Partha Pratim Roy, and Debi Prosad Dogra. Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm. *IEEE Transactions on Fuzzy Systems*, 27(5):956–965, 2018. (Cited on page 32)
- [60] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Cell phone-based biometric identification. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–7, 2010. (Cited on pages 14, 28 and 29)
- [61] Kristof Van Laerhoven and Ozan Cakmakci. What shall we teach our pants? In *Proceedings of the 4th IEEE International Symposium on Wearable Computers*, ISWC ’00, pages 77–83. IEEE Computer Society, 2000. (Cited on page 27)
- [62] Toby HW Lam and Raymond ST Lee. A new representation for human gait recognition: Motion silhouettes image (msi). In *International Conference on Biometrics*, pages 612–618. Springer, 2006. (Cited on page 30)
- [63] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999. (Cited on page 10)

- [64] Heesung Lee, Sungjun Hong, Imran Fareed Nizami, and Euntai Kim. A noise robust gait representation: Motion energy image. *International Journal of Control, Automation and Systems*, 7(4):638–643, 2009. (Cited on page 30)
- [65] Guto Leoni, Patricia Endo, Kayo Monteiro, Elisson Rocha, Ivanovitch Silva, and Theodore Lynn. Accelerometer-Based Human Fall Detection Using Convolutional Neural Networks. *Sensors*, 19:1644, 04 2019. (Cited on page 29)
- [66] H. Li, A. Shrestha, H. Heidari, J. Le Kernev, and F. Fioranelli. Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection. *IEEE Sensors Journal*, pages 1–1, 2019. (Cited on page 29)
- [67] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021. (Cited on page 13)
- [68] Aneta Lisowska., Alison O’Neil., and Ian Poole. Cross-cohort Evaluation of Machine Learning Approaches to Fall Detection from Accelerometer Data. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5 HEALTHINF: HEALTHINF*, pages 77–82. INSTICC, SciTePress, 2018. (Cited on page 29)
- [69] Aneta Lisowska, Gavin Wheeler, Victor Ceballos Inza, and Ian Poole. An evaluation of supervised, novelty-based and hybrid approaches to fall detection using silmee accelerometer data. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–16, 2015. (Cited on page 29)
- [70] Yushu Liu, Junping Zhang, Chen Wang, and Liang Wang. Multiple hog templates for gait recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2930–2933. IEEE, 2012. (Cited on page 30)
- [71] Francisco Luna-Perejon, Javier Civit-Masot, Isabel Amaya-Rodriguez, Lourdes Duran-Lopez, Juan Pedro Dominguez-Morales, Anton Civit-Balcells, and Alejandro Linares-Barranco. An Automated Fall Detection System Using Recurrent Neural Networks. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 36–41. Springer, 2019. (Cited on page 29)

- [72] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S. M. Makela, and H. A. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, 2005. (Cited on page 27)
- [73] M. J. Marín-Jiménez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer. Deep multi-task learning for gait-based biometrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2017. (Cited on pages 27 and 31)
- [74] Manuel J Marín-Jiménez, Francisco M Castro, Ángel Carmona-Poyato, and Nicolás Guil. On how to improve tracklet-based gait recognition systems. *Pattern Recognition Letters*, 68:103–110, 2015. (Cited on page 27)
- [75] Manuel J Marín-Jiménez, Francisco M Castro, Rubén Delgado-Escáño, Vicky Kalogeiton, and Nicolas Guil. Ugaitnet: Multimodal gait recognition with missing input modalities. *IEEE Transactions on Information Forensics and Security*, 16:5452–5462, 2021. (Cited on pages 5, 42, 48, 49, 54, 61, 63 and 64)
- [76] Taylor Mauldin, Marc Canby, Vangelis Metsis, Anne Ngu, and Coralys Rivera. SmartFall: A smartwatch-based fall detection system using deep learning. *Sensors*, 18(10):3363, 2018. (Cited on page 29)
- [77] Carlos Medrano, Raul Igual, Inmaculada Plaza, and Manuel Castro. Detecting Falls as Novelties in Acceleration Patterns Acquired with Smartphones. *PLOS ONE*, 9(4):1–9, 04 2014. (Cited on pages 4, 22 and 53)
- [78] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones (2017). *Applied Sciences*, 7(10), 2017. (Cited on pages 4, 22 and 53)
- [79] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *arXiv preprint arXiv:1902.03393*, 2019. (Cited on page 32)
- [80] J. Le Moing and I. Stengel. The smartphone as a gait recognition device impact of selected parameters on gait recognition. In *2015 International Conference on Information Systems Security and Privacy (ICISSP)*, 2015. (Cited on pages 3 and 52)

- [81] M. Muaaz and C. Nickel. Influence of different walking speeds and surfaces on accelerometer-based biometric gait recognition. In *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, pages 508–512, 2012. (Cited on page 23)
- [82] LINDASALWA MUDA, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010. (Cited on pages 1 and 51)
- [83] Mirto Musci, Daniele De Martini, Nicola Blago, Tullio Facchinetti, and Marco Piastra. Online fall detection using recurrent neural networks (2018). *arXiv preprint arXiv:1804.04976*, 2018. (Cited on page 29)
- [84] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. (Cited on page 17)
- [85] Thanh Trung Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition*, 47(1):228 – 237, 2014. (Cited on pages 3, 4, 6, 23, 29, 35, 53, 55 and 58)
- [86] Trung Thanh Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. Similar gait action recognition using an inertial sensor. *Pattern Recognition*, 48(4):1289 – 1301, 2015. (Cited on page 29)
- [87] C. Nickel and C. Busch. Classifying accelerometer data via hidden markov models to authenticate people by the way they walk. *IEEE Aerospace and Electronic Systems Magazine*, 28(10):29–35, 2013. (Cited on page 28)
- [88] Ishan Nigam, Mayank Vatsa, and Richa Singh. Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion*, 26:1–35, 2015. (Cited on pages 1 and 51)
- [89] T Oberg, A Karsznia, and K Oberg. Basic gait parameters : Reference data for normal subjects, 10-79 years of age. *Journal of rehabilitation research and development*, 30:210–23, 02 1993. (Cited on page 29)
- [90] M. J. O’Malley, M. F. Abel, D. L. Damiano, and C. L. Vaughan. Fuzzy clustering of children with cerebral palsy based on temporal-distance gait

- parameters. *IEEE Transactions on Rehabilitation Engineering*, 5(4):300–309, 1997. (Cited on pages 1 and 51)
- [91] Ahmet Ozdemir. An Analysis on Sensor Locations of the Human Body for Wearable Fall Detection Devices: Principles and Practice. *Sensors*, 16(8):1161, 07 2016. (Cited on pages 4, 22 and 53)
- [92] Jose Portillo-Portillo, Roberto Leyva, Victor Sanchez, Gabriel Sanchez-Perez, Hector Perez-Meana, Jesus Olivares-Mercado, Karina Toscano-Medina, and Mariko Nakano-Miyatake. Cross view gait recognition using joint-direct linear discriminant analysis. *Sensors*, 17(1):6, 2017. (Cited on page 30)
- [93] Johannes Preis, Moritz Kessel, Martin Werner, and Claudia Linnhoff-Popien. Gait recognition with Kinect. In *1st international workshop on Kinect in pervasive computing*, pages 1–4. New Castle, UK, 2012. (Cited on page 31)
- [94] J Peña Queraltà, TN Gia, Hannu Tenhunen, and T Westerlund. Edge-AI in LoRa-based Health Monitoring: Fall Detection System with Fog Computing and LSTM Recurrent Neural Networks. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 601–604. IEEE, 2019. (Cited on page 29)
- [95] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017. (Cited on pages 15 and 16)
- [96] Roberta de Melo Roiz, Enio Walker Azevedo Cacho, Manoela Macedo Pazinatto, Julia Guimarães Reis, Alberto Cliquet Jr, and Elizabeth Barasnevicius-Quagliato. Gait analysis comparing parkinson’s disease with healthy elderly subjects. *Arquivos de neuro-psiquiatria*, 68:81–86, 2010. (Cited on pages 1 and 51)
- [97] L. Rong, Z. Jianzhong, L. Ming, and H. Xiangfeng. A wearable acceleration sensor system for gait recognition. In *2007 2nd IEEE Conference on Industrial Electronics and Applications*, pages 2654–2659, 2007. (Cited on pages 23 and 28)
- [98] A. Samà, C. Pérez-Lopez, J. Romagosa, D. Rodríguez-Martín, A. Català, J. Cabestany, D. A. Pérez-Martínez, and A. Rodríguez-Molinero. Dyskinesia and motor state detection in parkinson’s disease patients with a single movement sensor. In *2012 Annual International Conference of the*

- IEEE Engineering in Medicine and Biology Society*, pages 1194–1197, 2012.  
(Cited on page 27)
- [99] Chao Shen, Yufei Chen, and Xiaohong Guan. Performance evaluation of implicit smartphones authentication via sensor-behavior analysis. *Information Sciences*, 430-431:538 – 553, 2018. (Cited on page 27)
- [100] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. (Cited on pages 16 and 30)
- [101] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2018. (Cited on page 31)
- [102] S. Sprager and M. B. Juric. An efficient hos-based gait authentication of accelerometer data. *IEEE Transactions on Information Forensics and Security*, 10(7), 2015. (Cited on page 28)
- [103] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Winter Conf. on Appl. of Computer Vision*, pages 625–634, 2020. (Cited on page 32)
- [104] Han Su and Feng-Gang Huang. Human gait recognition based on motion analysis. In *2005 International Conference on Machine Learning and Cybernetics*, volume 7, pages 4464–4468 Vol. 7, 2005. (Cited on page 27)
- [105] Angela Sucerquia, José López, and Jesús Vargas-Bonilla. SisFall: A fall and movement dataset. *Sensors*, 17(1):198, 2017. (Cited on pages 4, 22 and 53)
- [106] K. Sugandhi and G. Raju. Discriminative gait features based on signal properties of silhouette centroids. In Mayank Singh, P.K. Gupta, Vipin Tyagi, Jan Flusser, Tuncer Ören, and Rekha Kashyap, editors, *Advances in Computing and Data Sciences*, pages 680–688, Singapore, 2019. Springer Singapore. (Cited on page 30)
- [107] K. Sugandhi, F. F. Wahid, and G. Raju. Detection of human gait cycle: An overlap based approach. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pages 1–3, 2017. (Cited on page 29)
- [108] Hu Sun and Tao Yuao. Curve aligning approach for gait authentication based on a wearable accelerometer. *Physiological measurement*, 33:1111–20, 05 2012. (Cited on page 27)

- [109] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *Proc. NIPS*, 27, 06 2014. (Cited on page 29)
- [110] Seba Susan and Amitesh Kumar. The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art. *Engineering Reports*, 3(4):e12298, 2021. (Cited on page 25)
- [111] T Theodoridis, V Solachidis, N Vretos, and P Daras. Human fall detection from acceleration measurements using a Recurrent Neural Network. In *Precision Medicine Powered by pHealth and Connected Health*, pages 145–149. Springer, 2018. (Cited on page 29)
- [112] AS Tolba, AH El-Baz, and AA El-Harby. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103, 2006. (Cited on pages 1 and 51)
- [113] N. T. Trung, Y. Makihara, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi. Phase registration in a gallery improving gait authentication. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7, 2011. (Cited on page 27)
- [114] Ahmet Turan Özdemir and Billur Barshan. Detecting falls with wearable sensors using machine learning techniques. *Sensors*, 14(6):10691–10708, 2014. (Cited on page 27)
- [115] Sebastijan Šprager and Damjan Zazula. A cumulant-based method for gait identification using accelerometer data with principal component analysis and support vector machine. *WSEAS Transactions on Signal Processing*, 5, 11 2009. (Cited on page 27)
- [116] Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018. (Cited on pages 2 and 52)
- [117] X. Wang, R. Bai, X. Cui, T. Wu, and Z. Qian. Research on data fusion algorithm for attitude detection systems based on mems and magnetoresistive sensors. In *2017 9th International Conference on Advanced Infocomm Technology (ICAIT)*, pages 68–74, 2017. (Cited on page 32)
- [118] Y. Watanabe. Influence of holding smart phone for acceleration-based gait authentication. In *2014 Fifth International Conference on Emerging Security Technologies*, 2014. (Cited on page 28)

- [119] Z. Wei, W. Qinghui, D. Muqing, and L. Yiqi. A new inertial sensor-based gait recognition method via deterministic learning. In *2015 34th Chinese Control Conference (CCC)*, pages 3908–3913, 2015. (Cited on pages 23, 27 and 28)
- [120] I Wayan Wiprayoga Wisesa and Genggam Mahardika. Fall detection algorithm based on accelerometer and gyroscope sensor data using Recurrent Neural Networks. In *IOP Conference Series: Earth and Environmental Science*, volume 258, page 012035. IOP Publishing, 2019. (Cited on page 29)
- [121] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. (Cited on page 10)
- [122] Zifeng Wu, Yongzhen Huang, and Liang Wang. Learning representative deep features for image set analysis. *IEEE Trans. on Multimedia*, 17(11):1960–1968, Nov 2015. (Cited on page 31)
- [123] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1, 03 2016. (Cited on page 31)
- [124] C. Yan, B. Zhang, and F. Coenen. Multi-attributed gait identification by convolutional neural networks. In *International Congress on Image and Signal Processing (CISP)*, pages 642–647, Oct 2015. (Cited on page 31)
- [125] Haben Yhdego, Jiang Li, Steven Morrison, Michel Audette, Christopher Paolini, Mahasweta Sarkar, and Hamid Okhravi. Towards musculoskeletal simulation-aware fall injury mitigation: transfer learning with deep CNN for fall detection. In *2019 Spring Simulation Conference (SpringSim)*, pages 1–12. IEEE, 2019. (Cited on page 29)
- [126] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. (Cited on page 10)
- [127] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR’06)*, volume 4, pages 441–444. IEEE, 2006. (Cited on pages 4, 11, 24 and 53)

- [128] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. (Cited on page 32)
- [129] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. (Cited on page 31)
- [130] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. (Cited on page 32)
- [131] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2019. (Cited on page 32)
- [132] Yongjia Zhao and Suiping Zhou. Wearable device-based gait recognition using angle embedded gait dynamic images and a convolutional neural network. In *Sensors*, 2017. (Cited on pages 14, 23 and 28)
- [133] Y. Zhong and Y. Deng. Sensor orientation invariant mobile gait biometrics. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014. (Cited on pages 14, 23 and 28)
- [134] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang. Robust gait recognition by integrating inertial and rgbd sensors. *IEEE Transactions on Cybernetics*, 48(4):1136–1150, April 2018. (Cited on page 32)