IbPRIA 2023

BOOK OF ABSTRACTS

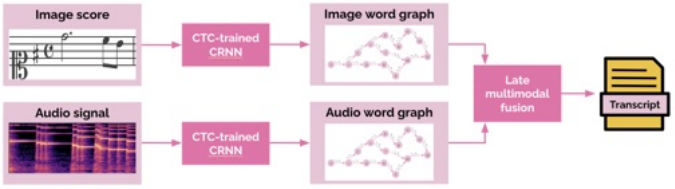
11th IBERIAN CONFERENCE ON PATTERN RECOGNITION AND IMAGE ANALYSIS

JUNE 27-30, 2023

UNIVERSITY OF ALICANTE, SPAIN

María Alfaro-Contreras

University of Alicante, Spain



Music transcription, which deals with the conversion of music sources into a structured digital format, is a key problem for Music Information Retrieval (MIR). When addressing this challenge in computational terms, the MIR community follows two lines of research: music documents, which is the case of Optical Music Recognition (OMR), or audio recordings, which is the case of Automatic Music Transcription (AMT). The different nature of the aforementioned input data has conditioned these fields to develop modality-specific frameworks. However, their recent definition in terms of sequence labeling tasks leads to a common output representation, which enables research on a combined paradigm. In this respect, multimodal image and audio music transcription comprises the challenge of effectively combining the information conveyed by image and audio modalities. In this work, we explore this question at a late-fusion level: we study four combination approaches in order to merge, for the first time, the hypotheses regarding end-to-end OMR and AMT systems in a lattice-based search space. The results obtained for a series of performance scenarios—in which the corresponding single-modality models yield different error rates—showed interesting benefits of these approaches. In addition, two of the four strategies considered significantly improve the corresponding unimodal standard recognition frameworks.

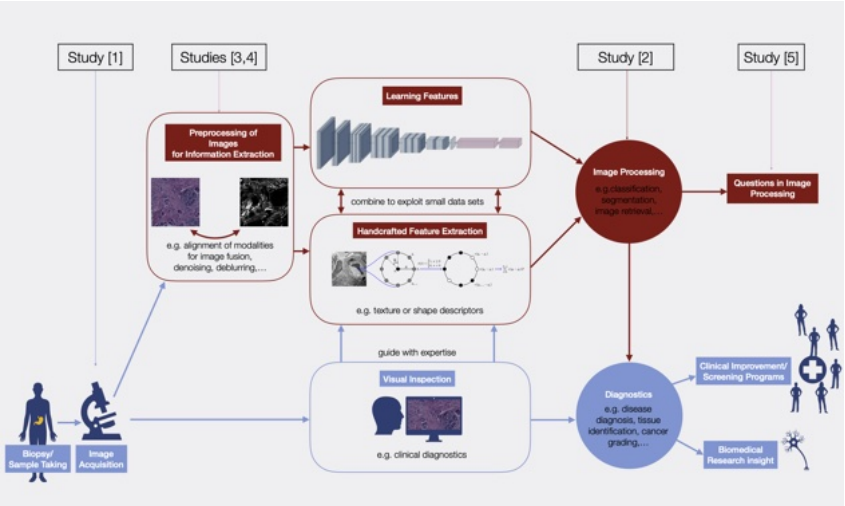
Ana Almeida

University of Alicante, Spain

Advanced learning and decision mechanisms are crucial to ensure optimal performance in a 5G and beyond system with various service verticals and virtual slices. These mechanisms help identify network trends and requirements, anomalies in expected behavior, and optimization of network resource usage, especially in dynamic scenarios. Moreover, such learning and prediction methods can benefit from application services like traffic profiling and studying traffic congestion. In a city scenario, these services can be provided in real-time to improve mobility and safety, given that large-scale learning and prediction approaches are being researched. It's important to note that mobility services and the network are interdependent, meaning that increasing urban traffic can lead to more network usage. This Thesis will focus on statistical, machine learning, and deep learning approaches to define, implement, and test network and application services in a city environment.

Elisabeth Wetzer

Uppsala University, Sweden



In recent years Machine Learning and in particular Deep Learning have excelled in object recognition and classification tasks in computer vision. As these methods extract features from the data itself by learning features that are relevant for a particular task, a key aspect of this remarkable success is the amount of data on which these methods train. Biomedical applications face the problem that the amount of training data is limited. In particular, labels and annotations are usually scarce and expensive to obtain as they require biological or medical expertise. One way to overcome this issue is to use additional knowledge about the data at hand. This guidance can come from expert knowledge, which puts focus on specific, relevant characteristics in the images, or geometric priors which can be used to exploit the spatial relationships in the images. My thesis presents machine learning methods for visual data that exploit such additional information and build upon classic image processing techniques, to combine the strengths of both model- and learning-based approaches. I have conducted five studies with applications in digital pathology. Two of them study the use and fusion of texture features within convolutional neural networks for image classification tasks. The other three studies explore rotational equivariant representation learning, and show that learned, shared representations of multimodal images can be used for multimodal image registration and cross-modality image retrieval.

Harry Dobbs

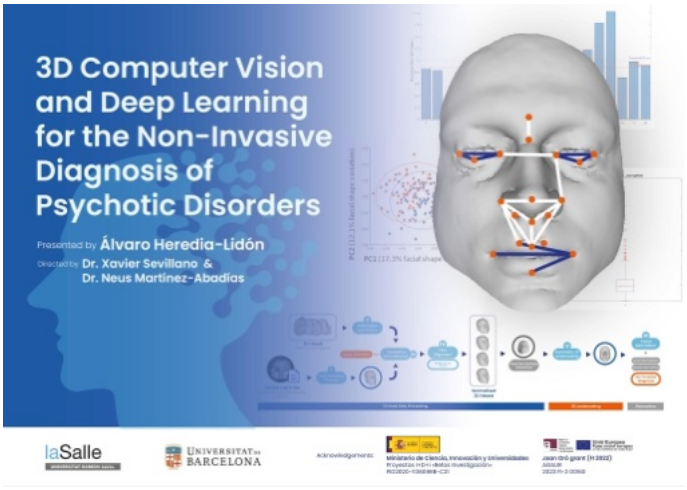
University of Canterbury, New Zealand

Accurate modelling of 3D tree geometry from point clouds is an open problem. These models have many applications such as biomass estimation, growth modelling, forestry management, urban micro-climate simulation and agritech applications such as robotic pruning and fruit picking.

We introduce Smart-Tree, a supervised method for approximating the medial axes of branch skeletons from a tree point cloud. Smart-Tree uses a sparse voxel convolutional neural network to extract the radius and direction towards the medial axis of each input point. A greedy algorithm performs robust skeletonization using the estimated medial axis. Our proposed method provides robustness to complex tree structures and improves fidelity when dealing with self-occlusions, complex geometry, touching branches, and varying point densities. We evaluate Smart-Tree using a multi-species synthetic tree dataset and perform qualitative analysis on a real-world tree point cloud. Our experimentation with synthetic and real-world datasets demonstrates the robustness of our approach over the current state-of-the-art method.

Álvaro Heredia-Lidón

La Salle - Universitat Ramon Llul, Spain



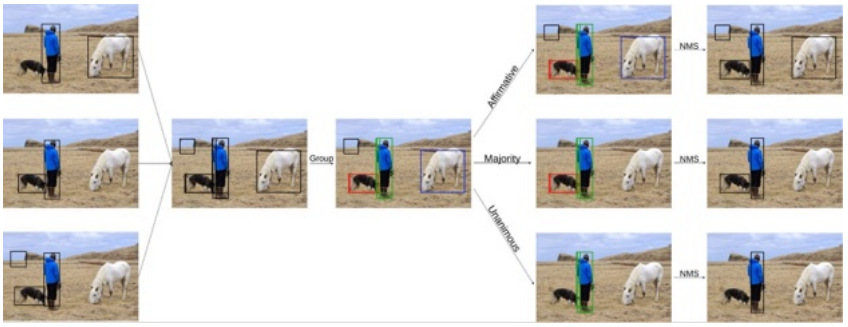
Psychotic disorders, such as schizophrenia and bipolar disorders, affect more than 3% of the world's population. However, there are no reliable clinical predictors to replace the classical clinical interview procedure for the diagnosis of these illnesses. This work is framed in the context of an innovative strategy using the potential of facial biomarkers as highly accessible and promising indicators of risk for psychotic disorder.

The aim of this project focused on the development and study of tools for the precise and automatic detection of anatomical biomarkers on three-dimensional images of the faces of sick patients. The work is divided into three distinct blocks. A first block of treatment and processing of the data to obtain a uniform facial mesh of the patient's face. A second block of implementation of a 3D landmarking method by means of point clouds registration methods, and a final block of evaluation of Deep Learning techniques in the field of automatic labelling by means of Convolutional Neural Networks.

Finally, the work concludes with a detailed analysis of the results obtained using the different methods employed and a proposal for combining their potential.

Ángela Casado-García

University of La Rioja, Spain

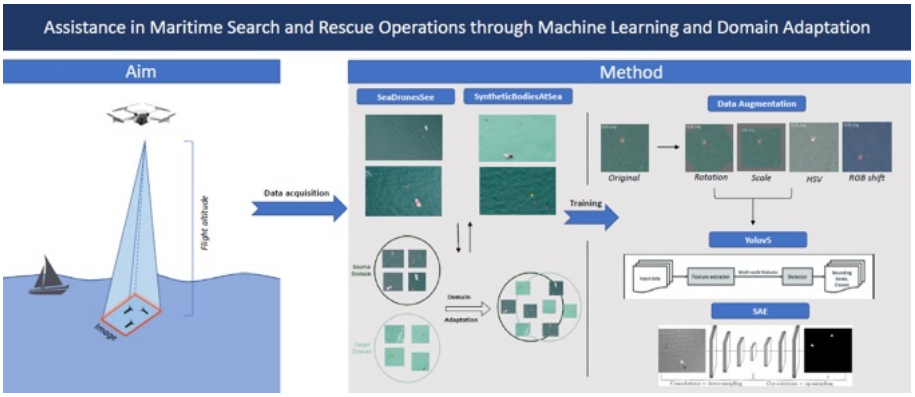


Object detection and semantic segmentation are two areas of computer vision that have numerous applications in various fields such as biology, agriculture, and medicine. Currently, the most successful techniques in these fields are based on deep learning methods. Although these methods have achieved excellent results, using such techniques in these contexts can be complex due to the large number of annotated images required (which can be difficult to obtain in the biomedical context), the resources needed to build models with them, and the technical difficulty of applying these techniques by experts. The aim of this thesis is to address these limitations through different theoretical developments and evaluate the proposed solutions in two contexts: plant physiology and precision agriculture.

The theoretical solutions fall within the development of new algorithms that improve the effectiveness of detection and segmentation models, regardless of their nature. Specifically, their improvement will be addressed in both the training and evaluation phases. All the developments made will not only remain in the theoretical domain but will also be implemented. Furthermore, the generability of the methods will be tested by evaluating them on widely used datasets within the community, such as Pascal VOC and COCO, as well as more specific contexts such as plant physiology and precision agriculture.

Juan Pedro Martínez Esteso

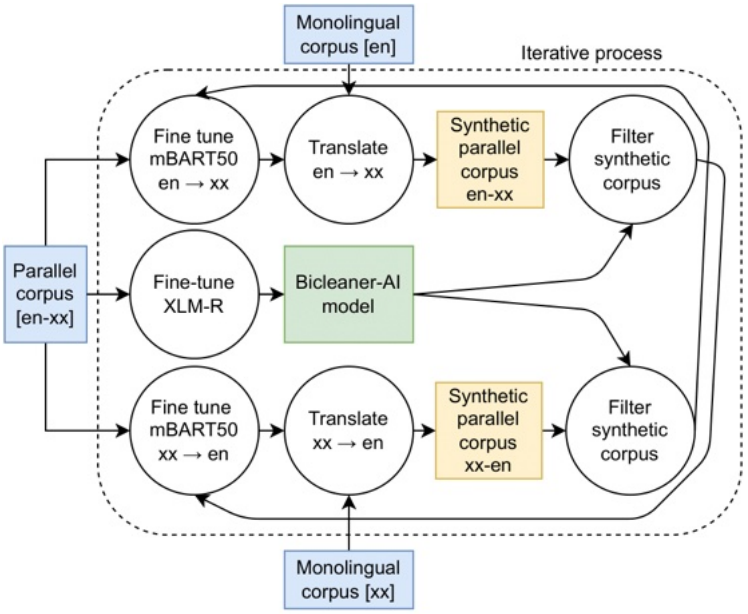
University of Alicante, Spain



The main objective of this thesis is the combination of Machine Learning techniques, domain adaptation and synthetic data generation for training, in order to generate robust models for the detection of bodies on the surface of the sea in maritime rescue tasks, which are able to respond to all kinds of scenarios and conditions.

Aarón Galiano-Jiménez

University of Alicante, Spain



Pre-trained models have revolutionized the natural language processing field by leveraging large-scale language representations for various tasks. Some pre-trained models offer general-purpose representations, while others are specialized in particular tasks, like neural machine translation (NMT). Multilingual NMT-targeted systems are often fine-tuned for specific language pairs, but there is a lack of evidence-based best-practice recommendations to guide this process. Additionally, deploying these large pre-trained models in computationally restricted environments, typically found in developing regions where low-resource languages are spoken, has become challenging.

We propose a pipeline to tune the mBART50 pre-trained model to 8 diverse low-resource language pairs, and then distill the resulting system to obtain lightweight and more sustainable NMT models. Our pipeline conveniently exploits back-translation, synthetic corpus filtering, and knowledge distillation to deliver efficient bilingual translation models that are 13 times smaller, while maintaining a close BLEU performance.

Rafael Aguilar-Ortega

University of Córdoba, Spain



Human pose estimation (HPE) is a widely researched topic in the field of Computer Vision. It involves extracting the position of individuals from static or moving images and accurately determining the locations of their body joints. Currently, HPE finds significant applications in medicine, particularly in assisting with non-invasive diagnosis and remote interactions between doctors and patients. In the medical field, precise joint estimation is essential for analyzing functional rehabilitation and identifying physical deficiencies in patients.

However, one of the challenges in pose estimation is the cost and controlled environment required for capturing high-quality images. To overcome this, there is a need for adapting and optimizing pose estimation methods to be used on low-cost devices like mobile terminals. This adaptation would allow patients to contribute their own image data, enabling remote or autonomous assistance. Developing metrics and techniques to ensure accurate pose estimation on affordable devices would not only reduce costs but also enhance accessibility and the potential for remote medical assistance.

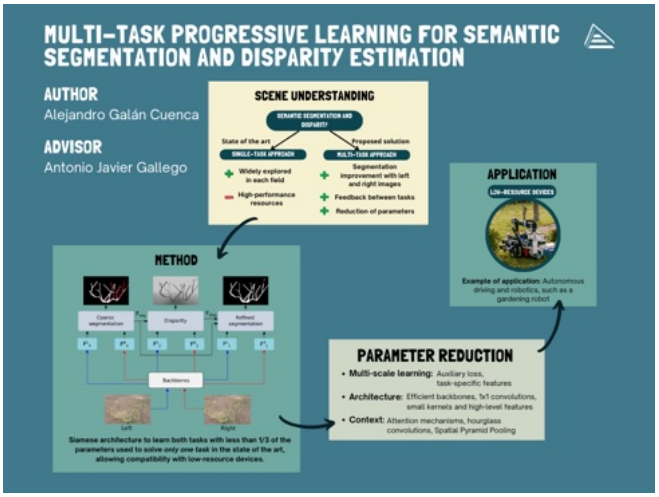
Pablo García Ruiz

University of Córdoba, Spain

Accurate object tracking is an important task in applications such as augmented/virtual reality and tracking of surgical instruments. We are gonna present a novel approach for object pose estimation using a dodecahedron with pentagonal fiducial markers. We propose a pentagonal marker that fits better in the dodecahedron figure. Our proposal improves marker detectability and enhances pose estimation with a novel pose refinement algorithm. Our experiments show the system's performance and the refinement algorithm's efficacy under different configurations.

Alejandro Galán Cuenca

University of Alicante, Spain



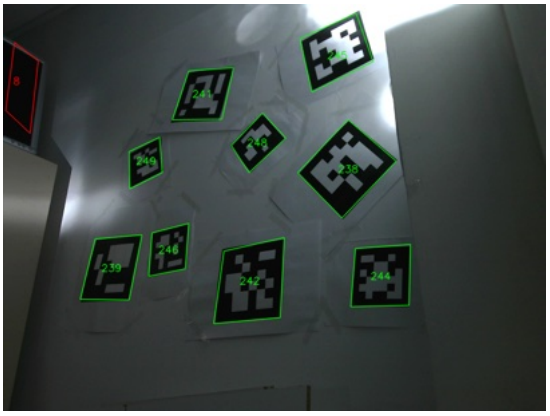
Scene understanding is an important area in robotics and autonomous driving. To accomplish these tasks, the 3D structures in the scene have to be inferred to know what the objects and their locations are. To this end, semantic segmentation and disparity estimation networks are typically used, but running them individually is inefficient since they require high-performance resources. A possible solution is to learn both tasks together using a multi-task approach. Some current methods address this problem by learning semantic segmentation and monocular depth together. However, monocular depth estimation from single images is an ill-posed problem. A better solution is to estimate the disparity between two images and take advantage of this additional information to improve the segmentation. This work proposes an efficient multi-task method that jointly learns disparity and semantic segmentation. The method extracts task-specific features and shares information progressively in a multi-scale fashion, achieving state-of-the-art results for joint segmentation and disparity estimation on the Cityscapes, TrimBot Garden, and S-ROSeS datasets, using only 1/3 of the parameters of previous approaches.

María Isabel Jiménez Velasco

University of Córdoba, Spain

Automatic analysis of human behaviour in video incorporating both RGB and 2D-3D pose information along with Deep Learning techniques.

Rafael Berral-Soler
University of Córdoba, Spain



Detection of fiducial markers in challenging lighting conditions can be useful in fields such as industry, medicine, or any other setting in which lighting cannot be controlled (e.g., outdoor environments or indoors with poor lighting). However, explicitly dealing with such conditions has not been done before. Hence, we propose DeepArUco, a deep learning-based framework that aims to detect ArUco markers in lighting conditions where the classical ArUco implementation fails. The system is built around Convolutional Neural Networks, performing the job of detecting and decoding ArUco markers. A method to generate synthetic data to train the networks is also proposed. Furthermore, a real-life dataset of ArUco markers in challenging lighting conditions is introduced and used to evaluate our system, which will be made publicly available alongside the implementation.

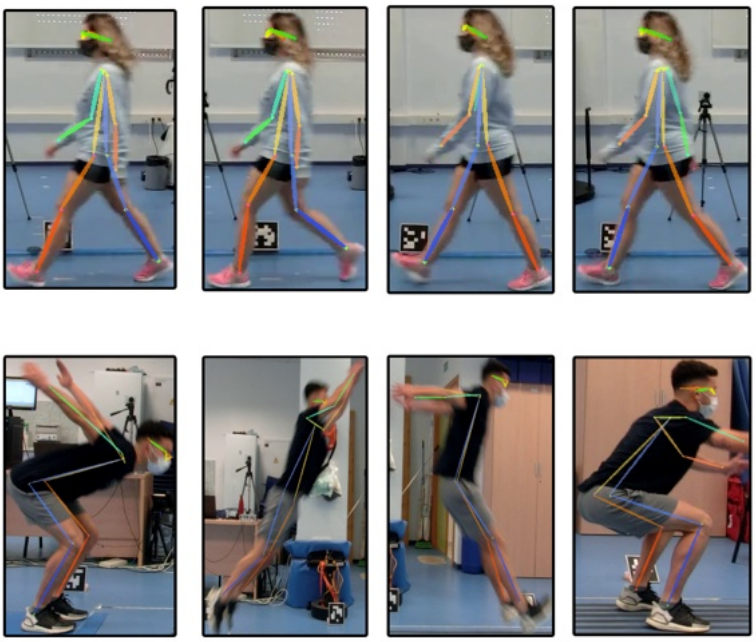
Antonio Ríos-Vila
University of Alicante, Spain

End-to-end Optical Music Recognition (OMR) is a field of research that focuses on transcribing intricate documents, including full-page, polyphonic, and composed music and lyric scores. Traditional techniques have long relied on multi-stage pipelines to tackle such complex transcriptions, focusing on Layout Analysis and staff-wise recognition. However, these approaches exhibit notable limitations.

The goal of this PhD study is to investigate the drawbacks associated with existing state-of-the-art methods for transcribing these complex music documents. Our research also aims to develop and explore end-to-end single-model approaches that can effectively transcribe these challenging documents while achieving satisfactory results. By addressing these limitations, our work contributes to shift OMR towards new projects that were recently considered out of reach.

Jorge Zafra-Palma

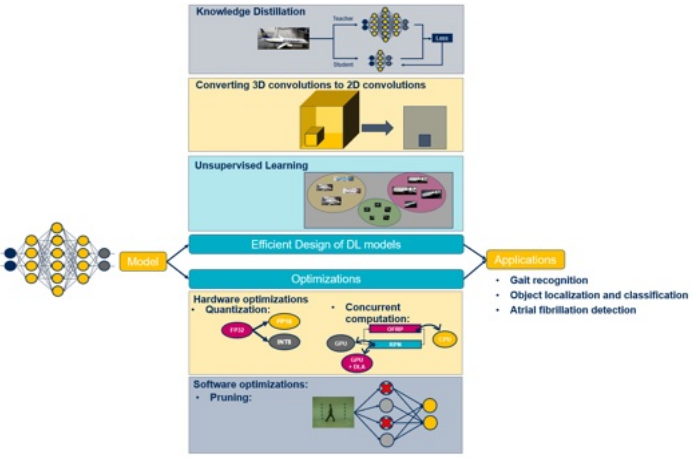
University of Córdoba, Spain



Human Pose Estimation (HPE) in Computer Vision aims to determine the joint positions in images or videos. Methods are classified based on the sensor, image or video input, inference method, and 2D or 3D information. HPE is used in medicine and sports for non-invasive diagnostics and assessments, as sports performance reflects health. Applying technological solutions in sports medicine, such as health and performance monitoring and improvement, is important. While specialized sensor-based solutions exist, affordable systems based on standard RGB cameras are desirable. Advancing HPE techniques in images and their application in this field is necessary.

Paula Ruiz Barroso

University of Málaga, Spain



Typically, the deployment of any deep learning (DL) application needs the use of high-performance architectures to fulfil both inference latency and energy consumption requirements. However, nowadays more applications need to be deployed in embedded platforms. Moreover, edge computing is gaining importance in the realm of Deep Learning due to the fact that powerful devices such as recent heterogeneous embedded systems have demonstrated remarkable skills for accelerating their challenging computational requirements.

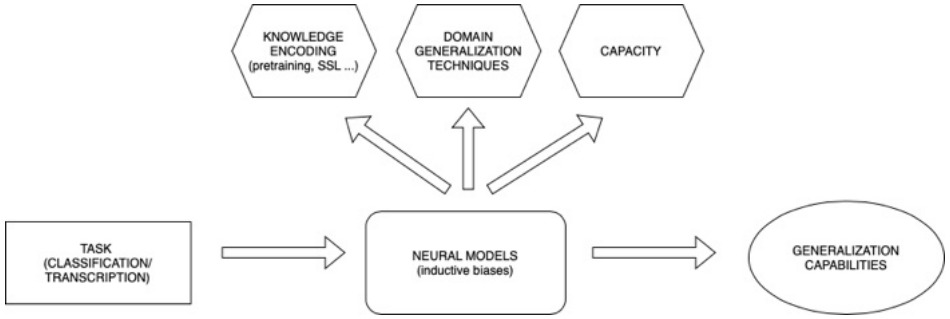
One of the main goals of my PhD is to deploy deep learning models such as CNNs and transformers in embedded devices consuming the minimum amount of energy and time possible. In this field, hardware (quantization and concurrent inference) and software optimizations (pruning) have been applied. These optimizations have demonstrated good results with state-of-the-art gait recognition models and for unsupervised object localization and classification in video.

Another topic of my PhD involves the design and training of DL models. I have proposed different unsupervised approaches. The first one is for object localization and classification at airports in real-time and the second one is with the aim of detecting atrial fibrillation in electrocardiograms. Furthermore, a knowledge distillation approach has been applied to replicate UMAP functionality.

Finally, one future objective of my PhD is the efficient deployment of a federated learning system using wearable devices.

Carlos Garrido Muñoz

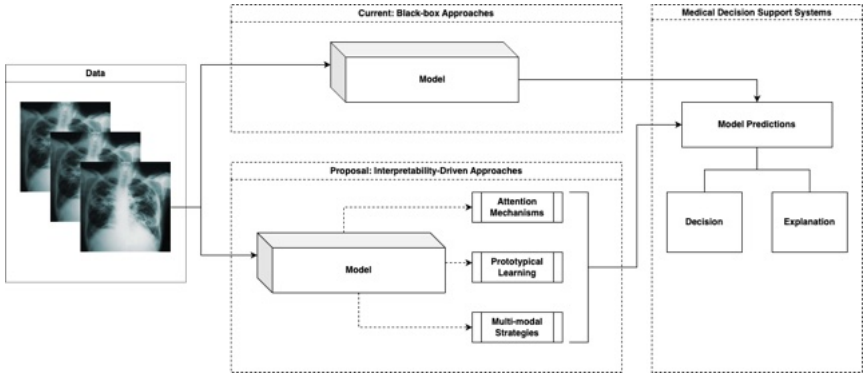
University of Alicante, Spain



My PhD focuses on inductive biases to enhance Deep Learning models by improving knowledge encoding, increasing capacity, and addressing generalization and out-of-distribution scenarios.

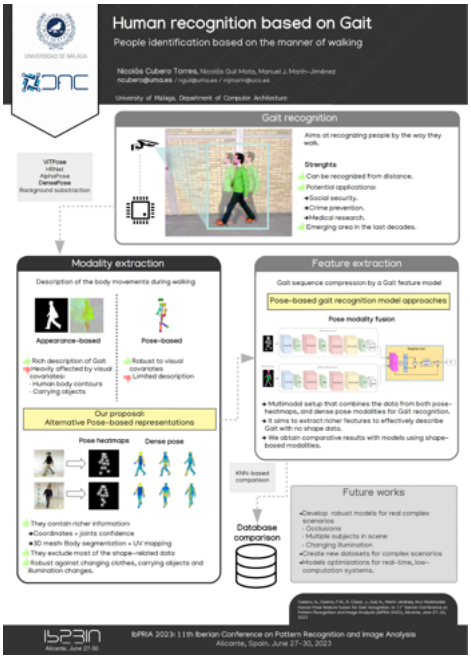
Tiago Filipe Sousa Gonçalves

Faculdade de Engenharia da Universidade do Porto and INESC TEC, Portugal



Artificial intelligence algorithms have enabled computers to learn directly from data to solve a specific problem. Regarding healthcare, this technology has powered medical decision support systems that aid health professionals in their clinical decisions. With the advent of deep learning, the performance of these algorithms has increased, especially in medical imaging cases. Due to their complexity, however, these methods are usually seen as black boxes that receive data and output a decision. This lack of transparency may jeopardise the adoption of these technologies by the medical community, which values the existence of explanations to justify a clinical decision. These issues promoted the emergence of explainable artificial intelligence, which studies the behaviour of these algorithms and tries to clarify their predictions. This project tackles the current open challenges in explainable artificial intelligence for healthcare using use cases based on medical imaging analysis. The first phase of this project focuses on an extensive study of the use and impact of attention mechanisms in deep learning algorithms for medical image analysis, given the high success of these architectures in natural language processing. Since medical decisions often rely on disparate clinical data sources, the second phase of this project employs multi-modal machine learning strategies that could leverage the information of several data types towards a clinical decision. In applications requiring high stake decisions, obtaining post-model explanations may not be enough. Therefore, in the third phase of this project, we intend to understand what the algorithms are already learning to propose quality assessment methods for the saliency maps obtained with post-model methods. Besides, our goal is to delve deeper into the field of intrinsically interpretable machine learning. Hence, we intend to propose novel interpretable algorithms for medical imaging use cases. The final deliverable of this project should be the implementation of all the algorithms into a prototype medical decision support system that can be used and validated in a real clinical context.

Nicolás Cubero
University of Málaga, Spain



Gait recognition aims at identifying people by their manner of walking. Unlike other biometrical features such as iris, or fingerprint, it can be recognized from distance without the subject cooperation. Hence, it owns very potential applications in social security, crime prevention and medical research, among others. However, there still exists several limitations that restrict its applicability to real scenarios.

One of the main limitations is related to the weaknesses of the commonly used gait modalities: On the one hand, the appearance-based modalities, such as silhouettes, gait energy images (GEI), or optical flow, contain rich descriptions of the body movement. Models based on that descriptors obtain generally good performance. However, as they are based on shape information, they contain information related to the body contours or carrying objects, that are not purely related to the gait movements (covariates). Therefore, this introduce a bias and performance may be penalized in scenarios with changing clothes or carrying objects. In contrast, the pose-based modalities such as 2D/3D skeletons, describe the position of the body joints or limbs at every instant of the gait cycle. Pose, typically arranges the 2D/3D coordinates of the body joints. Since pose does not contain appearance information it is robust to shape covariates, but the information provided is reduced. Hence, the models based on pose offers a lower performance in comparison to the shape-based models.

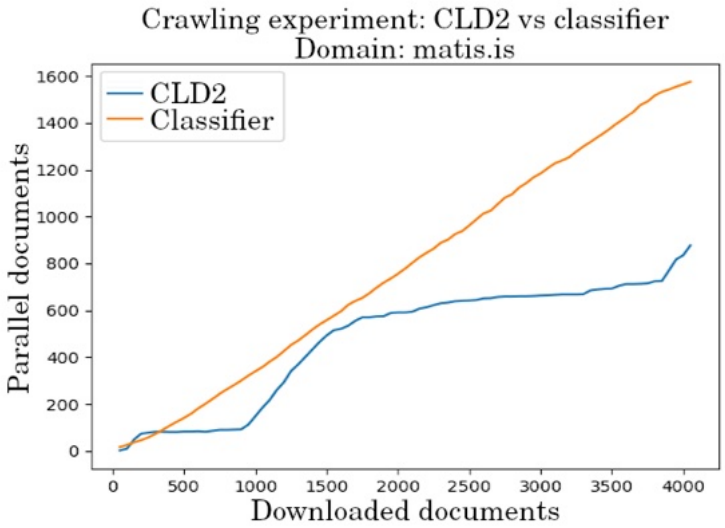
Another limitation is related to the applicability of the gait recognition models to real-life scenarios: Most of the datasets in the literature are recorded in controlled environments with a single subject in the scene, no occlusions and fixed illumination conditions. New datasets have recently been released that address gait recognition in wild environments, covering different complex scenarios with occlusions, truncations, or the presence of multiple subjects. The performance of the top state-of-the-art models on these datasets shows a significant loss of performance because of these hard scenarios.

The aim of this thesis is to propose solutions for the above limitations of gait recognition. Our current research focuses on exploring more robust gait modalities. In this line, we proposed two alternative pose-based representations: Firstly, the pose heatmaps, extracted from a keypoint extraction model. Heatmaps hold a probabilistic map per body joint that describes, not only the position of the joints but also, the confidence associated in the estimation of their location. Secondly, the dense pose, which encodes a 3D mesh that includes a body part segmentation, in addition to UV mapping. Both modalities contain richer information than the solely set of 2D/3D coordinates and do not use appearance information, making them more robust to visual covariates.

We propose a multimodal setup that effectively combines features derived from both pose representations (pose heatmaps and dense pose). Our proposal exploits multiple feature fusion strategies to combine the information from both pose representations, and obtains results comparable to the models using appearance-based modalities. In conclusion, we obtained a gait recognition model that is not dependent on appearance data, and performs optimally by only adding some minor changes to the architecture.

In future work, we plan to extend the present research by exploring more alternative modalities to describe the gait cycle. In more depth, we plan to study new frameworks and architectures to address the gait recognition in complex environments. Finally, we plan to study the optimization and deployment of models on real-time computation systems with limited computation capabilities.

Cristian García-Romero
University of Alicante, Spain



Parallel documents are documents that have the same content but are written different languages. Many Natural Language Processing tasks require these resources and their lack critically affects the accuracy and performance of these systems. The main problem is that parallel documents are scarce and difficult to obtain.

Although there are several sources from which parallel documents can be obtained (e.g. translations of books, subtitles of audiovisual content), the Internet is a great resource to be exploited due to the large amount of available data. Searching for parallel documents on the Internet has been done for more than two decades, and the largest corpora currently available has been built with data from the Internet (e.g. ParaCrawl corpus).

Retrieving parallel documents from the Internet involves a number of steps, including document downloading, preprocessing, alignment and postprocessing. While the documents are being downloaded, information from these documents which might be useful for guiding the downloading process, is ignored. This fact leads to a large number of downloaded documents being discarded because they do not contain parallel text. To avoid this, we hypothesise that parallel documents on the web are linked, and that URLs have patterns that allow parallel content to be detected. In order to discern whether two URLs will contain parallel text, we create a corpus of URLs from parallel documents using the corpus of the MaCoCu project and train a neural network that predicts whether two URLs represent a pair of documents that are mutual translation. The aim is to reduce the number of downloaded documents that will later be discarded and to retrieve a larger number of parallel documents from the Internet in less time than using established techniques. The use of URLs in order to detect parallel content is not new, but the techniques that are usually applied are intended to detect simple patterns (e.g. CCAligned corpus). Preliminary results show the usefulness of our classifier in a controlled environment.

IBERIAN 2023

BOOK OF ABSTRACTS

11th IBERIAN CONFERENCE ON PATTERN RECOGNITION AND IMAGE ANALYSIS

JUNE 27-30, 2023

UNIVERSITY OF ALICANTE, SPAIN