LOUBRYS LÁZARO ROJAS REINOSO

A real-time face analysis system for head poses classification.

Sao Paulo

2023

LOUBRYS LÁZARO ROJAS REINOSO

A real-time face analysis system for head poses classification.

Dissertation presented to the Polytechnic School of the University of São Paulo to obtain the title of Master's in Science.

Sao Paulo

2023

LOUBRYS LÁZARO ROJAS REINOSO

A real-time face analysis system for head poses classification.

Corrected Version

Dissertation presented to the Polytechnic School of the University of São Paulo to obtain the title of Master's in Science.

Concentration area:
Computer Engineering

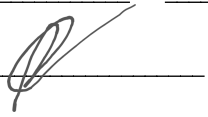
Tutor:
Prof. Graça Bressan, PhD

Sao Paulo

2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 15 de março de 2024

Assinatura do autor: 

Assinatura do orientador: _____

Catalogação-na-publicação

Rojas Reinoso, Loubrys Lázaro
A real-time face analysis system for head poses classification / L. L. Rojas Reinoso -- versão corr. -- São Paulo, 2024.
99 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Face (Análise) 2.Aprendizado Computacional 3.Processamento de Imagens I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

ACKNOWLEDGEMENTS

To my tutor Prof. Graça Bressan, Ph.D., for gave me the opportunity that changed so much more than my academic life.

To my dear friend Claudia, thank you for encouraging me and reaching out to me at the most difficult time.

To my beloved girlfriend for supporting me everyday since we are together.

To Fernando, Jose Carlos, Armando, and Patricia, a.k.a the LARC Cuban team, for receiving me as an old friend; and for teaching, helping, encouraging, and giving me advice so many times. Your contributions and teaching were very important in completing this work.

To all the professors and staff of LARC, for making so easy the adaptation process to a new country, culture, and language. It was my pleasure to meet all of you and work together.

To my sister, aunt, cousin, and the rest of the family and friends that I left in Cuba. Despite the time without seeing each other and the distance, I always felt loved and supported.

ABSTRACT

Head pose estimation attracts significant interest from the research community by the wide range of applications that rely on or are enhanced by a good head pose estimation system, including face recognition, liveness detection, facial animation, and more. Research has shown that face-related systems generally perform well in controlled scenarios, achieving excellent results for full and nearly frontal faces. However, the performance of face-related systems typically declines significantly for profile faces, presenting a major challenge.

We propose a multi-task, training-free face analysis system designed to classify face poses. After receiving an image, the system could be capable of detecting faces and their landmarks, estimating the head pose based on the human head's degrees of freedom, and calculating the angles using the Perspective-n-Point (PnP) problem, a solid alternative for those seeking a balance between robustness and speed. Some of the more robust methods often rely on deep learning solutions but they are generally unsuitable for real-time applications or require high-performance computers. After, estimating the angles, a supervised ML model has been trained to determinate faces as either frontal or non-frontal. The system will provide users with real-time feedback, such as visual cues, to help them correct their head position.

The Mean Absolute Error (MAE) obtained was 4.24, guarantying real-time running speed in low-power hardware. To evaluate the machine learning model trained, were implemented three protocols, yielding Protocol 2 the best results, with zero false positives, making it an excellent choice for real-world applications. Additionally, we tested the proposal in a proof of concept for accounts opening and verification. The system acted as an image preprocessor, allowing only images that meet specific requirements— frontal faces. We measured the impact of the proposal in other face analysis systems; the performance of face recognition and liveness detection systems improved significantly, increasing from 90% to 96% and from 62% to 87%, respectively.

Keywords: face analysis, head pose estimation (HPE), machine learning, support vector machine (SVM)

RESUMO

A estimativa de pose da cabeça é uma área de pesquisa ativa com uma ampla gama de aplicações, incluindo reconhecimento facial, detecção de vivacidade, animação facial e muito mais. As pesquisas mostram que sistemas relacionados a rostos geralmente funcionam bem em cenários controlados, obtendo excelentes resultados com rostos completos e quase frontais. No entanto, desafios surgem ao lidar com perfis.

Propomos um sistema de análise facial multitarefa, sem treinamento, projetado para classificar poses de rosto. Após receber uma imagem, o sistema pode ser capaz de detectar rostos e seus pontos de referência, estimar a pose do rosto através do grau de liberdade da cabeça humana, calculando os ângulos resolvendo o problema de Perspective-n-Point (PnP), uma sólida alternativa para aqueles que buscam um equilíbrio entre robustez e velocidade. Alguns dos métodos mais robustos geralmente dependem de soluções de aprendizado profundo, mas geralmente são inadequados para aplicações em tempo real ou exigem computadores de alto desempenho. Após estimar os ângulos, um modelo de ML supervisionado foi treinado para determinar rostos como frontais ou não frontais. O sistema fornecerá feedback em tempo real para os usuários corrigirem a pose da cabeça.

O Erro Absoluto Médio (MAE) obtido foi de 4,24, garantindo velocidade de execução em tempo real em hardware de baixa potência. Para avaliar o modelo de aprendizado de máquina treinado, foram implementados três protocolos, com o Protocolo 2 apresentando os melhores resultados, com zeros falsos positivos, tornando-o uma excelente escolha para aplicações no mundo real. Além disso, testamos a proposta em um protótipo de conceito para abertura e verificação de contas. O sistema funcionou como pré-processamento de imagens, permitindo apenas imagens que atendem a requisitos específicos - rostos frontais. Medimos o impacto da proposta em outros sistemas de análise facial; o desempenho dos sistemas de reconhecimento facial e detecção de vivacidade melhorou significativamente, aumentando de 90% para 96% e de 62% para 87%, respectivamente.

Palavras-chave: análise facial, estimativa de pose de cabeça (HPE), aprendizado de máquina, máquina de suporte vetorial (SVM).

"The face is the first thing we see, and it is often the last thing we forget."

Oliver Wendell Holmes, Sr.

LIST OF FIGURES

TABLES

ACRONYMS

FAS - facial analysis systems

FPS - frame per second

DOF - degrees of freedom

BB - bounding box

LBP - Local Binary Patterns

MB-LBP - Multi-block Local Binary Patterns

HOG - Histograms of Oriented Gradients (HOG)

SURF - Speeded-Up Robust Features

CF - Channel features

AFW - Annotated Face in-the-Wild

FDDB - Face Detection Dataset and Benchmark

CNN - Convolutional Neural Networks

DCNN- Deep CNNs

DPM - Deformable Parts-based Model

AI - artificial intelligence

ML - Machine Learning

R-CNN - Region CNN

RPNs - Region Proposal Networks

FNNs - feed-forward neural networks

HPE - head pose estimation

MTL - multi-task learning

RF - random forest

SUMMARY

CHAPTER 1

1.1. Introduction

This chapter will be present the foundation of the research proposal developed. It will introduce the context in which the research was conducted, analyzing the current importance of facial analysis systems (FAS). Furthermore, it´s going to explain the motivation behind the developing of a head pose estimator for head pose classification and/or to enhance the performance of other facial related systems. Hypotheses and a series of objectives have been formulated to quantify the impact of the mentioned research.

Additionally, this chapter will address other aspects such as the expected contributions, research methodology to be employed, the proposed itinerary, and the final structure of the work.

1.2. Research Context

With technological advances supported by better internet connections and greater internet availability, many online services have emerged to simplify users' lives. FAS are now in high demand and they´ve got here to stay. Facial analysis systems encompass a broad category of technologies and methods aimed at analyzing and understanding human faces (KHALIL *et al.*, 2020). These systems have simplified authentication processes, provided personalized experiences, enabled emotion analysis, social media interactions, supported health monitoring, and improved security measures.

This revolution has impacted various high-level security systems, including registration processes, payment apps, and bank operations. Biometric technologies for identifying individuals using physical characteristics have emerged as security tools, with face being preferred due to its less intrusive nature compared to other biometric technologies (KIM *et al.*, 2017).

The analysis of faces has posed significant challenges in computer vision and has been actively researched for its impact and contribution to society (RANJAN;

PATEL; CHELLAPPA, 2019). Research has shown that face-related systems typically perform well in constrained scenarios, yielding good results with full-frontal faces (WILLIAMS PONTIN, 2007). However, challenges arise when dealing with profiles. To overcome this limitation, various techniques analyze head pose to frontally align the face or create face classifiers that work with different face poses (KIM; OH; KIM, 2016). These solutions may introduce face distortions, artifacts or require more complex approaches.

Despite the progress made in recent years in face detection, estimating head position remains a challenging task. The research community's interest in head pose estimation is mainly driven by numerous applications that require or benefit from a reliable head pose estimation system, including face recognition, liveness detection, facial animation, human-computer interaction, people behavior understanding, virtual reality, driver assistance in the automotive field, and more (BORGHI *et al.*, 2020)

Head pose estimation involves determining the orientation and position of a person's head in relation to a reference frame or camera (DROUARD *et al.*, 2015, 2017). It is considered a part of the broader umbrella of facial analysis systems, aiming to estimate the rotation and translation of the head in three-dimensional space. **While head pose estimators are standalone systems, they can also serve as pre-processing steps for other facial analysis tasks**, for example: accurate head pose estimation can enhance the performance of face recognition systems by aligning faces to a canonical pose for better matching or verification.

In recent years, several approaches to estimating head pose have been developed, broadly categorized into geometric-based methods, appearance-based methods, and hybrid methods that combine both geometric and appearance cues.

1.3. Motivations

Due to the importance of face analysis systems in our daily lives, especially in high-level security applications as mentioned earlier, developing a system to analyze faces, making controlled the uncontrolled environments, is the wisest choice. Thanks to the remarkable progress in facial detection in recent years, faces can now be detected very quickly under various lighting conditions, with different poses, and even in the presence of occlusions.

Focusing on frontal faces, face analysis systems can establish a strong foundation for subsequent analysis steps, enhancing overall accuracy and reliability in various systems. **To streamline and automate, as much as possible, high-level security-related processes, without compromising security levels, a minimal user collaboration will be considered.** We aim for a minimally intrusive systems, striving to be as unobtrusive as possible, rather than a non-intrusive system. To improve the general users' experience, real-time performance capability will be crucial to improve the user experience. In computer vision, real time means that the system can process images and video as they are captured without significant delay. Systems that are used to detect objects in a video stream could use the reference of processing images at a near rate of 24 frames per second (FPS) minimum.

The central challenge and primary objective of this research was to determine, in real-time, whether a face is in the desired position, and if not, promptly guide the user to the desired face position. This strength could make the developed system capable of being used as an individual head pose estimation system or also used, as a pre-processing step in other FAS.

Our focus was to classify faces in frontal or not, given their relevance to other face analysis systems. To achieve the proposed face analysis system, was studied various face analysis systems with different tasks and applications. The intended system was designed to analyze faces and determine whether a face is in a frontal position or not. To assess the face pose, the degrees of freedom (DOF) of the human head was extracted, providing information about the head's position in relation to the camera. Additionally, a supervised machine learning algorithm was trained to quickly classify and identify, using the extracted DOF, whether the face is in the desired position.

1.4. Objectives

The main goal of this work was to develop a face analysis system for face pose classification capable of real-time working in low-power computers.

To achieve this goal, it was necessary to:

- Study the state-of-art of face analysis systems.

- Analyze the main alternatives for head pose estimation.
- Develop a supervised machine learning model for face pose classification.
- Evaluate the classification performance of the proposed system.
- Analyze the impact of the proposal in other face-related systems.

1.5. Expected Contributions

Two main contributions were expected from this research. Those contributions were:

- The implementation of Head Pose Estimator for face pose classification, capable of working in real-time in low-power computers.
- Demonstrate the performance improving of the developed Head Pose Classification system in other face-related systems.

1.6. Research Methodology

With the intention of obtaining the desired results, the methodology used walked behind of the following ideas:

- Define the research proposal scenario.
- A literature review of the main approaches in the face analysis system umbrella.
- A study of face detection and head pose estimation approaches in order to identify useful algorithms for our research proposal.
- Implementation of different algorithms for face detection with the intention of identifying a robust algorithm capable of working in real-time.
- Integration of a face detector with algorithms for detecting head movement degrees to develop a head pose estimator, capable of classifying frontal faces and work in real-time.

- Study, select, train and test supervised machine learning algorithms for frontal face classification.

- Performing simulations at different stages of the proposal system to identify the different characteristics and challenges of the analyzed scenario.

- Analysis of the results obtained by evaluating the performance of the system developed in the scenario studied. Analysis of compliance with expected contributions.

1.7. Anatomy of the Work

The following research is composed by six chapters, in addition to the references used as a theoretical basis for the development of the proposed method. A brief description of the main aspects discussed in each chapter will be given below:

Chapter 1 introduced the foundation of the academic research and described the research situation and the real work scenario for the approach developed. Also explained the motivations, objectives, and the methodology implemented to develop the exploration.

Chapter 2 will be analyzing the main works, methodologies and contributions related to the tasks under the FAS umbrella.

Chapter 3 will be presenting the computational resources used to develop the proposed approach. The hardware, programming language and libraries, will be mentioned in this chapter.

Chapter 4 will be presenting and detailing the approach, its architecture, and the relationship between every steps.

Chapter 5 will analyze and evaluate the performance of the proposed method implementation in different environments.

Chapter 6 will comment the main conclusions, considerations and will talk about future works.

References will be presenting the main state-of-art academic articles consulted for this work.

CHAPTER 2

This chapter will explore the key definitions, challenges, applications, methodologies, and algorithms that have been presented in state-of-art literature for the research and development of facial analysis systems. We have selected the most significant papers that form the foundation of this field and have influenced the development of our proposal.

In this chapter, we will examine both classic and state-of-the-art papers related to face analysis and the systems falling under the umbrella of facial analysis systems (FAS). We will analyze the characteristics of these systems, the challenges they address, the solutions they propose, and their overall performance. This analysis of existing works will aid in identifying, selecting, and implementing robust methods for facial analysis, with a focus on achieving real-time capability. At the conclusion of the chapter, we will summarize the main topics that have been analyzed.

1.1. Face Analysis Systems (FAS) Overview

The presence of facial images in our digital lives is a direct result of the widespread adoption of cameras, social media platforms, and smart devices. These technologies have seamlessly integrated into our daily routines, capturing and sharing countless images that include faces. Cameras are now embedded in various devices, from smartphones and laptops to security cameras and wearable gadgets. These devices and platforms encourage photo sharing and tagging, further amplifying the prevalence of facial images.

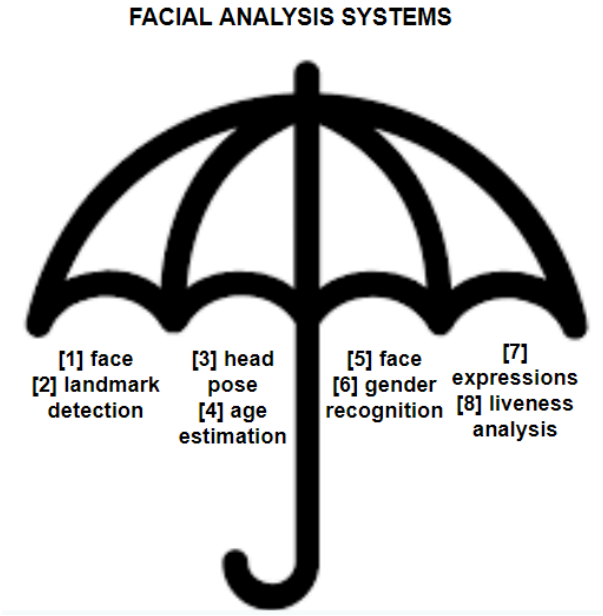
The context and relevance of face analysis systems lie at the intersection of computer vision, artificial intelligence, and human interaction. Faces are not only essential for human identification but also serve as rich sources of information about emotions, expressions, age, gender, identity, and even health conditions. This has led to the emergence of face analysis systems as a crucial component of various applications across diverse industries.

Face analysis systems are a critical subset of computer vision that focuses on extracting meaningful information from facial images and videos. By highlighting the importance of facial analysis, its relevance across domains such as security

(biometric authentication and surveillance), human-computer interaction (engagement and attention levels), healthcare (medical diagnostics, research, and patient care), and entertainment (emotion recognition and augmented reality), it encompasses countless real-world applications (ADHIKARI *et al.*, 2021; T *et al.*, 2023).

Facial image analysis encompasses a range of face perception tasks (see Figure 2-1), including but not limited to face detection, landmark detection, face recognition, head pose estimation, and more (BALAKRISHNAN *et al.*, 2021; BORGHI *et al.*, 2020a; BUOLAMWINI; GEBRU, 2018; KHALIL *et al.*, 2020; RANJAN; PATEL; CHELLAPPA, 2019a; WILLIAMS PONTIN, 2007)

Figure 2-1 Tasks under the Facial Analysis Systems denomination.

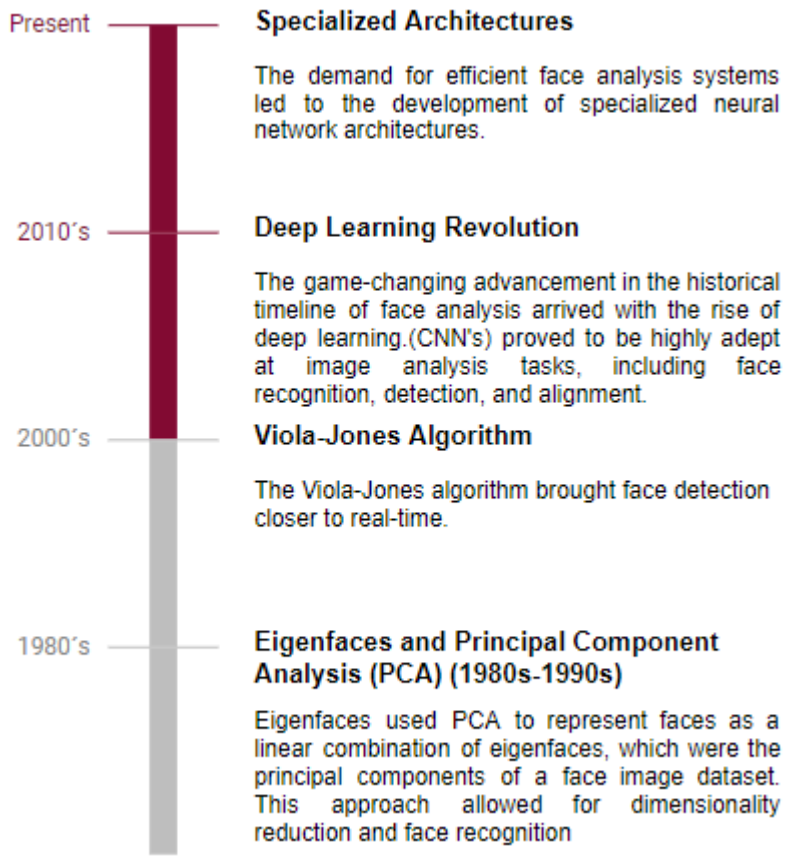


Source: The author (2023)

The historical development of face analysis systems has been marked by a remarkable transition from traditional methods to the groundbreaking surge of deep learning-based approaches (see Figure 2-2). This evolution has revolutionized the accuracy, speed, and versatility of face analysis technologies, ushering in a new era of capabilities and applications. In the early stages of face analysis, traditional methods primarily relied on hand-crafted features, such as edge detection, texture analysis, and geometric measurements, to identify and analyze facial attributes. These approaches, while pioneering at the time, were limited by their sensitivity to lighting conditions, pose variations, and expressions. The turn of the millennium saw the emergence of the Viola-Jones algorithm (VIOLA; JONES, 2003, 2001), a pivotal

moment in face analysis. This algorithm introduced Haar-like features and a cascade of classifiers, enabling rapid face detection. The game-changing advancement in the historical timeline of face analysis arrived with the rise of deep learning. Convolutional Neural Networks (CNNs) proved to be highly adept at image analysis tasks, including face recognition, detection, and alignment. Deep learning further solidified its dominance with the introduction of face embeddings and triplet loss functions.

Figure 2-2 Face analysis approaches timeline evolution.



Source: The author (2023)

Despite the advancements, face analysis systems keep facing challenges. Variations in lighting, pose, expressions, and occlusions can impact accuracy (MAGDIN; BENKO; KOPRDA, 2019; SAVCHENKO, 2022). Ethical considerations surrounding privacy, bias, and potential misuse have raised important questions about responsible deployment. The widespread adoption of face analysis systems prompts discussions about ethical considerations, privacy concerns, and potential

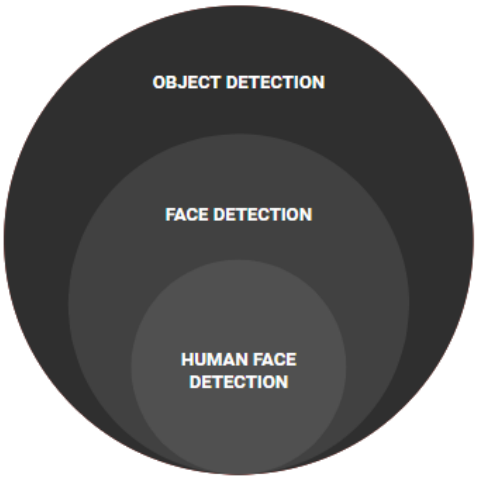
biases. Striking a balance between innovation and ethical use is crucial to ensure that these systems contribute positively to society.

1.2. Face Detection Overview

Over the last two decades, face detection in images has been a highly researched topic in the field of computer vision (RANJAN; PATEL; CHELLAPPA, 2019b; TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018). Object detection involves identifying instances of objects from specific classes (such as people, cars, buildings, or faces) in images or videos (DANG; SHARMA, 2017). Face detection is a subdomain of object detection (as illustrated in Figure 2-3), which allows to recognize faces among other objects. And human face detection differentiates people faces from animal faces.

Accurate face detection plays a crucial role in a wide spectrum of applications, ranging from security systems to personalized marketing. It underpins facial recognition, emotion analysis, and human-computer interaction, showcasing its significance in these real-world applications.

Figure 2-3 Human face detection as part of object detection domain.



Source: The author (2023).

Due to its high level of usability in commercial applications and products, face detection is currently considered a mature technology, but not yet saturated (MATHIAS *et al.*, 2014). It is regarded as the cornerstone of all facial analysis algorithms (ZHANG *et al.*, 2018). Numerous applications require face detection as

their initial step for subsequent processing, including face alignment, face recognition, head pose tracking, and many others (ZHANG; ZHANG, 2010).

According to (LIAO; JAIN; LI, 2015), the primary objective of a face detector is to locate and identify faces in images and, if found, return the dimensions (bounding boxes, BB) of each face. Performing this task remains challenging in images taken in uncontrolled environments. Various challenges, as depicted in Figure 2-4, such as variations in scale, the absence of structural components (occlusion), significant pose variations, facial expressions, varying lighting conditions, low image resolution, and others, can degrade the performance of facial detectors and other facial analysis systems (LI *et al.*, 2015; LIAO; JAIN; LI, 2015; RANJAN; PATEL; CHELLAPPA, 2015; ZHANG; ZHANG, 2010; ZHANG *et al.*, 2018) .

Figure 2-4 Face detection challenges.

Source: The author (2023).

In the early days, training a facial detector was a time-consuming task, primarily due to the limited computational resources available. However, recent research in face detection has seen significant advancements thanks to the widespread availability of image datasets captured in unrestricted conditions and the computational power of modern computers (TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018). Consequently, achieving real-time speed while maintaining high performance and robustness remains one of the remaining challenges for practical face detectors on CPU devices (ZHANG *et al.*, 2018)

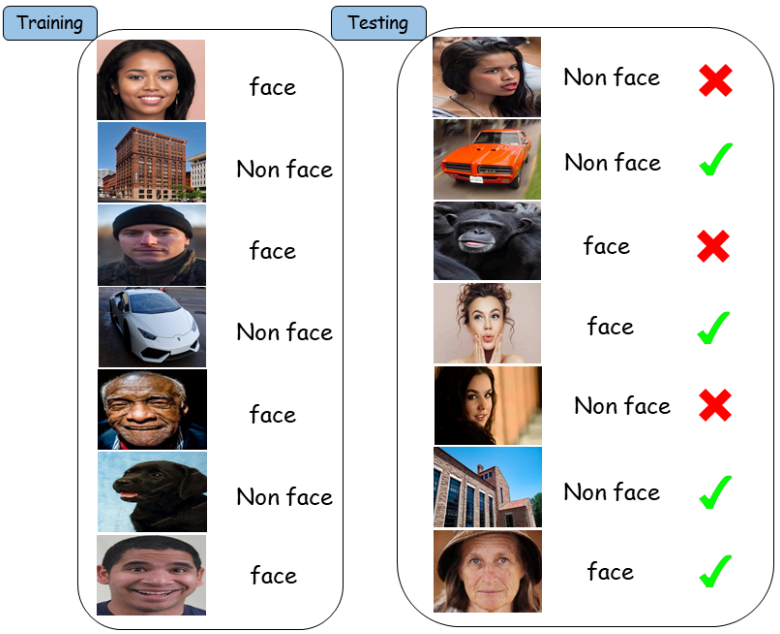
The historical evolution of face detection, transitioning from rule-based methods such as Haar cascades to the emergence of deep learning, highlights the dynamic nature of the field. This historical perspective not only sheds light on the progress made in overcoming previous limitations but also paves the way for exploring more advanced methods.

1.2.1. Features Extraction

As previously mentioned, the rapid growth in storage and processing power has led to the dominance of appearance-based methods in advancing face detection. Consequently, two key issues have emerged in this process: determining which features to extract and selecting the appropriate learning algorithm (ZHANG; ZHANG, 2010). Feature extraction serves the purpose of describing facial appearance (ZAFEIRIOU; ZHANG; ZHANG, 2015) or distinguishing facial patterns from other patterns (DANG; SHARMA, 2017).

In the context of face detection, hand-crafted features refer to manually designed and engineered characteristics of facial images used to differentiate and identify faces by their appearances (see Figure 2-5). Before the emergence of deep learning and neural networks, traditional face detection heavily relied on these manually crafted features to detect faces in images. These features are carefully chosen based on an understanding of facial structures, patterns, and variations. Some common hand-crafted features used in traditional face detection methods include Haar-like Features, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and others.

Figure 2-5 Appearance-based methods. The appearance and its features differentiate human face form other objects. Left: the definition of what is a face for training; right: the final classification of the method. In green: images classified correctly; in red: images misclassified.

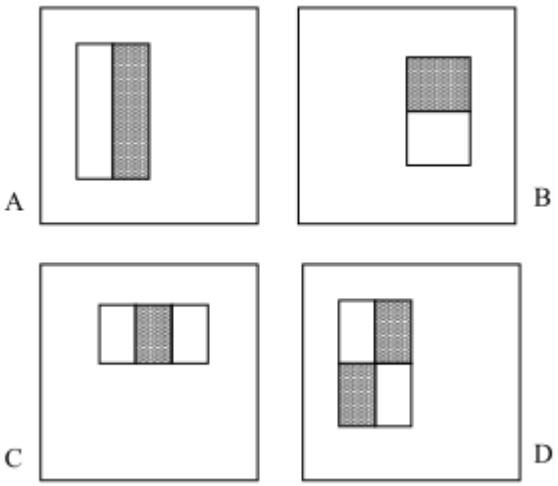


Source: The author (2023).

The seminal work presented by (VIOLA; JONES, 2001) is one of the best-know object detection algorithms since it allowed the use of detectors feasible for real-world applications. Their object detection classifies images based on the value of simple features. They used three kinds of Haar-like features: two, three and four-rectangle features (see Figure 2-6). Rectangle features are sensitive to the presence of edges, bars, and other image structures, and provide a rich image representation which supports effective learning. These features can be computed very rapidly using a representation for the image, called integral image (Equation 1). The integral image at location x, y contains the sum of the pixels above and to the left of x, y, inclusive: where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i\left(x', y'\right) \quad (1)$$

Figure 2-6 Haar-like features. Two, three and four rectangles' features presented by Viola-Jones. Sensitive features to the presence of edges, bars and others image structures.



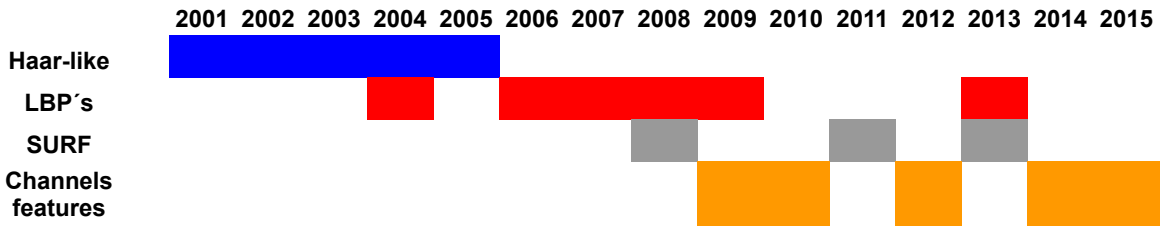
Source: (VIOLA; JONES, 2001)

LIENHART; MAYDT, (2002) introduced a set of 45° rotated Haar-like features, extending what was presented by (VIOLA; JONES, 2001).This set of features can also be computed rapidly and improves upon its predecessor by reducing the false alarm rate by 10%. In (VIOLA; JONES, 2003) the authors introduced a fourth type of rectangular filter. This addition was necessary because the three types of filters they had previously introduced were insufficient for accurately detecting non-upright and non-frontal faces. The new filter operates in the same way as the others but enhances the detector's capability to handle various face poses encountered in real images while maintaining its original speed advantage.

In (MITA; KANEKO; HORI, 2005), a joint Haar-like feature approach for face detection was introduced, based on the co-occurrence of three Haar-like features. Feature co-occurrence captures structural similarities within the face class, enabling the construction of an effective classifier. The joint Haar-like detector exhibited better performance than the Viola and Jones detectors, demonstrating robustness against changes in illumination and the addition of noise.

The algorithm presented in (VIOLA; JONES, 2001) enabled real-time face detection but had limited efficiency due to the challenges mentioned earlier. Over the years (as shown in Figure 2-7), new forms of feature extraction have been developed in pursuit of a more robust and effective face detector.

Figure 2-7 Evolution of features extraction techniques.

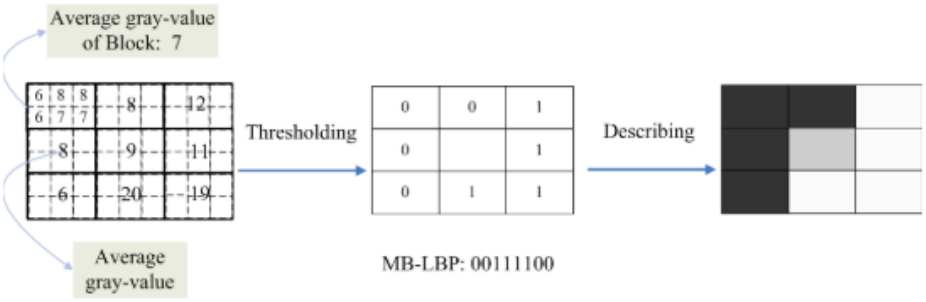


Source: The author (2023).

In their work, (ZHANG *et al.*, 2007) introduced a novel set of distinctive rectangular features known as Multi-block Local Binary Patterns (MB-LBP). MB-LBP encodes intensities using the local binary pattern operator, allowing it to capture more information about image structure. This approach exhibits more distinctive performance compared to traditional Haar-like features and the original LBP features.

In another study, (WANG; HAN; YAN, 2009) combined LBP and Histograms of Oriented Gradients (HOG). While this combination was not used for face detection but for human detection, which falls under the object detection umbrella, it proved effective. This method captures edge and local shape information from HOG and texture information from LBP. The primary advantage of this approach is its ability to handle cases of partial occlusion.

Figure 2-8 Multi-block LBP feature for image representation. MB-LBP features encode rectangular regions intensities by local binary patterns. The resulting binary patterns can capture and describe diverse image structures.



Source: (ZHANG et al., 2007).

Speeded-Up Robust Features (SURF) was initially introduced in (BAY *et al.*, 2008). SURF is a scale and rotation-invariant detector and descriptor. J. Li and Zhang used SURF features to describe local patches. First, they applied it for face detection

in (LI; ZHANG; WANG, 2011), and later, for general object detection in (LI; ZHANG, 2013). In both cases, they achieved results comparable to the state-of-the-art in terms of performance and speed.

Channel features (CF) (MATHIAS *et al.*, 2014) have proven to be robust for object detection and similar tasks. Integral channel features (ICF) involve registering multiple image channels computed using linear and nonlinear transformations of the input image. These resulting features can be efficiently computed using the integral image technique (DOLLÁR *et al.*, 2009). Depending on the features used, such as gradient histograms, color, gradient magnitude, and others, this approach can work in real-time.

(BIN YANG *et al.*, 2014) noted that different combinations of channel types have a significant impact on performance. For face detection, they found that the color channel in LUV space, along with the gradient magnitude channel and gradient histograms channels in RGB space, produced the best results. Combining pixel values from all subsampled channels provided a powerful representation capacity, resulting in a multi-view face detector with performance comparable to the state-of-the-art in challenging datasets.

Another notable paper in the literature utilizes channel features effectively for face detection. In this case, (MATHIAS *et al.*, 2014) used only the gradient magnitude channel and LUV color space channel to build the Head-Hunter face detector. They employed ICF and a rigid template learning algorithm, achieving excellent results for various classes, particularly when ample training data is available. They also developed a Deformable Parts Model (DPM) face detector for comparison. Both detectors achieved top performance on face detection tasks, even in challenging databases such as Pascal Faces (FENG *et al.*, 2022; NADA *et al.*, 2018), Annotated Face in-the-Wild (AFW) (KÖSTINGER *et al.*, 2011a), and Face Detection Dataset and Benchmark (FDDB) (JAIN; LEARNED-MILLER, 2010).

1.2.2. Learning algorithms

As mentioned earlier, the main keys that allowed the development of facial detectors in recent years were the extraction of the most robust features that define

the human face and, the development of learning algorithms that learn quickly and efficient the features extracted.

(ZAFEIRIOU; ZHANG; ZHANG, 2015) organized the face detection learning algorithms into two large groups:

- algorithms based on rigid templates. In this group of rigid templates are algorithms from the Boosting´s family, which includes the algorithm presented by Viola and Jones and variations. This group also includes algorithms based on Convolutional Neural Networks (CNN) and Deep CNNs (DCNN), since their learning architectures are researched and used in face detection, after demonstrating great performance in multi-object class detection.
- algorithms that learn and apply Deformable Parts-based Model (DPM). In this other group are the family of algorithms that can combine face detection with the location of parts of the face to model potential deformations between those parts of the face.

Methods such as Boosting and DCNN have problems handling unseen views, this is relieved due to a large amount of data in different views that exist, but it directly influences the amount of data and times to train the algorithm, also highlighting the high computational cost it demands. However, DPM-based methodologies show better generalization in the face of novel views, which requires a smaller amount of data when training (ZAFEIRIOU; ZHANG; ZHANG, 2015).

1.2.2.1. Deformable Parts-based Model (DPM)

DPM´s, also known as pictorial structures modeling, is a widely adopted choice for developing generic object detectors (ZAFEIRIOU; ZHANG; ZHANG, 2015). Part-based face representation has been explored in previous works, yielding impressive results (RANJAN; PATEL; CHELLAPPA, 2015; YAN *et al.*, 2013, 2014). However, one challenge of DPMs is their high computational complexity. Researchers have taken various steps to address this challenge, optimizing the approach to reduce complexity without significantly compromising performance.

In (YAN *et al.*, 2013) an effective DPM for face detection in the wild was presented, involving the extraction and calculation of HOG features. This approach achieved near real-time performance for frontal faces and surpassed the accuracy of both academic and commercial detectors in the FDDB benchmark. The outstanding performance on this challenging dataset highlighted the algorithm's ability to handle one of the most difficult issues in face detection: occlusion.

(YAN *et al.*, 2014) also trained a DPM for object detection, including face detection testing. They used look-up tables for HOG feature extraction, and the algorithm required parallelization to achieve real-time performance. However, the approach outperformed state-of-the-art methods in both pedestrian and face detection tasks.

(GIRSHICK *et al.*, 2015) were among the first to combine DPM and CNN models to enhance face detection performance. They replaced the standard image features used in DPM models with a learned function extractor. This hybrid model, known as Deep Pyramid DPM, outperformed traditional DPM models based on HOG features and R-CNN detection systems.

In (RANJAN; PATEL; CHELLAPPA, 2015), the authors also substituted HOG features with deep pyramidal features and introduced a normalization layer to deep CNN to bridge the training-testing gap in DPM models. This method successfully detected faces at various scales, poses, and under occlusion in uncontrolled environments. While it delivered good performance, it came with a high computational cost, taking over 20 seconds to process a face. While this limitation can be mitigated with the use of GPUs, it still presents challenges for real-time applications.

1.2.2.2. Rigid templates

To improve the performance of detectors, researchers have focused on enhancing the learning algorithms (ZAFEIRIOU; ZHANG; ZHANG, 2015). Various variants of boosting algorithms exist, including Discrete Adaboost (MITÉRAN *et al.*, 2005; NOCK *et al.*, 2006), Real AdaBoost (ÇOŞKUN; ÇETIN, 2022; SHAHRAKI; ABBASI; HAUGEN, 2020), and Gentle AdaBoost (SHAHRAKI; ABBASI; HAUGEN, 2020). While all these variants share similar computational complexity from a

classification perspective, they differ in their learning algorithms (LIENHART; KURANOV; PISAREVSKY, 2003).

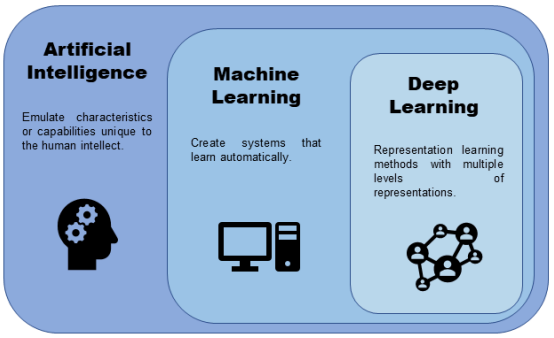
(VIOLA; JONES, 2001) used the AdaBoost learning algorithm in their seminal work to boost the classification performance of a weak learning algorithm. The key insight is that generalization performance is related to the margins of the examples, and AdaBoost rapidly achieves large margins.

A new boosting algorithm called FloatBoost was introduced in (LI *et al.*, 2002; LI; ZHANG, 2004; ZHANG *et al.*, 2002) to overcome limitations in AdaBoost learning. FloatBoost learns the component detectors in the pyramid and achieves similar or higher classification accuracy than AdaBoost with a smaller number of weak classifiers. The authors claimed to have created the world's first real-time face detection system.

In their papers, (BRUBAKER *et al.*, 2008; LIENHART; KURANOV; PISAREVSKY, 2003) provided simple comparisons of the three boosting algorithms. The former chose RealBoost for the overall cascade learning strategy, while the latter preferred GentleBoost, as it outperformed the other two boosting algorithms and required, on average, fewer features.

Deep Learning is making significant strides in solving problems that have long challenged the artificial intelligence (AI) community (LECUN; BENGIO; HINTON, 2015). The term "Deep Learning" refers to a set of Machine Learning (ML), see Figure 2-9, algorithms that utilize complex architectures with multiple levels to extract multiple abstractions from their input data (TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018).

Figure 2-9 Relationship between Artificial Intelligence, Machine Learning and Deep Learning.



Source: The author (2023).

Deep Learning approaches continue to evolve, continually improving the state-of-the-art year after year. In (JIANG; LEARNED-MILLER, 2017), the authors used a Faster Region CNN (R-CNN) for face detection, a method initially proposed and proven effective by (REN *et al.*, 2015) for object detection. The adoption of Region Proposal Networks (RPNs) for efficient and accurate region proposal generation has led to state-of-the-art performance. RPNs are fully convolutional networks that simultaneously predict object bounds and objectivity scores at each position. They are trained end-to-end to generate high-quality region proposals, which are then used by Fast R-CNN for detection.

CNN´s were described as "black-box" non-linear classifiers by (GIRSHICK *et al.*, 2015; REN *et al.*, 2015; TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018) defined CNNs as a sub-class of feed-forward neural networks (FNNs). The primary objective of FNNs is to learn a parameterized function that models the relationship between its input data x and the desired output y. Empirical evidence has shown that adding more layers between the input and the output in FNNs enhances the network's performance, reigniting scientific interest in Deep Learning techniques.

In (TRIANTAFYLLIDOU; NOUSI; TEFAS, 2018) a lightweight CNN for face detection was proposed. It is simple enough to be deployed on mobile applications and accurate enough to be comparable to state-of-the-art face detection methods. To train this lightweight deep network without compromising its efficiency, a new training method of progressive positive and hard negative sample mining was introduced. The model begins by learning from easier positive examples, such as frontal faces, and progressively moves on to harder positive examples. This progressive learning

approach mimics the learning process of humans when tackling difficult tasks, significantly improving training speed and accuracy.

Balancing speed and accuracy are a crucial aspect of designing effective face detection systems. It is essential to understand the unique demands of the system and comprehend the implications of one-stage and two-stage frameworks for optimizing the FAS´s performance:

- One-Stage Detection Frameworks: These are celebrated for their simplicity and efficiency. These models directly predict object classes and bounding box coordinates in a single pass through the network. This streamlined approach ensures faster inference times, making them well-suited for real-time applications like video surveillance or interactive systems.
- Two-Stage Detection Frameworks: These divide the task into distinct stages: region proposal and object classification. The first stage generates a pool of potential regions likely to contain a face, and the second stage classifies these proposed regions and refines the bounding box predictions. This method enhances accuracy as the model performs a more refined analysis of each region, capturing intricate facial features. While two-stage detectors generally offer heightened precision, they can be computationally heavier due to the two-pass process, leading to relatively slower inference times compared to one-stage models.

1.3. Head Pose Estimation Overview

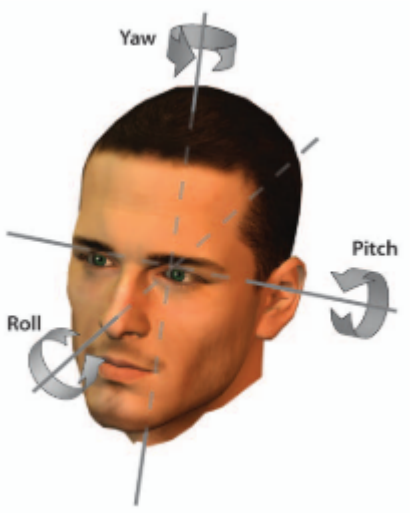
Despite the progress made in recent years in face detection, estimating the position of the head remains a challenging task (SHAO *et al.*, 2020). Computer vision researchers have long explored the topic of head pose estimation, yet an approach capable of estimating head pose in all situations remains elusive (CZUPRYŃSKI; STRUPCZEWSKI, 2014). Nowadays, numerous applications rely on, and demand robust head pose estimation for future analyses, such as facial behavior analysis (analyzing motivations and intentions, focus of attention) and human-computer interaction (face verification, face tracking, or face recognition) (BALTRUSAITIS;

ROBINSON; MORENCY, 2016; LIU *et al.*, 2016; PAPAZOV; MARKS; JONES, 2015; VALLE *et al.*, 2016). (KHAN *et al.*, 2021; SHAO *et al.*, 2020) also assert that head pose estimation is an essential preprocessing step for accurately inferring many facial attributes, such as age, gender, race, identity, or facial expressions.

In the computational vision field, head pose estimation can be defined as the orientation of a person's head relative to the camera's view (DROUARD et al., 2015, 2017; MURPHY-CHUTORIAN; TRIVEDI, 2009). It is typically a highly complex task due to the myriad of factors that can alter the image, including camera distortion, projective geometry, resolution, changes in people's biological appearance (expressions, age, race, gender), and environmental factors (occlusions, lighting, background) (DROUARD et al., 2017; PAPAZOV; MARKS; JONES, 2015).

Head pose estimation is closely related to the degrees of freedom (DOF) of the human head (KHAN *et al.*, 2021; MURPHY-CHUTORIAN; TRIVEDI, 2009). These angles (pitch, yaw, roll) (see Figure 2-10) describe the orientation with respect to a head-centered frame (DROUARD et al., 2017), with yaw rotation being the most informative angle due to its significant impact on the expressive parts of the face (VALLE *et al.*, 2016). Measuring the DOF of the human head helps determine when a face is frontal, slightly inclined, or in profile to the camera (KHAN et al., 2021).

Figure 2-10 The degree of freedom (DOF) of the human head.



Source: (MURPHY-CHUTORIAN; TRIVEDI, 2009)

HPE is usually linked to visual gaze estimation (ability to detect the direction and focus of a person´s eyes) (MURPHY-CHUTORIAN; TRIVEDI, 2009). Sometimes, this association is right, since the eyes usually move in the same direction of the head, but there are times they are not in the same direction, when this occurs the direction of the head is more important, this effect could be observed in Wollaston illusion (see Figure 2-11) (WOLLASTON, 1824).

Figure 2-11 Wollaston illusion: Although the eyes are the same in both images, the perceived gaze direction is dictated by the orientation of the head.

Source: (WOLLASTON, 1824).

1.3.1. Head Pose Estimation Applications

Head pose carries rich information. The position of a person's head can convey a wealth of information, with changes in head posture indicating emotions such as agreement, disagreement, understanding, confusion, or surprise. As mentioned earlier, HPE has numerous application areas in human-computer systems, including human social behavior analysis, driving safety, surveillance, and targeted advertising, among others. The use of HPE approaches benefits many other face-related systems, such as face verification, face recognition, and even liveness detection.

(FRISCHHOLZ; WERNER, 2003) presented an HPE technique based on a single camera input designed to prevent replay attacks. HPE verifies whether the user's head movements match the random instructions given to them. After challenge-response tests, the user is guided to look directly into the camera, and the final image is acquired for face recognition.

Face recognition is an application where the computer either classifies human identity based on facial features (face identification) or verifies whether two images belong to the same person (face verification). It typically involves two steps: enrollment and matching. In the enrollment process, facial features are extracted from the image to obtain the subject's signature, which is then compared in the matching process. When combined with head pose analysis, the accuracy of recognition and facial expression analysis becomes more robust.

Frontal faces have proven to be valuable for various face-related systems, including face and gender recognition, age estimation, entertainment filters, and more. While frontal face recognition is considered a solved problem (XU *et al.*, 2018), it's worth noting that important papers in the literature still perform face frontalization to achieve pose-invariant face recognition (KAKADIARIS *et al.*, 2017; TAIGMAN *et al.*, 2014; XU *et al.*, 2018). Experiments in these papers demonstrate that as faces move toward extreme poses, the similarity index with the frontal face (face verification) decreases. Typically, papers use extreme poses for testing and frontal poses for the matching process.

Face frontalization was introduced to address non-frontal faces, either through face alignment or generating a part of the face with less visibility, using data from the most visible part of the face. Frontalization is intended to facilitate better feature extraction (TAIGMAN *et al.*, 2014). However, both alignment and generation processes have drawbacks to consider alignment can introduce distortions and affect the proportions of the generated frontal face, while the generation process may create a non-real symmetrical frontal face by filling in the missing part with known data.

Several works, (BORGHI et al., 2020; MURPHY-CHUTORIAN; DOSHI; TRIVEDI, 2007; RAY; TEIZER, 2012; SCHULZ et al., 2011) also explore the benefits of HPE systems for driver attention monitoring. HPE can detect in real-time whether the driver is drowsy or distracted, such as looking at their phone instead of the road. Various applications in this area include providing contextual alert signals for pedestrians outside the driver's field of view, predicting pedestrian paths, and detecting blind spots.

Systems built on the assumption that people's attention in a scene can serve as an indicator of interesting areas and events can be employed in targeted

advertising and surveillance and safety applications. (SMITH *et al.*, 2008) proposed a method to track visual attention in wandering people, counting individuals looking at specific outdoor advertisements (targeted advertising). (SANKARANARAYANAN; CHANG; KRAHNSTOEVER, 2011) tracks the visual attention of different subjects from various fixed surveillance cameras.

1.3.2. Head Pose Estimation Approaches

Over the years, various approaches have been developed to address the task of HPE. Some of these are end-to-end systems, capable of estimating head positions from input images (BALTRUSAITIS; ROBINSON; MORENCY, 2016; RANJAN *et al.*, 2017; ZHANG *et al.*, 2015). Others require input images with detected faces (YAN *et al.*, 2016), and some employ facial landmark localization to enhance HPE results (BALTRUSAITIS; ROBINSON; MORENCY, 2016; PAPAZOV; MARKS; JONES, 2015; RANJAN *et al.*, 2017; RANJAN; PATEL; CHELLAPPA, 2019b; ZHANG *et al.*, 2015).

End-to-end systems and those combining facial landmark detection for improved HPE typically report better results than methods that handle each task separately. This performance boost is attributed to multi-task learning (MTL), an approach that enhances generalization by leveraging domain information from related tasks. In MTL, tasks are learned in parallel, with knowledge gained from each task benefiting the learning process of other tasks (CARUANA, 1997). Recent studies have demonstrated that jointly learning correlated tasks can significantly improve the performance of individual tasks in various computer vision problems. Authors argue that a unified approach may simplify the problem (CHEN *et al.*, 2014; RANJAN *et al.*, 2017; ZHU; RAMANAN, 2012).

The table (see Table 2-1) summarizes papers focused on HPE approaches. It highlights the primary methods used in recent years and their advantages, such as real-time capability and solved challenges. Articles published in the early 2000, were used to mention the main quality challenges presented in the images for testing. Over the years, the papers information about the characteristics of the images used was absorbed by the emergence and standardization of image databases for training and testing algorithms, with this, comparison between solutions became easier.

Also, was analyzed if the state-of-the-art approaches took advantage of MTL implementation.

Table 2-0-1 Head Pose Estimation Approaches state-of-art. Main capabilities and solved challenges.

Reference	Technique	Face Detection	Landmark Localization	HPE	Real-time	Solved callenges
(FANELLI et al.*, 2011)*	Random Regression Forest	-	-	X	X	Pose variation, partial occlusions, facial expressions.
(ZHU; RAMANAN, 2012)	Mixture of trees	X	X	X	X	-
(YANG et al.*, 2015)*	CNN	X	X	X	-	
(PAPAZOV; MARKS; JONES, 2015)	Triangular Surface Patch	-	X	X	X	Robust to noise, rotation and translation invariant.
(LIU et al.*, 2016)*	CNN	-	-	X	X	-
(VALLE et al.*, 2016)*	*RF*	*X*	-	*X*	*X*	*Facial expressions, ilumination, oclussions, blur*
(BALTRUSAITIS; ROBINSON; MORENCY, 2016)	*Mixture of machine learning algorithms*	*X*	*X*	*X*	*X*	-
(RANJAN et al.*, 2017)*	CNN	X	X	X	-	Extreme poses, illumination and resolution
(BORGHI et al.*, 2018)*	CNN	X	-	X	X	-
(ZHOU; GREGSON, 2020)	CNN	-	-	X	X	Extreme poses
(TRUONG; LAO; HUANG, 2020)	Multi-cameras	-	-	X	X	-

(VALLE; BUENAPOSADA; BAUMELA, 2020)	CNN	-	X	X	X	Extreme poses
(THAI et al.*, 2022)*	CNN	X	-	X	X	-
(HEMPEL; ABDELRAHMAN; AL-HAMADI, 2022)	CNN	-	-	X		-

Source: The author (2023).

Analyzing the papers presented in the table, we see that many approaches that implement MTL (CHEN *et al.*, 2014; ZHANG *et al.*, 2014; ZHU; RAMANAN, 2012) can perform in real-time, demonstrating not only good results in performance but also in speed. The approach presented by (RANJAN *et al.*, 2017) also implemented MTL, but it cannot work in real-time due to the double deep (vertical and horizontal connections) CNN used and the algorithms to generate face region proposals.

Random Forest (RF) is a simple but powerful classification tool and has been used successfully in many computer vision problems. RF is a machine learning algorithm formed by a set T decision tree, whose prediction is the combination of the outputs of all trees (VALLE *et al.*, 2016). Classification and regression trees are powerful tools capable of mapping complex input spaces into discrete or respectively continuous output spaces (FANELLI *et al.*, 2011). Fanelli et al. (2011) also said that many approaches have been used RF to solve the tasks, due to its ability to handle large datasets, high generalization problem, can be trained with a moderate amount of data and the resulting models can work in real-time (SHAO *et al.*, 2020).

Convolutional neural network (CNN) has been successfully applied to computer vision tasks such as image classification, landmark localization (ZHANG *et al.*, 2014), and HPE (BORGHI *et al.*, 2018; LIU *et al.*, 2016; RANJAN *et al.*, 2017). (YANG *et al.*, 2015; ZHANG *et al.*, 2015) used CNN to learn the mapping from the head poses and eye images to gaze directions in the camera coordinate system. Another CNN was trained on synthetic head images to estimate head poses (LIU *et al.*, 2016). (RANJAN *et al.*, 2017) developed an end-to-end system formed by a double deep CNN for face detection, landmark localization, gender recognition, and HPE tasks,

outperforming the state-of-the-art in the last task, in quite challenging datasets such as AFW and AFLW.

CNN´s have strong learning abilities for image processing and often they can achieve the desire effect. In the other hand, they require long training time and high computational cost, most of the time, needing a good GPU or even a server drive. Also, the models are getting more complicated with time, the number of layers is getting deeper and deeper. The amount of data to train these models are getting bigger and bigger.

In recent years, the method based on manifold learning has completed excellent results in pose data modeling. The view of manifold learning is that the data we can observe is mapped from a low-dimensional manifold to a high-dimensional space. One of the main applications of manifold learning is ”non-linear dimensionality reduction” (HONG *et al.*, 2019). The manifold can describe the essence of the data. (BENABDELKADER, 2010), reducing from high-dimensional space to low-dimensional space without losing information, so this mapping can input the original data and output more essential characteristics of the data.

Training techniques are more numerous than training-free techniques methods. The reason can be found in the popularity of training techniques as CNN´s, RNN´s, etc., gained in last years, that lead to higher accuracies. Both of those methods usually can test an image in more than 20 fps (to be considered real-time). Training techniques require a considerable time for training and testing the model, before the use as final product, but, in case that the environment or the characteristics of subject change, several hours of training will be again necessary to adjust the model to the new data. In the other hand, training-free techniques methods have good generalization properties, so it can be remodeled in much less time. As mentioned above, training techniques methods demand a lot more computational power, making more suitable training-free techniques for low-power and real-time computational applications. (ABATE *et al.*, 2022) made a comparison using two works, one from each technique, and it could see how a training-free technique can be better not only in time and hardware requirement, but in performance, getting pretty similar results in the AFLW2000 dataset (HASSNER *et al.*, 2015; YIN *et al.*, 2017; ZHU *et al.*, 2015).

1.4. Summary and Conclusions

Face analysis systems play a vital role in the field of computer vision by extracting valuable insights from facial images and videos. Their significance spans across various domains, including security (biometric authentication and surveillance), human-computer interaction (engagement and attention tracking), healthcare (medical diagnostics, research, and patient care), and entertainment (emotion recognition and augmented reality), among others.

Facial image analysis encompasses a wide range of tasks, such as face detection, landmark detection, face recognition, head pose estimation, and more. Face detection serves as the foundational step for many facial analysis algorithms, as it is often the first essential task in processing facial data. The Viola-Jones algorithm, introduced at the turn of the millennium, marked a significant milestone in face analysis by introducing Haar-like features and a cascade of classifiers, enabling fast face detection.

However, the field continued to evolve with the introduction of new features and learning algorithms aimed at enhancing both the speed and accuracy of face detection. The real game-changer in the history of face analysis was the advent of deep learning, which revolutionized not only face detection but also various other components of facial analysis systems.

Head pose estimation (HPE) systems have also benefited from the popularity and advancements in deep learning. Over the years, increasingly accurate HPE systems have been developed, with a wide range of applications, including face recognition, liveness detection, driver monitoring assistance, targeted advertising, and more. HPE is closely linked to the degrees of freedom (DOF) of the human head, which include pitch, yaw, and roll angles, describing the head's orientation in a head-centered frame. Measuring these DOF helps determine whether a face is frontal, slightly inclined, or in profile to the camera.

Modern methods for HPE include training techniques that have gained popularity due to their higher accuracy, although they often require significant computational power and time. These advancements collectively contribute to the

ongoing progress and importance of face analysis systems in diverse real-world applications.

CHAPTER 3

2.1. Computational Infrastructure and Libraries

The proposed system was developed in Python programming language on a machine with 64-bit Windows 11 operating system, with Intel Core i5 processor at 2.40 GHz and 16GB of RAM. The use of public computer vision libraries was necessary for easy code development, and future improvements. And, the public libraries used, will allow to other researchers, an easy tunning and implementation of the proposal.

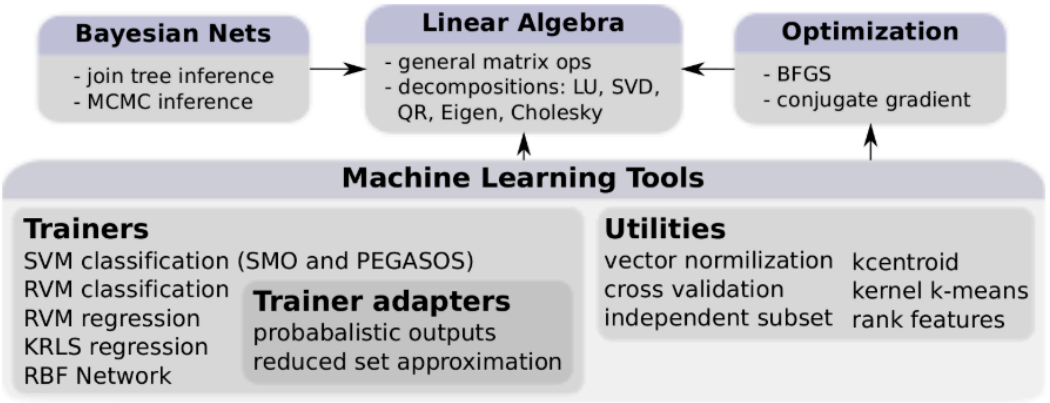
2.1.1. Computer Vision Libraries

(BALTRUSAITIS; ROBINSON; MORENCY, 2016) said: “There is a big gap between state-of-the-art algorithms and freely available toolkits”. Sometimes re-implement approaches could require a significant amount of effort due to the few available source codes or the lack of the details in the paper. Considering this, the proposed research will be developed with publicly available machine learning libraries that facilitate the use or replication of the approach.

Three of the most important and computer vision libraries will be used:

- Dlib: is a cross-platform open-source software library written. The library is intended to be useful in both research and real-world commercial projects (KING, 2009). Dlib can be implemented on images, column vectors or any form of structured data. Implementation of the algorithm is entirely different from the data on which they operate. Dlib library has a lot of components, graphical user interfaces, data structures, linear algebra, machine learning, image processing, data mining, and many other tasks (see Figure 3-1) (SHARMA; SHANMUGASUNDARAM; RAMASAMY, 2017). The use of those machine learning tools allows user to make innumerable functions such as face detection, face alignment, facial landmark detection, face recognition.

Figure 3-1 Main tasks from Dlib library. Arrows show dependencies between components.



Source: (KING, 2009).

- OpenCV: is an open-source available computer vision library. There are active interfaces for Python, Matlab, and other languages. OpenCV was designed for computational efficiency and with a strong focus on real-time applications. It contains over 500 functions that span many areas in computer vision, also include a full general-purpose machine learning library (BRADSKI; KAEHLER, 2008). It incorporates a mix of low-level image-processing functions and high-level algorithms such as face detection, pedestrian detection, feature matching, and tracking (PULLI *et al.*, 2012). The library has found use in different fields from interactive art to robotics (DRUZHKOV *et al.*, 2011). The OpenCV offers a free and easy way for people to get started in computer vision. It creates a way to grow the developer community and encourages innovation in a space where many of the algorithms and methods for computer vision systems are locked behind the corporate and R&D laboratory doors (ZELINSKY, 2009). OpenCV can be used in academic and commercial applications.
- Scikit-learn: This Machine Learning Python open-source library is quite powerful when evaluating models with few lines of code (ANTONA CORTÉS, 2017). They are simple and efficient tools for predictive data analysis, being accessible to everybody, and reusable in various contexts. It integrates a wide range of state-of-the-art machine learning algorithms for medium scale supervised and unsupervised problems

(see Table 3-1). Since it relies on the scientific Python ecosystem, it can easily be integrated into applications outside the traditional range of statistical data analysis (PEDREGOSA *et al.*, 2011; VAROQUAUX *et al.*, 2015).

Table 3-1 Performance variations between different machine learning algorithms ran in Python. Time in seconds shows faster execution for scikit-learn in the application of different estimators.

	scikit-learn	mlpy	pybrain	pymvpa	mdp	shogun
Support Vector Classification	**5.2**	9.47	17.5	11.52	40.48	5.63
Lasso (LARS)	**1.17**	105.3	-	37.35	-	-
Elastic Net	**0.52**	73.7	-	1.44	-	-
k-Nearest Neighbors	0.57	1.41	-	**0.56**	0.58	1.36
PCA (9 components)	**0.18**	-	-	8.93	0.47	0.33
k-Means (9 clusters)	1.34	0.79	⋆	-	35.75	**0.68**
License	BSD	GPL	BSD	BSD	BSD	GPL

-: Not implemented. ⋆: Does not converge within 1 hour.

Source: (PEDREGOSA *et al.*, 2011)

CHAPTER 4

3.1. Real-Time Face Analysis System Proposal

In this chapter, the real-time face position classifier will be presented and described. The author intention was to implement the proposed approach for head pose estimation as an individual alternative, and to test the approach, as a pre-processing step for filtering image, to improve the accuracy of other face-related systems.

The chapter was structured into three different sections. Firstly, the scenario in which the proposed system was implemented, is explained. Subsequently, every phase of the system that has been proposed will be described. This description included, how the system works in general, and the algorithms selected for implementation in each phase. Additionally, a mobile alternative was introduced and expounded upon. This solution was considered for testing and enhancing the user experience. Finally, the main aspects of the proposal were summarized in the concluding section of the chapter.

3.2. Scenario

In the Introduction chapter, detailed information was presented regarding the research context and motivations for this academic study. Summarizing the chapter mentioned before, we said, that over the past few years, studies and research have been seeking solutions to identify and validate users, avoiding the need of human intervention. People faces replaced digital impressions and became our main biometric key. As the time passed by, financial institutions apps, like many other applications, have begun to require photo from users during the sign-up, login, and validation processes. Considering the nature of these institutions, efficiently processing the user faces without compromising security and reliability had become a significant challenge.

Face recognition and liveness detection systems are usually implemented in high-security applications. It is widely reported in state-of-the-art papers that frontal face recognition has largely been solved, and frontal faces are consistently saved and used as reference points in the face verification matching process. Therefore, by

focusing on frontal faces, a robust foundation can be established for subsequent analysis steps, enhancing the overall accuracy and reliability across various systems. To streamline and automate as far as is achievable, maintaining security levels, a low user collaboration was considered. The aim was to opt for a less intrusive system, being as unobtrusive as possible, instead of a non-intrusive system. Furthermore, when it comes to helping or increasing the user experience, the real-time performance capability played a key role.

Automatically validating the quality of users' pictures before sending, was the most practical solution. Before users send the images, it is necessary to ensure that they met all the requirements for automatic analysis on the server. The systems automatically take the photo when the image met the requirements of lighting conditions and a correct head positioning, in this case, a frontal position relative to the camera. Aiming at improving the user experience, the system will provide the user with real-time guidance to fulfill all the image requirements quickly.

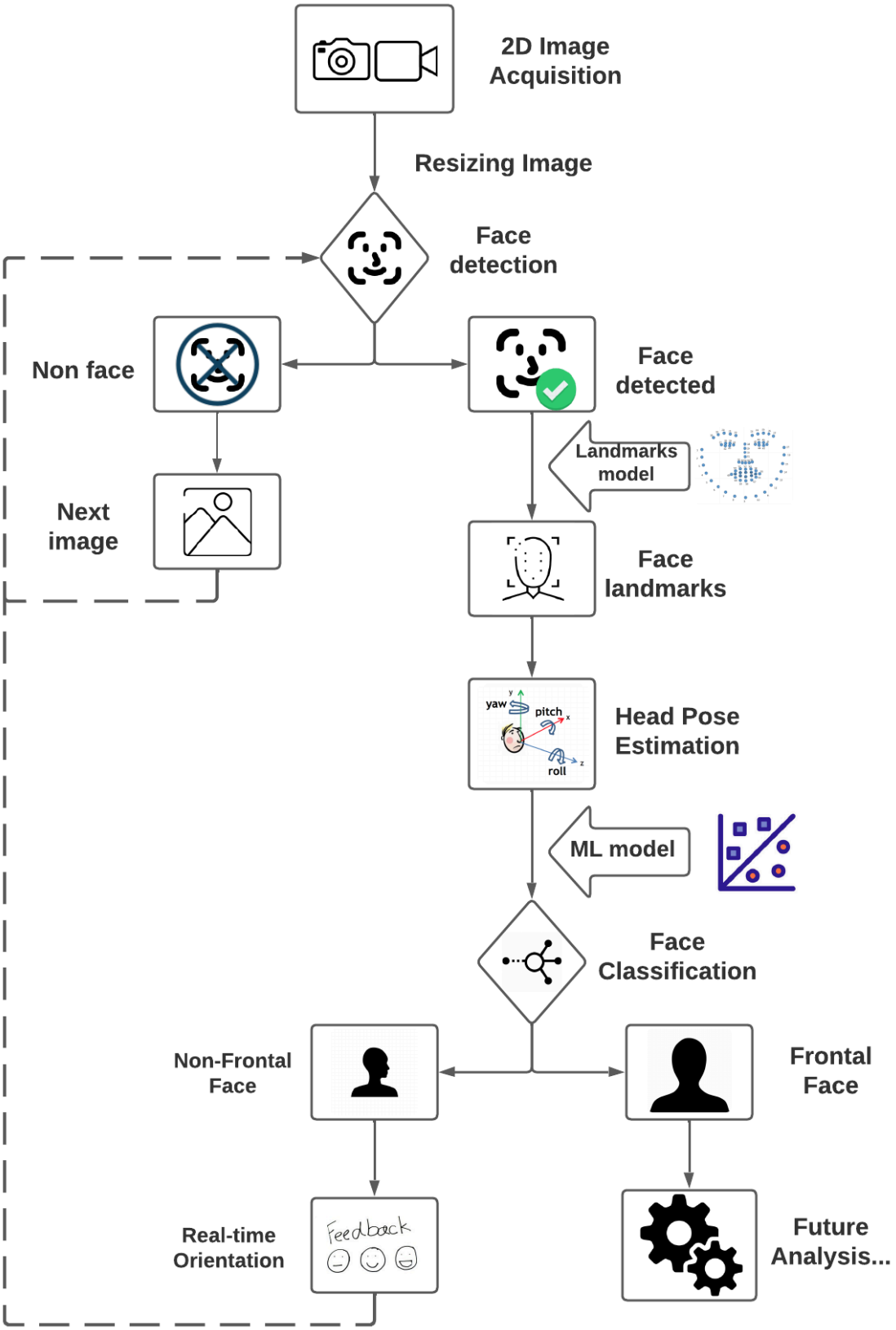
All face analysis systems have a common problem; they are all very sensitive to illumination and lighting conditions, especially when only 2D images are being used. In conjunction with this work, a desired image quality model was developed. The authors studied the impact of the image quality on face processing algorithms (ALEJANDRA PACHECO REINA *et al.*, 2021). They analyzed different distortions that affect the overall image quality and created a model for real-time image quality analysis. This model validates the values of the different distortions and determines instantly whether the quality is suitable for future processing. The combination of these solutions, image quality and head pose estimation with face pose classifying, should enhance the performance of other face analysis systems, mitigating the primary challenges of face-related systems.

3.3. Proposed Method

To handle the scenario presented above, a proposal was developed. Its main functions were estimating the face pose and to classify whether the faces in the acquired 2D image are frontal or not. The proposed approach has two main functions and works as follows: an 2D image (photo or video) is obtained by some acquisition media, could be a smartphone, tablet, webcam, IP camera, and others. Depending

on the media used the resolution could be different, so, the image size will be resized to standardize the process and increase the speed of image analysis, while maintaining an excellent quality and the aspect ratio of the original image. Next, face detection was performed by a face detector selected from a study of the field. In addition to having good precision, the detector should also work in real-time. Various face detectors were evaluated to select the best considering the balance between speed and accuracy. Then, facial landmarks were localized, to aid in calculating the euler angles; the 68-face points model was used. Euler angles or rotation degrees of freedom (DOF) of the human head, they are used to estimate the position of the head respect to the camera and capture information about the three axes of rotational movement: yaw, pitch, and roll. First function is completed here, head pose estimation. To determine whether a face was frontal or not, a supervised machine learning model was trained. Understanding and using the DOF of the head, a model can be trained to classify when a face was frontal or not. The strategical combination of euler angles and ML allowed to develop an approach capable of estimating the head pose and quickly classify the head pose estimated, extending the serviceability of this combination to several areas and applications. Summarizing, the system can extract the Euler angles that informs the head pose estimation, and with these angles a supervised ML was trained to classify the head poses in frontal or not. The ML model is easy to train and to readjust, following the presented idea, more models for different face poses classification can be trained. Real-time feedback and guidance was provided to the user to correct his face position. In case that, the face position was classified as frontal, then, the face image analyzed is ready to be processed by other FAS. The Figure 4-1 represents the entire operation of the system, as it was recently described.

Figure 4-1 Proposed system operation.



Source: The author (2023).

3.3.1. Face Detection

As previously commented, facial detection is the cornerstone of many systems that involve human-computer interaction, including HPE and face recognition. The main function is to identify the presence of faces in each image and provide the position and size information for each detected face.

To develop our system, a series of face detection algorithms will be analyzed, seeking to use a robust algorithm capable of addressing the various challenges that may arise during image capture while ensuring real-time performance. The image size will be downscaled while maintaining the aspect ratio. This measure is taken for standardization and faster image processing without compromising quality, which could affect the effectiveness of the pose estimation process. The impact of the image size on both speed and effectiveness will also be evaluated.

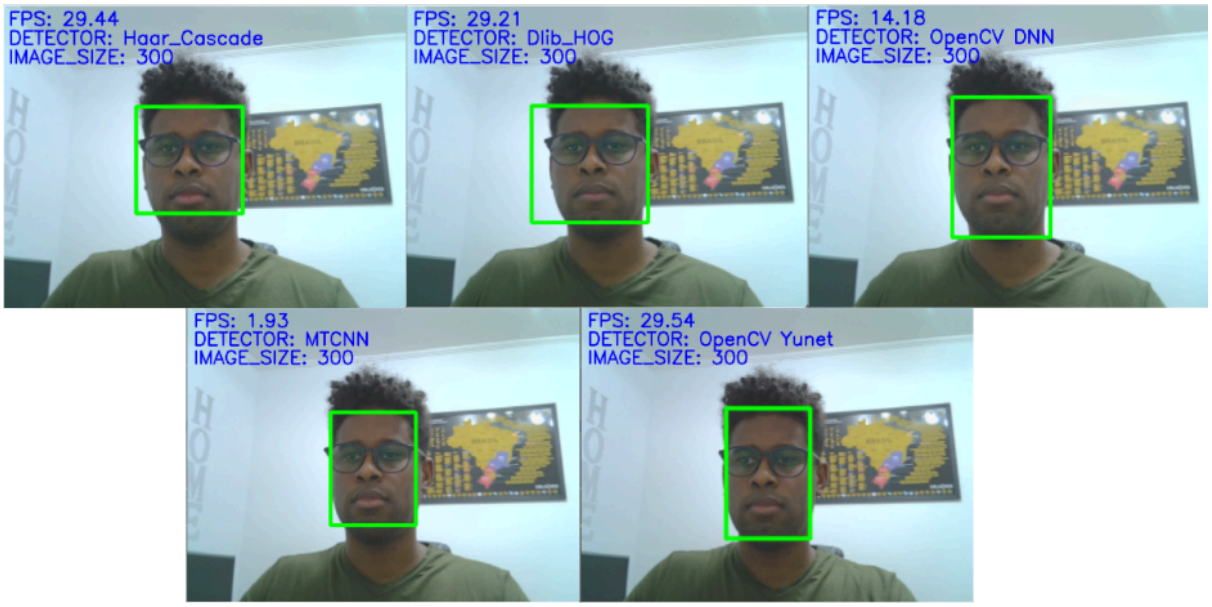
Many state-of-the-art academic papers present algorithms with excellent performances in public databases, However, at the same time, many authors do not share the repository of the algorithms. Often, replicating the results of state-of-the-art algorithms is very difficult due to the lack of details in papers. Five free and publicly available face detection Python algorithms were collected for comparison. The selected algorithms were:

- Viola and Jones Haar Cascade. The most iconic and well-known face detector. It laid the foundation for real-time face detection and can be implemented using the OpenCV Python library.
- HOG + Linear SVM face detector is a popular face detection algorithm based on the Histogram of Oriented Gradients (HOG) feature descriptor and a linear support vector machine (SVM) classifier. It is a relatively simple, yet highly effective algorithm often used in real-time face detection applications. It can be implemented using the Python library Dlib.
- OpenCV DNN is a module in OpenCV that allows you to use deep learning models for image processing tasks. It supports a variety of deep learning frameworks, including Caffe and TensorFlow. In testing sessions, using the pre-trained Caffe model works quicker than the TensorFlow ones.

- MTCNN. It is Multi-task Cascaded Convolutional Network for face detection and alignment. The MTCNN model is a powerful and effective face detection algorithm that can detect faces in a variety of conditions, including low resolution, occlusion, and pose variation. It can be implemented with mtcnn library in Python.
- Yunet. Yunet is a deep learning face detector based on the You Only Look Once (YOLO) algorithm. It is a single-stage detector, which means that it can detect faces in a single pass through the image. Although Yunet is a relatively new face detector, it has demonstrated high accuracy and efficiency. It comes readily available as a pre-trained model, making it simple to use it with OpenCV.

Knowing the face detectors for comparison, it is time to conduct some tests. Firstly, the speed and real-time capability will be analyzed. The algorithms are going to detect faces in pre-saved videos and in a live webcam video (see Figure 4-1).

Figure 4-2 Live webcam video face detection comparison.



Source: The author (2023).

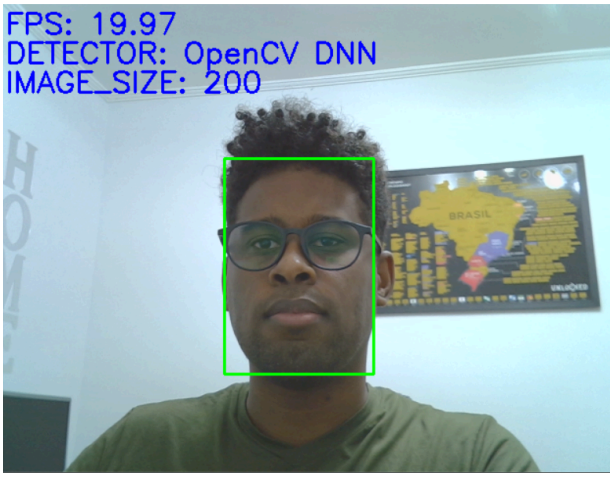
The images in the previous figure present the execution time of all detectors in a frame captured by the webcam. The original frames had a resolution of 640x480, but for speeding-up the detection in all detectors, the frames were downscaled to a width of 300, maintaining the aspect ratio. The webcam used is a Logitech C270, a

commonly used and affordable webcam. The speed is measured in FPS (frames per second). To ensure that a facial detector can work in real time, it must achieve a frame rate greater than 20 fps. Obviously, the higher FPS the better.

MTCNN face detector runs at nearly 2 fps on a 2.40 GHz CPU, making it the slowest algorithm in the list. In the original paper, the authors reported 16 fps on a 2.60 GHz CPU and 99 fps on an Nvidia GPU. Further reducing the image size has little impact on the running time. Deep Learning models, specially cascaded ones, are more suitable for GPU assistant applications.

OpenCV DNN face detector is another algorithm that could not achieve the desired minimum FPS for real-time consideration. Unlike the previous detector, if the image size is further reduced, it could operate at 20 fps (see Figure 4-2).

Figure 4-3 The impact of image size reducing analysis in OpenCV DNN face detector to reach real-time capability.



Source: The author (2023).

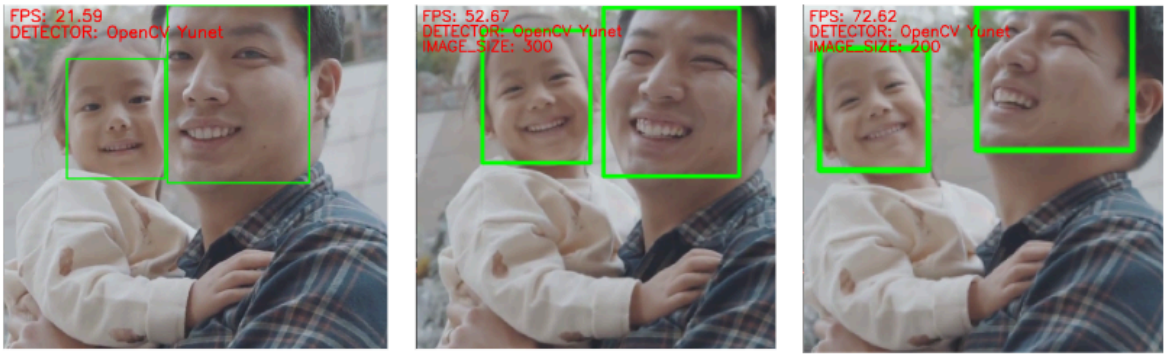
It is worth noting that the maximum speed that detectors could reach in webcam video tests was 30 fps. Face detection for live video captured by the webcam is limited to the maximum FPS that the webcam offers. The logitech C270 records videos at 30 FPS. When pre-saved videos are being analyzed, the number of FPS increases (see Figure 4-3).

Figure 4-4 Face detection speed running comparison. Left: live webcam video analysis. Right: pre-recorded video analysis.

Source: The author (2023).

Images can be downscaled to speed up the process. How much should the image be reduced? There is not a unique answer to this question. The appropriate downsizing should be determined after several tests until finding an image size that enhances processing speed without affecting system accuracy. Many aspects should be considered when selecting the right image size, including the number of FAS implemented after face detection, the main objective of the whole system, the intended environment for system usage, among others. In the figure below, the impact of downscaling the image size on the processing speed is presented.

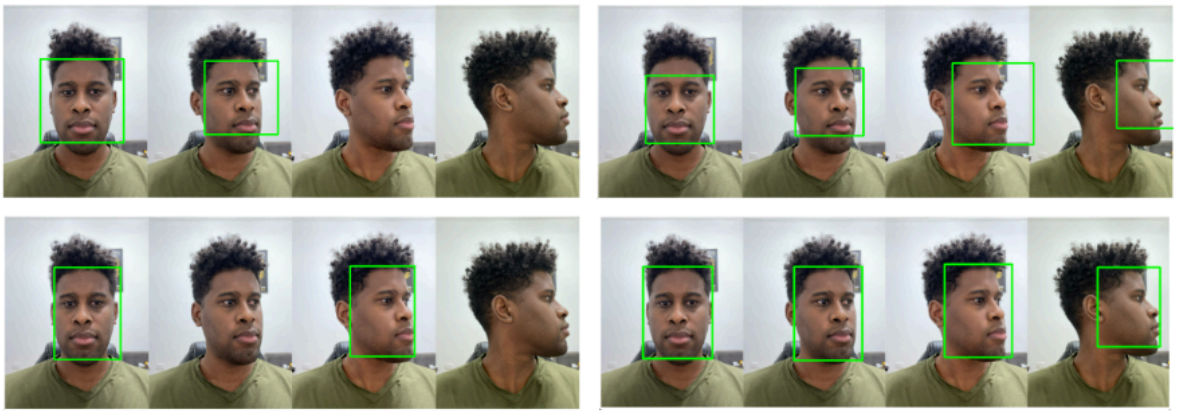
Figure 4-5 Images size vs Speed.



Source: The author (2023).

After assessing the speed and real-time capabilities, some accuracy tests were also conducted on the detectors. In the figure below, the performance of detectors in different face poses can be observed. The Haar Cascade detector, the oldest in the list, is known to have limitations in handling non-frontal faces, as discussed in papers. cv2.CascadeClassifier.detectMultiScale() function from OpenCV allows refining the detection process. However, in some cases, adjusting the image scale or the number of neighbors for each candidate rectangle does not resolve all the issues with this detector. OpenCV DNN had a negative result; it failed to detect two face poses in the images. The DNN model is considered very robust in the literature; therefore, it is difficult to understand why it could not detect face pose #2 but it did with face pose #3, which is in a more profiled position than #2 (see Figure 4-5). On the other hand, it is remarkable to see how an old and simple detector such as HOG + SVM from Dlib can handle very extreme face poses, successfully detecting all tested poses. Finally, Yunet face detector continues to be a very reliable face detector, passing all the tests comfortably.

Figure 4-6 Face poses detection comparison. Top left: Haar Cascade. Top right: HOG Dlib. Bottom left: MTCNN. Bottom right: Yunet.

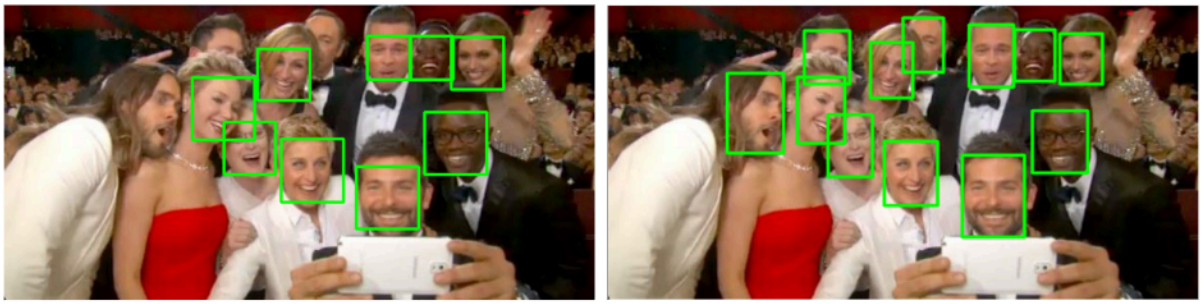


Source: The author (2023).

The final test for the detectors will take place in uncontrolled environments. So far, three detectors were disregarded: MTCNN due to its running speed, Haar Cascade due to the issues with non-frontal poses, and OpenCV DNN due to a combination of the two reasons. For this concluding round, an image was selected that presents a multitude of challenges in a single visualization. The figure below illustrates the comparison between HOG + SVM Dlib and Yunet in handling

challenges such as facial expressions, occlusion, lighting conditions, different poses, diverse skin colors and faces at different scales.

Figure 4-7 Face detection challenges in uncontrolled environments. Left: HOG Dlib. Right: Yunet.



Source: The author (2023).

After evaluating all face detectors, only one passed successfully every task. Yunet face detector seems to be the most reliable option. The detector can easily achieves over the 50 fps, making it a bonus for real-time applications. Yunet detector also detects all faces poses, even the most extreme profile ones, and in uncontrolled environments, it works very well, excellently handling numerous challenges encountered by face analysis systems.

Honorable mention for HOG + SVM Dlib face detector. Considering its longevity, around two decades, the simplicity, and its results, it can be said that it is one of the best face detectors out there, even today, the era of deep learning face analysis systems.

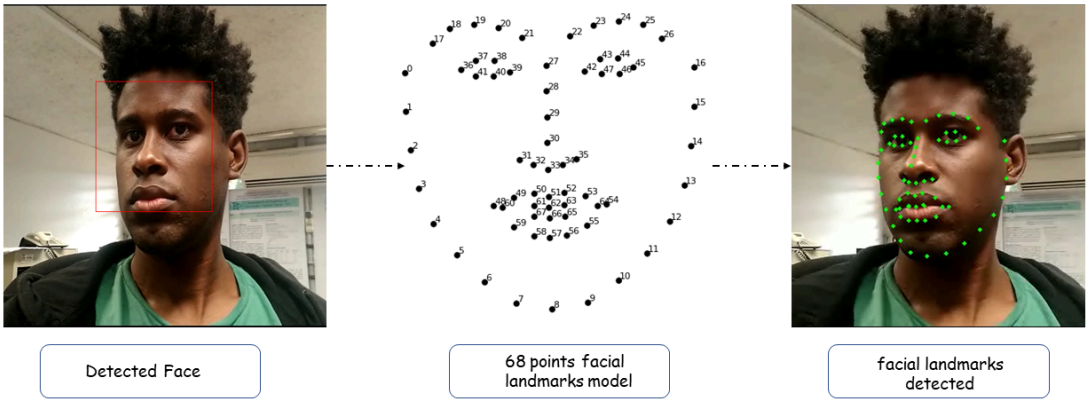
3.3.2. Facial Landmarks Localization

Face landmark detection is the process of finding points of interest of a human face in an image. It has recently seen rapid growth in the computer vision community because it has many compelling applications in other FAS, such as face alignment, HPE, face recognition, and others (KÖSTINGER et al., 2011). In this case, facial landmarks detection will be helpful in the next process of finding the DOF of human head.

In fact, there are different types of annotations for facial landmarks: 21 points (KÖSTINGER *et al.*, 2011b), 80 points (AIFANTI; PAPACHRISTOU; DELOPOULOS,

2010), 68 points (MATAS *et al.*, 1999), and others. The number of facial points detected depends on the poses of the faces and the annotation policy used during system training or in the dataset where it was implemented. The annotation used in this approach is one of the most used in the academic community. The 68 facial landmarks could be accurately estimated thanks to the predictor function found in the dlib library (see Figure 4-7).

Figure 4-8 Facial landmark detection with Dlib.



Source: The Author.

The facial landmarks will further help to calculate the DOF of the detected face. Six out of the 68 detected landmarks are used for estimating the head pose, solving the Perspective-n-Point (PnP) problem. A similar implementation of this approach was used in (BALTRUSAITIS; ROBINSON; MORENCY, 2016).

3.3.3. Head Pose Estimation

In computer vision, the pose of an object refers to its relative orientation and position with respect to a camera. This pose can be altered by either moving the object with respect to the camera or the camera with respect to the object. Various papers in the literature leverage the rotation degrees of freedom (DOF) of the human head to estimate its position (KUHNKE; OSTERMANN, 2019; RIEGLER *et al.*, 2013; RUIZ; CHONG; REHG, 2018). A 3D rigid object exhibits only two kinds of motion with respect to the camera: translation and rotation. Initially, translation does not affect the head's position with respect to the camera, but rotation does. Rotation motion encompasses three degrees of freedom, and can represented using Euler angles

(roll, pitch, and yaw) (MALLICK, 2016). Euler angles are a mathematical way to describe the orientation of a rigid body in three-dimensional space. In the case of the human head, yaw is left/right, pitch is up/down and roll is the rotation of the head when we try to put the head over the shoulders (see Figure 2-10).

The performance of head pose estimation is commonly assessed using the mean absolute error (MAE) (Kuhnke & Ostermann, 2019). MAE quantifies the difference between the measured angles and the "true" value of those angles (DROUARD *et al.*, 2017a; VENTURELLI *et al.*, 2017). The next formula defines MAE:

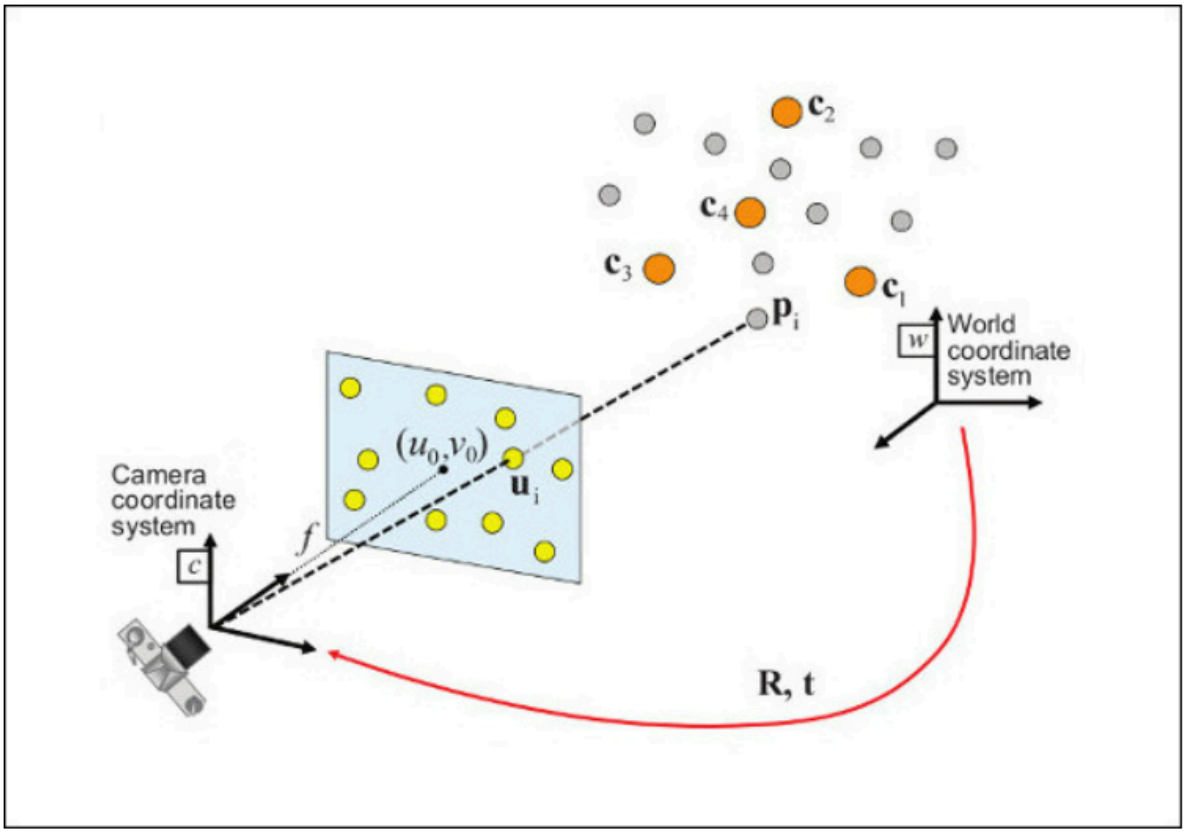
$$\frac{1}{N}\sum_{i=1}^{n}\left|x_i - x\right| \tag{2}$$

One of the challenges in pose estimation involves solving the PnP problem. In this context, the goal is to find the pose of an object when provided with a calibrated camera, and when knowing the locations of n 3D points on the object and the corresponding 2D projections in the image (MALLICK, 2016).

To calculate the Euler angles of the detected face, algorithms from two libraries described in the previous chapter will be implemented: Dlib and OpenCV. To achieve this, certain information is necessary, including 2D coordinates of specific points, 3D locations of the same points, and intrinsic parameters of the camera used. The 2D coordinates of these points are obtained through the landmark's detection process. Out of the 68 points obtained with Dlib, only 6 will be used for head pose estimation: the tip of the nose, the chin, the left corner of the left eye, the right corner of the right eye, the left corner of the mouth, and the right corner of the mouth. Additionally, 3D locations of these specific points are required. Ideally, a 3D model of the person in the photo would be necessary, but in practice, it is not. Having a 3D model of the head for every person analyzed is very difficult or requires the use of special equipment for 3D image acquisition. Fortunately, the computer vision community offers a generic 3D model of the head that will suffice the requirements. If the head pose estimation is designed for a specific scenario that demands a particular camera, the actual intrinsic parameters of that camera should be used. However, if the HPE system is intended for general purposes, the optical center can

be approximated as the center of the image, the focal length can be estimated based on the width of the image in pixels, and the absence of radial distortion can be assumed.

Figure 4-9 Perspective-n-Points problem.



Source: The Author.

To solve the Perspective-n-Point (see Figure 4-8), it is necessary to transform the points in the world coordinate system (3D head model) into 3D points in camera coordinates. To do this, the translation and rotation vectors are used. After having the points in camera coordinates, they can be projected onto the image plane using the intrinsic parameters of the camera. In summary, if the 3D head model points, 2D image plane points, and intrinsic parameters of the camera are known, the rotation and translation vectors can be calculated. To assist with this procedure, the OpenCV library has a function named ‘solvePnP’. This function returns the translation and rotation vectors. By decomposing the rotation vector, the Euler angles can be accessed. The figure below shows the results of the head pose estimation using this method.

Figure 4-10 Euler angles detection example.

Source: The author.

3.3.4. Supervised Machine Learning Model

After Euler angles detection, an SVM model will be trained for frontal face classification. Support vector machines (SVMs) are a particularly powerful and flexible class of supervised algorithms for both classification and regression. They offer very high accuracy compared to other classifiers such as logistic regression and decision trees. SVMs are used in a variety of applications such as face detection, intrusion detection, classification of emails, news articles, and web pages, classification of genes, and handwriting recognition.

SVM is an exciting algorithm, and the concepts are relatively simple. The classifier separates data points using a hyperplane with the maximum margin. SVM finds an optimal hyperplane, which helps in classifying new data points. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. That's why an SVM classifier is also known as a discriminative classifier.

During training, frontal faces will be distinguished from non-frontal ones. In this supervised learning process, the classifier will learn the behavior of the DOF (yaw, pitch, roll) in both frontal and non-frontal faces. In ideal cases, the measurements of

the three angles should be close to 0 degrees for frontal faces. The core idea of SVM is to find a maximum margin hyperplane (MMH) that effectively separates the dataset into frontal and non-frontal classes.

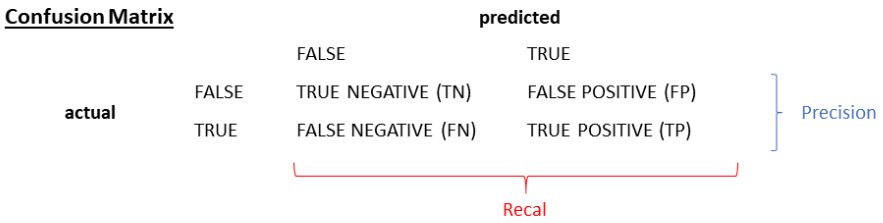
The SVM algorithm is implemented in practice using a kernel. A kernel transforms an input data space into the required format. The kernel converts a low-dimensional input space into a higher-dimensional one. In simpler terms, it converts the non-separable problems into separable problems by adding extra dimensions. Its utility is most evident in tackling non-linear separation problems. The kernel trick enables the construction of a more accurate classifier.

There are various kernel tricks for constructing accurate classifiers. Due to the characteristics of the analyzed data, the most appropriate kernel trick will be the radial basis function (RBF). RBF can map an input space into an infinite-dimensional space. This kernel is associated with hyperparameters such as regularization (C) and gamma:

- C is the penalty parameter, representing a term for misclassification or errors. This parameter guides the SVM optimization by indicating how much error is tolerable. A smaller value of C results in a small-margin hyperplane, while a larger value of C leads to a larger-margin hyperplane.
- Gamma: A higher value of gamma will perfectly fit the training dataset, which causes overfitting. Gamma=0.1 is a recommended default value. The value of gamma should be manually specified in the learning algorithm.

Tuning the parameters of the SVM can be a challenging task. One effective solution could be the implementation of GridSearchCV, a function provided by the Scikit-learn library. GridSearchCV helps determine the best parameters for the SVM. By passing various values of C, gamma, and the desired kernel into the function, it can identify the best parameter combination. In addition to training and generating an SVM model, the Scikit-learn library also allows evaluating the model through the confusion matrix and the classification report. These metrics will be used to measure the accuracy of our model.

Figure 4-11 Confusion Matrix



Source: The Author.

Precision and recall are the measures used to measure the relevance of the algorithm. Precision also is known as the positive predictive value which states the fraction of retrieved instances relevant and gives information about the false alarms. Recall gives information about the number of objects detected or relevant instances retrieved. For the perfect system the more it is close to 1 more accurate the results of the algorithm (DANG; SHARMA, 2017).

$$Recall = \frac{No.\ of\ correctly\ frontal\ face\ detected}{No.\ of\ frontal\ face\ in\ databse} \tag{3}$$

$$Precision = \frac{No.\ of\ correctly\ frontal\ face\ detected}{Total\ frontal\ face\ detected} \tag{4}$$

3.3.5. Mobile Implementation

With the rapid advancement of technology, almost every person has a cellphone at their disposal or has been in contact with one. Analyzing real-world scenarios for the application of the developed approach, it was decided to create a mobile version of the system to evaluate the user experience. Cellphones are the most popular devices for image acquisition, and virtually every major financial institution has at least one mobile app.

There are numerous brands and models available in the market today. Despite the wide variety, only two mobile operating systems exist: Android and iOS. Android is an open operating system used by almost all brands except Apple. Currently,

Android maintains its dominance in the global operating system market, holding a 70.89% market share in 2023.

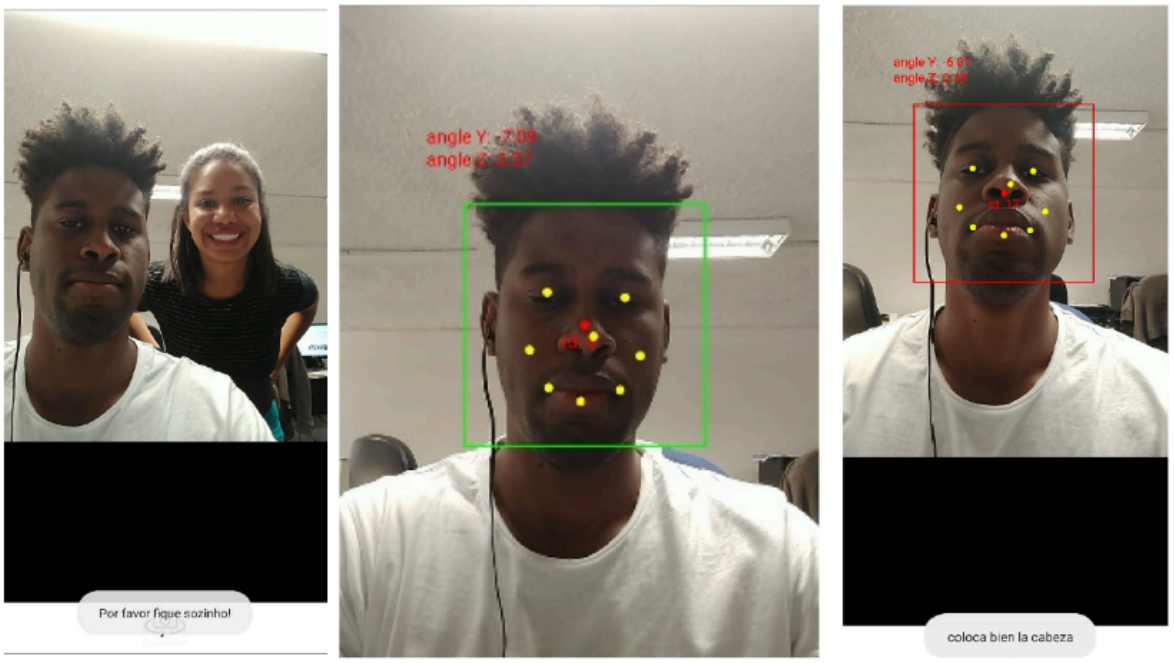
Considering the aforementioned aspects, a mobile HPE system based on the fundamentals of regular Python programming will be initiated. The mobile operating system chosen is Android due to its significant market share in the smartphone industry. To create an HPE application for Android, a different programming language is required, and Java has been selected for this purpose.

Google, the owner of android, offers a series of machine learning capabilities through its platform Firebase for image processing. Google Firebase ML Kit is a machine learning SDK that provides user-friendly APIs for integrating various machine learning features into mobile and web applications. Firebase ML Kit includes pre-trained models and APIs that enable tasks such as image labeling, object detection, face recognition, among others, without the need to create and train models from scratch. ML Kit's APIs are all designed for on-device use, supporting real-time applications like processing a live camera stream. Furthermore, this functionality is available for offline use as well.

Earlier in this chapter, it was mentioned that this work was developed in conjunction with an image quality processing model, which will complement it. Importing the image quality model from Python to Android, along with the necessary libraries for its operation, required additional adaptation work. To ensure real-time functionality and ease of installation on any Android device, the head pose estimation functionality was developed using the Firebase ML Kit. This approach resulted in a final application with both functionalities working in real time.

Google Firebase ML Kit offers a range with different computer vision applications. In the Face Detection section key tasks and capabilities, including face tracking, landmark detection, contour detection, facial expression recognition, and face orientation can be found. The Face Detection API in ML Kit provides measurements of Euler Y and Euler Z. However, it does not provide information about Euler X for detected faces. Euler Y represents the yaw angle, which indicates the rotation of the face around the vertical axis. Euler Z represents the roll angle, which indicates the rotation of the face around the horizontal axis. Notably, the pitch angle is not provided by the Face Detection API.

Figure 4-12 Head Pose Estimation with Google Firebase ML Kit.



Source: The Author.

The mobile application utilizing Firebase ML Kit (see Figure 4-11) functions effectively. It can be configured to send notifications to the user when issue arise. As shown in the images, when it detects more than one face, it refrains from estimating the pose and displays a notification recommending the user to be alone. Additionally, when the face is in a non-frontal pose, it alerts the user to look straight to the camera. In this case, the classification into frontal or non-frontal was configured manually, as the supervised machine learning model cannot be directly integrated into the mobile app. However, the angles to define frontal faces were determined by training and testing the model created in a computer environment.

3.4. Summary and Conclusions

In this chapter, the proposed FAS system was presented, which comprises a face detector, a landmark detector, and a head pose estimator. The entire approach will be capable of operating in real-time and the number of FPS will vary depending on whether the analyzed image is from a pre-saved video or a live webcam feed. The webcam analysis videos will always be limited by the characteristics of the equipment used. For processing pre-saved videos, the image size can be

downscaled in order to increase the (FPS). It is essential to experiment with different image sizes to determine the optimal size that improves processing speed without affecting the system accuracy.

Several free and public face detectors were evaluated, and the most promising option tested was the Yunet face detector, which can be implemented using OpenCV. This detector achieved a processing speed of 22 fps when applied to a video with a resolution of 720 x 720 without downscaling. However, when downscaling to 300 x 300, it achieved a remarkable speed of more than 50 fps. The detector also performed exceptionally well in accuracy tests, successfully detecting faces in various poses and handling challenging images captured. Additionally, honorable mention should be made of the HOG + SVM face detector with Dlib, which, despite its simplicity and years since its origin, outperformed some deep learning-based approaches.

After face detection, the facial landmarks were localized. To obtain 68 facial landmarks, a pre-trained model with the dlib library was used. The extracted landmarks help to estimate the head pose by calculating the euler angles.

To calculate the Euler angles and estimate the head pose, it is necessary to solve the PnP problem. In this problem, the goal is to find the pose of an object when we have a calibrated camera, and we know the locations of n 3D points on the object and their corresponding 2D projections in the image. To solve the Perspective-n-Point, it is necessary to transform the points from the world coordinates system (3D head model) into 3D points in camera coordinates. This transformation involves using translation and rotation vectors. After obtaining the points in camera coordinates, they can be projected onto the image plane using the intrinsic parameters of the camera. In summary, if the 3D head model points, the 2D image plane points, and the intrinsic parameters of the camera are known; then, the rotation and translation vectors can be calculated. To assist with this procedure, the OpenCV library has a function named after the problem: solvePnP. This function returns the translation and rotation vectors and by decomposing the rotation vectors, the Euler angles can be accessed.

Knowing the Euler angles of a face, a supervised machine learning model can be trained for image classification. Images are categorized as either frontal or

non-frontal based on their appearance. Subsequently, the model is trained to learn the Euler angles that describe a frontal face. Following this idea, the machine learning model can be trained for several user monitoring applications such as driver monitoring assistant, students' attention analysis, among others. The selected algorithm was a Support Vector Machine (SVM), known for its simplicity and effectiveness, making it a popular choice for both classification and regression tasks.

To test the user experience and applications of the system, a mobile app was developed for Android using the Java programming language. The app follows the concepts of detecting faces and estimating poses in real-time. Since cellphones are generally less powerful than computers and considering that the proposed FAS should work in conjunction with an image quality analysis model, the head pose estimation process for the mobile app was developed using native Android machine learning tools instead of the libraries used in the computer environment. With the Google Firebase ML Kit, faces and Euler angles can be detected. The real-time capability allows the app to provide guidance to the user when the image does not meet the minimum requirements in terms of face poses and lighting conditions.

CHAPTER 5

4.1. Applications and Results

In this chapter, will be presented the results of the proposed approach. The system was tested and validated in Head Pose Estimation databases, comparing with state-of-the-art solutions and evaluating the impact of the proposal on other facial analysis systems. The results obtained will be compared with the selected works using MAE, the most important metric for Head Pose Estimation approaches evaluation.

4.2. State-of-the-art and similar works comparison

Head Pose Estimation is a computer vision task with almost two decades of actively researching. Many papers were produced in this space of time, due the benefits of application and contributions of HPE systems. To identify the published proposals with the highest results, two important sources were used: paperswithcode.com site and an extensive survey published by Carmen Bisogni and Aniello Castiglione (ABATE *et al.*, 2022).

Papers with code was created for easier research papers reading, summary, code implementation, results and benchmarking. It is a very robust tool for finding and comparison state-of-the-art papers and their codes. It is very helpful to save time for manually browsing. Three approaches were selected from this source for analyzing and comparison: (ALBIERO *et al.*, 2021; HEMPEL; ABDELRAHMAN; AL-HAMADI, 2022; VALLE; BUENAPOSADA; BAUMELA, 2020). These solutions presented the highest results for 2D image processing.

HEMPEL; ABDELRAHMAN and AL-HAMADI (2022) proposed a landmark-free head pose estimation method that uses rotation matrix representation for accurate head pose orientation. Authors use the geodesic loss instead of the commonly used mean squared error loss, this way they can penalize the network in the training process using the distance angle between predicted and the ground truth orientation. The results reported, outperformed almost 20% on the AFLW2000 test dataset with a MAE = 3.97 and achieved the state-of-the-art MAE = 3.47 in BIWI database. Despite the results, authors do not inform the time spent for training and

validation of the neural network, the running speed of the model trained, and the hardware used to develop the network.

In (VALLE; BUENAPOSADA; BAUMELA, 2020) a deep learning-based multi-task approach for head pose estimation in images was presented. The authors presented an encoder-decoder multi-task CNN with residual blocks and lateral skip connections and showed that the combination of head pose estimation and landmark-based face alignment improve the performance of the former task. Authors reported top results for the three tasks, head pose, face alignment and visibility (see Figure 5-1). For model evaluation in head pose estimation task, the authors tested the model in four benchmark datasets, obtaining a MAE = 3.83 in AFLW2000 and a MAE = 3.66 in Biwi, outperforming at that time the state-of-the-art approaches. The runtime of their implementation infers head pose at a rate of 62.5 fps using a NVidia GeForce GTX1080Ti (11GB) GPU and a dual Intel Xeon Silver 4114 CPU at 2.20GHz (2×10 cores/40 threads, 128 GB).

Figure 5-1 Multi-task approach. Simultaneous head pose estimation, facial landmark location and their visibility predictions.

Source: (VALLE; BUENAPOSADA; BAUMELA, 2020).

An unusual solution was presented in (ALBIERO *et al.*, 2021), a real-time approach capable of estimating face pose without detecting face or landmark localization. Also, they can obtain accurate 2D face bounding boxes after predicted 3D faces poses, with minimal computational overhead. Authors presented a Faster R-CNN-based framework, reporting real-time running time using a Titan XP GPU and a four-day training consumption time. Results on AFLW2000 and Biwi benchmark datasets presented MAE of 3.91 and 3.78, respectively.

As mentioned before, the study of (ABATE *et al.*, 2022) was very advantageous, the article confirmed Biwi and AFLW2000 as the most popular benchmark datasets,

they were the databases more used for testing HPE applications. The article also compares training and free-training methods, resalting the better results obtained for training ones, but its disadvantages considering the huge difference in the computational time required to build, readjust and optimize the model, the necessity of GPU for training and running, the lack of generalization and real-time running capability in most of the cases. Authors also analyzed the state-of-the-art in three most popular benchmark database: Biwi, AFLW2000 and Pointing´04.

In 2020, the authors of the survey mentioned before, presented a solution being the best model in Biwi dataset. The proposal presented in (ABATE *et al.*, 2020), use a regression algorithm applied to Web-Shaped Model (WSM) algorithm to rapid predict the pose of the face, without training any neural network and without the need for a large amount of data. The proposal benefits of the extraction of 68 face landmarks coordinates. Results obtained in Biwi dataset showing a very low MAE of 2.43. In (ABATE *et al.*, 2022) the authors made references to fast processing and running time, but in this work (ABATE *et al.*, 2020), there is no mention of the estimation speed, or the hardware used in the process.

In (XIA *et al.*, 2019), the authors achieved the best result in AFLW2000 database, according to (ABATE *et al.*, 2022), combining the advantages of model-based methods and appearance-based methods. The landmarks are localized and used as input of CNN, consisting of a task simplifier, a heatmap generator and a feed-forward neural network for face pose estimation. The hardware configuration is as follows: NVIDIA TITAN Xp graphics card, 12GB GPU memory, i9-7900X @3.60GHz × 10 processor and 32GB RAM. Authors mention no information about the running time. Remarkable results are presented in the previous mentioned database, a MAE of 1.46 was achieved.

So far, papers from 2019 to 2022 were described, each one reporting the best result in benchmark databases. The main conclusions of this state-of-the-art papers comparison are: first, even though each new article says that it is better than what was previously published in a certain database, there may be older articles with better results that were not considered in the comparison. Second, training-free solutions are less popular and researched on these days, but they still are an excellent option. Third, multi-task approaches are a good choice both in training and

training-free solutions. Fourth, training solutions such as neural networks, in general, could lead to slightly better results, but they require a huge amount of computational power, making hard for replication and implementation in real-world problems solutions. Five, in the last years papers were focusing on lowest MAE acquisition in benchmark datasets; this could be an erratic way to address the research, to resolve a practical real-world problem could be a better way to develop HPE proposals (ABATE *et al.*, 2022).

As was detailed in Chapter 4, our proposal consists in a multi-task free-training system. Our method was designed to detect faces on images, knowing the face area, 68 landmarks points were localized and finally, we estimated the pose solving the PnP problem. The approach is easy to implement and replicate, also it can run in real-time. The proposal was tested in two benchmark databases, obtaining a MAE of 4.24, as its best result.

Initially, to analyze and validate head pose estimation performance, the approach was tested in FEI Face Database. This database was specially created to train and test face analysis algorithms. The MAE obtained here was excellent. Considering that our approach is a training free solution, a MAE of 4.24 was pretty good, there was almost no difference between the estimated angles and the true value of the angles. This result was briefly compared to state-of-the-art solutions mentioned before, and compared with another landmark based head pose estimation solution (see Table 5-1).

Table 5-0-1 Head Pose Estimation MAE´s Comparison.

RESULTADO EM FEI FACE DATABASE = 4.24	
Artigos	MAE (detecção yaw)
Estado-da-arte	
Nossa proposta	**4.24**
(Hempel et al., 2022)	3.97
(Albiero et al., 2021)	3.91
(Valle et al., 2020)	3.83
Soluções baseadas em pontos de referência	
HAR (Bertók & Fazekas, 2016)	10.06
Nossa proposta	**4.24**
OpenFace (Baltrusaitis et al., 2016)	2.8

Despite the good result obtained, the database initially used has the disadvantage of only capturing face pose movements in the yaw axis. Furthermore, the FEI Face database is not widely used for head pose estimation solutions comparison. Considering the mentioned before, was decided to test our approach in one the most popular benchmark database for head pose estimation, Pointing´04. This database was specifically created for head pose estimation task. The database contains more than 2700 face images captured with highly precision in a laboratory, with poses between -90 and +90, for horizontal and vertical movements.

The result obtained at Pointing´04 was also pretty good (see Table 5-2). Training solutions presented amazing performance when they were trained and tested in the same database. Training free solutions or solutions that only used Pointing´04 for testing presented good numbers but very distant from what was mentioned above. The approach obtained a MAE of 9,78. Comparing the MAE obtained with state-of-the-art solutions can be seen that the mark reached is second between solutions that only used the database for testing, remembering that the presented proposal is a training free solution.

Table 5-0-2 Head Pose Estimation MAE´s Comparison at Pointing´04 Benchmark Database.

RESULTADO EM POINTING´04 DATABASE = 9.78	
Artigos	MAE (detecção yaw)
Treinamento e Teste	
(Mekami et al., 2020)	**1.30**
(Xu et al., 2019)	**3.92**
(Gao et al., 2016)	4.64
(Hara & Chellappa, 2014)	4.83
(Geng & Xia, 2014)	6.45
(Valle et al., 2016)	7.84
Somente Teste	
(Barra et al., 2020)	8.49
Nossa proposta	**9.78**
(Barra et al., 2018)	15.00

Only analyzing Head Pose Estimation performances, the MAE obtained is a little bit higher than state-of-the-art approaches. Our proposal handled well the challenges of one of the most popular benchmark databases for HPE task. It cannot

be considered the best alternative, but it real-time capability, low computational power necessity, good generalization and easy tunning and adjusting, turned our proposal more suitable for real world applications, in comparison with more complex state-of-the-art approaches.

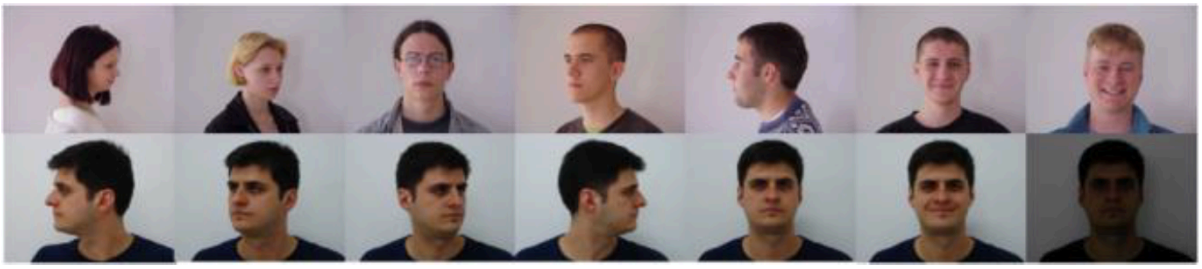
4.3. Face classification model

After estimating the pose of faces, it was decided to create a supervised ML model capable of classification the poses estimated. The head pose estimation is only given by angles, the Euler angles; they define the position of the human head respect to the camera. Face poses of interest can be quickly detected after training a model with the desired poses. The model is trained with the Euler angles of the desired poses, later the model could easily and quickly classificate if the new face pose estimated belongs to the desired position or not.

The process of registration and account opening/verification requires images with good resolution, acceptable lighting conditions, good image quality in general, and faces as frontal as possible, with the user's attention focused on the camera. So, it was necessary to develop a model for frontal face classification. The performance level of the classifier depends greatly on the images it was trained on, so training the classifier with the appropriate databases for the desired scenario of using, is a crucial matter.

After conducting a literature review, two image databases were selected: the CVL Face Database (PEER *et al.*, 2014) and the FEI Face Database (TENORIO; THOMAZ, 2011; THOMAZ; GIRALDI, 2010) (see Figure 5-2). Although these databases were not created under ideal conditions for our scenario, they meet most of the requirements. These databases were primarily created to train and test facial detectors under different conditions, for facial expression detection and analysis, and face recognition purposes.

Figure 5-2 Images preview from selected databases. Top: CVL Face Database. Bottom: FEI Face Database.



Source: The Author (2023).

The CVL Face Database is a database of approximately 800 images created by the University of Ljubljana. In the database's creation, 114 individuals, 90% of whom were men, storing 7 photos of each of them. The photos were taken under uniform lighting conditions with a white background. All images have a resolution of 640x480 pixels and were taken with the same camera. The database was designed for face detection and recognition, feature detection, and 3D modeling of faces.

The FEI Face Database is a database containing 2800 images of 200 individuals, half of them men and half women. It was created by the Artificial Intelligence Laboratory at FEI in São Bernardo do Campo, São Paulo, Brazil. The photos feature a consistent white background and a frontal vertical position, with profile rotation of up to 180 degrees.

To create the head pose classifier, the aforementioned databases were used. In this case, a SVM model, was trained to determine whether a face is frontal or not. In this trial, face and landmarks detection was performed using the open-source library dlib. With the face and the 68 detected landmarks, the open-source computer vision library OpenCV will be used to determine the different Euler angles, solving the PnP problem, to define the three DOF of head movement. The Euler angles of a frontal face are close to zero.

The HPE procedure described above will be applied to the chosen databases. According to the projected scenario of account opening and/or verification, the images were manually divided by face positions, into frontal and non-frontal categories. The SVM model is generated after training to learn and to classify the head position as frontal or non-frontal based on the detected Euler angles. This

ensures that when a new image is inputted using the generated model, it can recognize whether the image is frontal or not, see Figure 5-3. The system estimates the pose of every detected face using the Euler angles. The SVM model then classifies the head position by validating whether the Euler angles correspond to a frontal face or not.

Figure 5-3 Head pose classification process.

Source: The Author (2023).

Three experiment protocols were conducted to find the best possible SVM classifier configuration. Firstly, the classifier was trained with the CVL Face Database and tested on the FEI Face Database in **protocol 1**. Then, in **protocol 2**, the reverse process was carried out to assess the classifier's behavior variation. Lastly, both databases were combined for training and testing in **protocol 3**.

The results of protocols are shown in Table 5-3. As explained previously, the CVL Face Database consists of 7 images for each person. Among these images, three are frontal, and four are non-frontal. The non-frontal images include two images taken at a complete profile angle (+90 degrees and -90 degrees) and two images with the user looking at +45 degrees and -45 degrees. The FEI Face Database contains 6 frontal and 8 non-frontal images for each user. The non-frontal images in this database also cover a range of 180 degrees (from +90 to -90 degrees), but unlike CVL Face Database, it records more non-frontal angles.

It's worth noting that during both the training and testing processes, dlib only detects 482 out of 789 images (328 frontal and 154 non-frontal) in the CVL Face Database and 2627 out of 2800 images (1177 frontal and 1450 non-frontal) in the FEI Face Database. This occurs because it doesn't detect fully profile faces and faces in very low lighting conditions.

Table 5-0-3 Performance metrics. Protocol 1.

	Metrics		
	Precision	*Recall*	*F1-score*
Non-Frontal	0.96	0.91	0.93
Frontal	0.90	0.95	0.92

Source: The Author (2023).

The classifier's performance is indicated by the metrics: Precision, Recall, and F1-score. *Precision* is the percentage of instances that our model correctly classified as frontal or non-frontal. *Recall* is the percentage of all frontal or non-frontal instances that it successfully identified. The *F1-score* is the harmonic mean of precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and its worst at 0.

In the previous table, the classification report of protocol 1 is shown. Despite being trained with only a few hundred images and limited variation in face angles, the model manages to achieve good performance. Although it has an overall accuracy of 92%, the most critical aspect is the 10% of images that it classified as frontal but were not (false positives). In our scenario, reducing the number of false positives is crucial, as it greatly defines the level of reliability and security of our system.

When the roles are reversed, and training is conducted with the FEI Face database, the results improve significantly (see Table 5-4), the model is trained with a larger number of images and variations in angles. This provides more information during training, resulting in a more robust model.

Table 5-0-4 Performance metrics. Protocol 2.

	Metrics		
	Precision	*Recall*	*F1-score*
Non-Frontal	0.97	1.00	0.98
Frontal	1.00	0.98	0.99

Source: The Author (2023).

In this case, the model successfully eliminates false positives and accurately classifies images that were previously categorized in the same way. The model's accuracy in this training was only slightly affected by a small group of frontal images that it misclassified, possibly due to the presence of facial expressions in many images in the databases, which can be a factor in misclassification. These facial expressions may have a minor impact on Euler angles detection, leading to misclassification.

In protocol 3, a larger database was created by combining the images from the two previously studied databases. This merging was done to harness the strengths of both databases for the new model. The results are analyzed in Table 3.

Table 5-0-5 Performance metrics. Protocol 3.

	Metrics		
	Precision	*Recall*	*F1-score*
Non-Frontal	0.96	0.97	0.96
Frontal	0.97	0.96	0.97

Source: The Author (2023).

In the previous results, initially, a model with good performance was achieved, but there was room for improvement in reducing false positives due to training with the weaker database. Subsequently, a much better result was obtained compared to protocol 1. However, it's worth noting that protocol 2 reported better results than protocol 3.

The training, validation, and testing process was carried out in all the protocols, involving splitting the image database into 80% for training and validation and the remaining 20% for testing. This can be easily and efficiently done using the **train_test_split** method from the scikit-learn machine learning library for Python.

The main lesson learned from these testing protocols is that in a supervised learning classifier like SVM, having detailed and descriptive information is much more important than having a larger amount of data with limited description. The model generated from protocol 2 is well-suited for real-world application tests, as it

does not allow the passage of non-frontal faces, and any frontal faces that might be erroneously rejected could be addressed with real-time feedback to the user.

4.4. Contributions to other facial analysis systems

After generating a valid model for facial classification and testing it on image databases, it's essential to validate its operation in projects like the ideal scenario. The creation of the Head Pose Estimator (HPE) and the frontal face classifier in this work, are also part of a research project on computer vision systems at the Laboratory of Computer Networks (LARC). In the mentioned project, alternatives were studied to ensure high levels of security in account opening and login processes while aiming to automate these services on the server. This should be done without compromising the initially provided security levels through reviews and employee validations.

Before developing our system, the project achieved a 92% success rate in identity recognition and a 62% success rate in identifying spoofing attacks. The facial recognition system used a face alignment process before extracting face encodings with a neural network. For spoofing detection, the system used information captured in the HSV and YCbCr color spaces and classified it using a random decision tree called Extra Trees. This approach was another research from LARC, studied and developed as an alternative for non-user collaboration spoofing detection.

The proposed approach was introduced as a preprocessing stage before face recognition and spoofing detection. The primary objective is to analyze only the images that fulfill the requirements. Since the photo to be sent is entirely frontal, validated by the developed classifier, the face alignment process in facial recognition has been eliminated, as it was solely used for frontalizing the face. Frontalization methods are to turn in frontal faces the non-frontal ones. Frontal faces are widely established as the best approach for facial recognition. Furthermore, the literature reports that facial alignment systems or the generation of the missed parts of non-frontal faces can introduce distortions and artifacts in the resulting frontal face.

A server was set up to host the face recognition and spoof detection systems. A mobile application was developed in Java for Android devices, which constitute over 70% of the global mobile market. The enabled server could receive images from

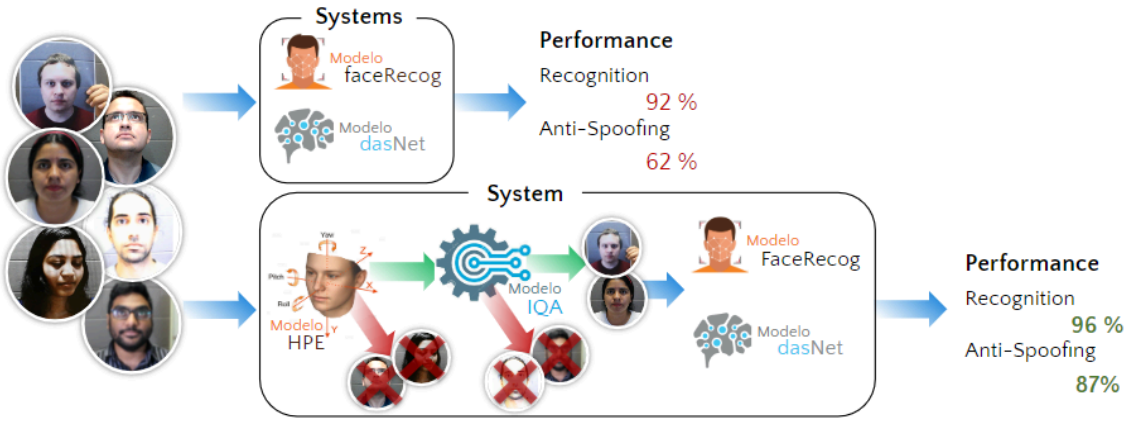
both the mobile application and the system running on the computer, both with the same premise: sending only frontal images with good quality. In both systems, in real-time, the systems provided a response and guidance to the user if the frames of the video being captured now did not meet the minimum requirements.

After testing the image preprocessing and validation systems, we were able to increase the face recognition precision to 96% and spoof detection to 87%. It's worth noting that the improvement in the systems is not only attributed to the pose estimator but also to the combined use with the image quality model. We created a real-time system that validated the face pose, image quality, and provided user-friendly guidance.

Facial analysis systems are generally developed to perform optimally in uncontrolled and challenging environments. They are often only validated using image datasets. Systems developed with this intention can sometimes be impractical for real-world applications. Additionally, many of them require significant computational resources. Currently, there seems to be a trend in academic articles and research focusing more on developing systems to address specific problems rather than creating novel computational tools with no practical real-life applications.

It is very pleasing to see that nowadays, many mobile applications from financial institutions are following a similar methodology to the one presented, allowing only the submission of frontal images with good lighting conditions for analysis. This has resulted in shorter account opening times or verification in sensitive transactions.

Figure 5-4 Measuring the impact of the proposed HPE in other FAS.



Source: LARC (2020).

4.5. Summary and Conclusions

In the previous sections, the main results of the testing and applications, were presented. The values obtained for each of the metrics used in the classification stage are considered encouraging and promising results. Metrics such as precision, recall, and F1-Score represent values above 0.98, which can be considered as a system that achieves an excellent result. Three protocols were tested to evaluate the best way to train the SVM classifier. The protocol 2 outperformed the others achieving a Recall of 1.00 in non-frontal faces, meaning, zero false positive.

It is possible to observe that the chosen databases represent different and detailed face positions that ensure robustness in the system under the desired condition. The images contained in the databases presented different lighting conditions and facial expressions. We observed in some testing that some facial expressions in frontal faces in the border with non-frontal ones, could be misclassified.

After having a supervised ML classifier properly trained, we decided to prove its importance in other FAS. The main idea was using the proposed approach of estimating the head pose as pre-processing stage in face recognition and liveness detection. Face recognition systems usually have alignment methods before features extraction and face codification. The alignment methods are necessary for face recognition systems, but, in the frontalization process they create distortions and artifacts. A wise decision could be to substitute the face alignment with the HPE system with frontal face classification.

A Proof of Concept was produced. With the face recognition and liveness detection system hosted in a server, we used the proposed approach, the pc and mobile versions, to pre-process the images before sending to the server. The images pre-processing works in real-time to help and improve the user experience. The HPE approaches worked in association to image quality verification model, designed to overcome illumination challenges. The results achieved were very good, the face recognition system passed from 92% to 96% of accuracy, and the liveness detection system passed from 62% to 87% of accuracy.

CHAPTER 6

5.1. Final Conclusions and Future Works

In this work, we presented a facial analysis system for the classification of frontal faces. The proposed system utilizes a pose estimation system by calculating Euler angles to classify faces as either frontal or non-frontal using a Support Vector Machine (SVM). One of the main advantages of our system is its real-time operation on low-performance computers and its training-free nature, which simplifies maintenance and adaptation to new scenarios.

We conducted an exhaustive review of the state-of-the-art in facial analysis systems, spanning from face detection and landmark detection to pose estimation. We needed to analyze each of these areas because, typically, estimating facial pose requires first detecting the face, and the chosen estimation procedure benefits from facial landmark detection. Relevant works were selected for each area, and a table comparing key articles and solutions for pose estimation was created. It was observed that while some solutions were reported as the most robust, they often didn't function in real-time.

Our research in the state-of-the-art led us to build a system that analyzes and controls each of its components. If the goal was for the entire system to function in real-time, each part had to possess this capability. The system first detects faces and their landmarks, then estimates the pose using the facial reference points to calculate Euler angles, which are subsequently classified by the SVM. The method chosen to calculate Euler angles and estimate pose is commonly referred to in computer vision as the Perspective-n-Point problem. The objective of this method is to determine the pose of a 3D object when we know the locations of n 3D points and their corresponding 2D projections in an image. This method does not require prior training, operates in real-time, and offers good accuracy, making it an excellent solution for those seeking a balance between robustness and speed.

Various metrics were considered to evaluate our system's performance: Mean Absolute Error (MAE) for angle prediction and Precision, Recall, and F1-score for SVM classifier. The MAE in Euler angle prediction was 9,78. Compared to similar

works, the results are satisfactory, though not significantly improved. These results are based on tests conducted in controlled conditions, as the system was designed for use in controlled environments. Three training and testing protocols were also designed to validate the SVM classifier, with Protocol 2 yielding the best results. This protocol used a larger image database and various facial poses for training, along with a less robust database for testing. Notably, it achieved an amazing result with a Recall score of 1.00, thus avoiding 0 false positives.

One of the motivations behind this work was to develop a system that enhances the performance of other facial analysis systems. Our system was used in combination with an image quality validation model in a proof of concept for an account opening and verification system. When the systems were used as preprocessing steps, the server-based recognition system's performance improved from 90% to 96%, and the attack prevention system increased from 62% to 87%. The use of the pose estimator eliminated the need for face alignment within the recognition system. While frontalization are useful in the recognition process, they introduce distortions and artifacts in the generated frontal face.

In the proof of concept, a mobile version of the pose estimator was also used. The system was developed using Google's Firebase ML Kit for Android. The trained and generated SVM model was not used directly to ensure a lightweight and real-time application. However, testing with this model provided insights into the Euler angle ranges that determine whether a face is frontal or not.

As future works, we plan to explore new methods for estimating facial pose without sacrificing real-time functionality. It would be interesting to find robust methods and alternatives to reduce the MAE in Euler angle prediction. Additionally, we aim to train other supervised learning models and compare their performance with the SVM generated in this research. Training and testing new models with image databases that encompass a greater number of images and facial poses will be essential, always ensuring high similarity between the images in the database and the intended usage scenario of the designed model. Furthermore, we intend to test the applicability of the developed system in other real-world problems, such as engagement, attention, and focus analysis.

REFERENCES

ABATE, A. F.; BARRA, P.; PERO, C.; TUCCI, M. Head pose estimation by regression algorithm. **Pattern Recognition Letters**, v. 140, p. 179–185, 1 dez. 2020.

ABATE, A. F.; BISOGNI, C.; CASTIGLIONE, A.; NAPPI, M. Head pose estimation: An extensive survey on recent techniques and applications. **Pattern Recognition**, v. 127, 1 jul. 2022. Disponível em: <https://www.researchgate.net/publication/358647531_Head_Pose_Estimation_An_Extensive_Survey_on_Recent_Techniques_and_Applications>. Acesso em: 8 set. 2023.

ADHIKARI, B.; NI, X.; RAHTU, E.; HUTTUNEN, H. Towards a Real-Time Facial Analysis System. **IEEE 23rd International Workshop on Multimedia Signal Processing, MMSP 2021**, 2021. . Acesso em: 7 set. 2023.

AIFANTI, N.; PAPACHRISTOU, C.; DELOPOULOS, A. The MUG facial expression database. Em: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, 2010, [...]. IEEE, 2010. p. 1–4.

ALBIERO, V.; CHEN, X.; YIN, X.; PANG, G.; HASSNER, T. img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. Em: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, [...]. 2021. p. 7617–7627.

ALEJANDRA PACHECO REINA, P.; MANUEL GUTIÉRREZ MENÉNDEZ, A.; CARLOS GUTIÉRREZ MENÉNDEZ, J.; BRESSAN, G.; RUGGEIRO, W. Understanding the Impact of Image Quality in Face Processing Algorithms. Em: IMPROVE, 2021, [...]. 2021. p. 145–152.

ANTONA CORTÉS, C. **Herramientas modernas en redes neuronales: la librería Keras**. 2017. 2017. Disponível em: <https://repositorio.uam.es/handle/10486/677854>. Acesso em: 30 set. 2018.

BALAKRISHNAN, G.; XIONG, Y.; XIA, W.; PERONA, P. Towards Causal Benchmarking of Biasin Face Analysis Algorithms. **Advances in Computer Vision**

and Pattern Recognition, p. 327–359, 2021. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-74697-1_15>. Acesso em: 7 set. 2023.

BALTRUSAITIS, T.; ROBINSON, P.; MORENCY, L.-P. OpenFace: An open source facial behavior analysis toolkit. Em: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, [...]. IEEE, 2016. p. 1–10.

BAY, H.; ESS, A.; TUYTELAARS, T.; VAN GOOL, L. Speeded-Up Robust Features (SURF). **Computer vision and image understanding**, v. 110, n. 3, p. 346–359, 2008. Disponível em: <www.sciencedirect.com>. Acesso em: 2 set. 2019.

BENABDELKADER, C. Robust head pose estimation using supervised manifold learning. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 6316 LNCS, n. PART 6, p. 518–531, 2010. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-15567-3_38>. Acesso em: 8 set. 2023.

BIN YANG; YAN, J.; LEI, Z.; LI, S. Z. Aggregate channel features for multi-view face detection. Em: IEEE International Joint Conference on Biometrics, 2014, [...]. IEEE, 2014. p. 1–8.

BORGHI, G.; FABBRI, M.; VEZZANI, R.; CALDERARA, S.; CUCCHIARA, R. Face-from-Depth for Head Pose Estimation on Depth Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2018. Disponível em: <https://ieeexplore.ieee.org/document/8567956/>. Acesso em: 14 out. 2019.

BORGHI, G.; FABBRI, M.; VEZZANI, R.; CALDERARA, S.; CUCCHIARA, R. Face-from-Depth for Head Pose Estimation on Depth Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 42, n. 3, p. 596–609, 1 mar. 2020a. . Acesso em: 7 set. 2023.

BORGHI, G.; FABBRI, M.; VEZZANI, R.; CALDERARA, S.; CUCCHIARA, R. Face-from-Depth for Head Pose Estimation on Depth Images. **IEEE Transactions**

on Pattern Analysis and Machine Intelligence, v. 42, n. 3, p. 596–609, 1 mar. 2020b. . Acesso em: 7 set. 2023.

BRADSKI, G.; KAEHLER, A. **Learning OpenCV: Computer vision with the OpenCV library**. [s.l.] O'Reilly Media, Inc., 2008. 557 p.

BRUBAKER, S. C.; WU, J.; SUN, J.; MULLIN, M. D.; REHG, J. M. On the design of cascades of boosted ensembles for face detection. **International Journal of Computer Vision**, v. 77, n. 1–3, p. 65–86, maio 2008. . Acesso em: 13 nov. 2019.

BUOLAMWINI, J.; GEBRU, T. **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender ClassificationProceedings of Machine Learning Research**PMLR, 21 jan. 2018. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. Acesso em: 7 set. 2023.

CARUANA, R. Multitask Learning. **Machine Learning**, v. 28, n. 1, p. 41–75, 1997. . Acesso em: 27 out. 2019.

CHEN, D.; REN, S.; WEI, Y.; CAO, X.; SUN, J. Joint cascade face detection and alignment. Em: European conference on computer vision, PART 6., 2014, [...]. Springer Verlag, 2014. v. 8694 LNCS, p. 109–122.

ÇOŞKUN, K.; ÇETIN, G. A COMPARATIVE EVALUATION OF THE BOOSTING ALGORITHMS FOR NETWORK ATTACK CLASSIFICATION. **J. of 3D Printing Tech. Dig. Ind**, v. 6, n. 1, p. 102–112, 2022. . Acesso em: 7 set. 2023.

CZUPRYŃSKI, B.; STRUPCZEWSKI, A. High accuracy head pose tracking survey. Em: International Conference on Active Media Technology, 2014, [...]. Springer Verlag, 2014. p. 407–420.

DANG, K.; SHARMA, S. Review and comparison of face detection algorithms. Em: Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering, 2017, [...]. 2017.

DOLLÁR, P.; TU, Z.; PERONA, P.; BELONGIE, S. Integral Channel Features. 2009. Disponível em:

<https://authors.library.caltech.edu/60048/1/dollarBMVC09ChnFtrs.pdf>. Acesso em: 2 set. 2019.

DROUARD, V.; BA, S.; EVANGELIDIS, G.; DELEFORGE, A.; HORAUD, R. Head pose estimation via probabilistic high-dimensional regression. **Proceedings - International Conference on Image Processing, ICIP**, v. 2015- December, p. 4624–4628, 9 dez. 2015. . Acesso em: 7 set. 2023.

DROUARD, V.; HORAUD, R.; DELEFORGE, A.; BA, S.; EVANGELIDIS, G. Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions. **IEEE Transactions on Image Processing**, v. 26, n. 3, p. 1428–1440, mar. 2017a. Disponível em: <http://ieeexplore.ieee.org/document/7819497/>. Acesso em: 20 jun. 2019.

DROUARD, V.; HORAUD, R.; DELEFORGE, A.; BA, S.; EVANGELIDIS, G. Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions. **IEEE Transactions on Image Processing**, v. 26, n. 3, p. 1428–1440, 1 mar. 2017b. . Acesso em: 7 set. 2023.

DRUZHKOV, P. N.; ERUKHIMOV, V. L.; ZOLOTYKH, N. Y.; KOZINOV, E. A.; KUSTIKOVA, V. D.; MEEROV, I. B.; POLOVINKIN, A. N. New object detection features in the OpenCV library. **Pattern Recognition and Image Analysis**, v. 21, n. 3, p. 384–386, set. 2011. . Acesso em: 3 dez. 2019.

FANELLI, G.; GALL, J.; VAN GOOL, L.; ZURICH, E. Real time head pose estimation with random regression forests. Em: CVPR 2011, 2011, [...]. IEEE, 2011. p. 617–624.

FENG, Y.; YU, S.; PENG, H.; LI, Y. R.; ZHANG, J. Detect Faces Efficiently: A Survey and Evaluations. **IEEE Transactions on Biometrics, Behavior, and Identity Science**, v. 4, n. 1, p. 1–18, 1 jan. 2022. . Acesso em: 7 set. 2023.

FRISCHHOLZ, R. W.; WERNER, A. Avoiding replay-attacks in a face recognition system using head-pose estimation. Em: 2003 IEEE International SOI Conference. Proceedings, 2003, [...]. 2003. p. 234--235.

GIRSHICK, R.; IANDOLA, F.; DARRELL, T.; MALIK, J. Deformable Part Models are Convolutional Neural Networks. Em: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, [...]. 2015. p. 437--446.

HASSNER, T.; HAREL, S.; PAZ, E.; ROEE ENBAR, †. **Effective Face Frontalization in Unconstrained Images**2015. Disponível em: <www.openu.ac.il/home/hassner/projects/frontalize>. Acesso em: 8 set. 2023.

HEMPEL, T.; ABDELRAHMAN, A. A.; AL-HAMADI, A. 6D Rotation Representation For Unconstrained Head Pose Estimation. Em: 2022 IEEE International Conference on Image Processing (ICIP), 2022, [...]. 2022. p. 2496–2500.

HONG, C.; YU, J.; ZHANG, J.; JIN, X.; LEE, K. H. Multimodal Face-Pose Estimation With Multitask Manifold Deep Learning. **IEEE Transactions on Industrial Informatics**, v. 15, n. 7, p. 3952–3961, 1 jul. 2019. . Acesso em: 8 set. 2023.

JAIN, V.; LEARNED-MILLER, E. **FDDB: A Benchmark for Face Detection in Unconstrained Settings**. [s.l: s.n.]. Disponível em: <http://news.yahoo.com>. Acesso em: 7 set. 2023.

JIANG, H.; LEARNED-MILLER, E. Face Detection with the Faster R-CNN. **Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge**, p. 650–657, 2017.

KAKADIARIS, I. A.; TODERICI, G.; EVANGELOPOULOS, G.; PASSALIS, G.; CHU, D.; ZHAO, X.; SHAH, S. K.; THEOHARIS, T. 3D-2D face recognition with pose and illumination normalization. **Computer Vision and Image Understanding**, v. 154, p. 137–151, 1 jan. 2017. . Acesso em: 2 nov. 2019.

KHALIL, A.; AHMED, S. G.; KHATTAK, A. M.; AL-QIRIM, N. Investigating Bias in Facial Analysis Systems: A Systematic Review. **IEEE Access**, v. 8, p. 130751–130761, 2020. . Acesso em: 7 set. 2023.

KHAN, K.; KHAN, R. U.; LEONARDI, R.; MIGLIORATI, P.; BENINI, S. Head pose estimation: A survey of the last ten years. **Signal Process. Image Commun.**, v. 99, 1 nov. 2021. . Acesso em: 7 set. 2023.

KIM, E.-H.; KIM, B.-Y.; OH, S.-K.; KIM, J.-Y. Design of Robust Face Recognition System Realized with the Aid of Automatic Pose Estimation-based Classification and Preprocessing Networks Structure. **Journal of Electrical Engineering and Technology**, v. 12, n. 6, p. 2388--2398, 2017. Disponível em: <http://doi.org/10.???/JEET.2017.12.3.1921>. Acesso em: 29 out. 2019.

KIM, S. H.; OH, S. K.; KIM, J. Y. Design of Face Recognition System Realized with the Aid of PCA-Based RBFNN. **Proceedings - 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2016**, p. 754–758, 2016.

KING, D. E. Dlib-ml: A machine learning toolkit. **Journal of Machine Learning Research**, v. 10, n. Jul, p. 1755–1758, 2009.

KÖSTINGER, M.; WOHLHART, P.; ROTH, P. M.; BISCHOF, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. **Proceedings of the IEEE International Conference on Computer Vision**, p. 2144–2151, 2011a. . Acesso em: 7 set. 2023.

KÖSTINGER, M.; WOHLHART, P.; ROTH, P. M.; BISCHOF, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. **Proceedings of the IEEE International Conference on Computer Vision**, p. 2144–2151, 2011b.

KUHNKE, F.; OSTERMANN, J. Deep Head Pose Estimation Using Synthetic Images and Partial Adversarial Domain Adaption for Continuous Label Spaces. Em: Proceedings of the IEEE International Conference on Computer Vision, 2019, [...]. 2019. p. 10164--10173.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, 28 maio 2015. Disponível em: <http://www.nature.com/articles/nature14539>. Acesso em: 4 out. 2018.

LI, H.; LIN, Z.; SHEN, X.; BRANDT, J.; HUA, G. A Convolutional Neural Network Cascade for Face Detection. Em: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, [...]. 2015. p. 5325–5334.

LI, J.; ZHANG, Y. Learning SURF Cascade for Fast and Accurate Object Detection. Em: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, [...]. 2013. p. 3468--3475.

LI, J.; ZHANG, Y.; WANG, T. Face detection using SURF cascade. Em: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, [...]. IEEE, 2011. p. 2183–2190.

LI, S. Z.; ZHANG, Z. Q. FloatBoost learning and statistical face detection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 26, n. 9, p. 1112–1123, set. 2004. . Acesso em: 7 set. 2023.

LI, S. Z.; ZHANG, Z.; SHUM, H.-Y.; ZHANG, H. FloatBoost Learning for Classification. **Advances in Neural Information Processing Systems**, v. 15, 2002. Disponível em: <http://research.microsoft.com/>. Acesso em: 7 set. 2023.

LIAO, S.; JAIN, A. K.; LI, S. Z. A Fast and Accurate Unconstrained Face Detector. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 38, n. 2, p. 211–223, 1 fev. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7130626/>. Acesso em: 27 mar. 2019.

LIENHART, R.; KURANOV, A.; PISAREVSKY, V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Em: Joint Pattern Recognition Symposium, 2003, [...]. 2003. p. 297–304.

LIENHART, R.; MAYDT, J. An extended set of Haar-like features for rapid object detection. Em: Proceedings. International Conference on Image Processing, 2002, [...]. IEEE, 2002. v. 1, p. I-900-I–903.

LIU, X.; LIANG, W.; WANG, Y.; LI, S.; PEI, M. 3D head pose estimation with convolutional neural network trained on synthetic images. Em: 2016 IEEE International Conference on Image Processing (ICIP), 2016, [...]. IEEE, 2016. p. 1289–1293.

MAGDIN, M.; BENKO, L.; KOPRDA, Š. A Case Study of Facial Emotion Classification Using Affdex. **Sensors 2019, Vol. 19, Page 2140**, v. 19, n. 9, p. 2140, 9 maio 2019. Disponível em: <https://www.mdpi.com/1424-8220/19/9/2140/htm>. Acesso em: 7 set. 2023.

MALLICK, S. **Head Pose Estimation using OpenCV and Dlib | Learn OpenCV**. Disponível em: <https://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/>. Acesso em: 4 dez. 2019.

MATAS, J.; LUETTIN, J.; MESSER, K.; MATAS, J.; KITTLER, J.; LUETTIN, J.; MAITRE, G. XM2VTSDB: The extended M2VTS database. Em: Second international conference on audio and video-based biometric person authentication, 1999, [...]. 1999.

MATHIAS, M.; BENENSON, R.; PEDERSOLI, M.; VAN GOOL, L. Face detection without bells and whistles. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 8692 LNCS, n. PART 4, p. 720–735, 2014. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-10593-2_47>. Acesso em: 7 set. 2023.

MITA, T.; KANEKO, T.; HORI, O. Joint Haar-like features for face detection. Em: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, [...]. IEEE, 2005. p. 1619-1626 Vol. 2.

MITÉRAN, J.; MATAS, J.; BOURENNANE, E.; PAINDAVOINE, M.; DUBOIS, J. Automatic hardware implementation tool for a discrete Adaboost-based decision algorithm. **Eurasip Journal on Applied Signal Processing**, v. 2005, n. 7, p. 1035–1046, 11 maio 2005. Disponível em: <https://link.springer.com/articles/10.1155/ASP.2005.1035>. Acesso em: 7 set. 2023.

MURPHY-CHUTORIAN, E.; DOSHI, A.; TRIVEDI, M. M. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. **IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC**, p. 709–714, 2007. . Acesso em: 7 set. 2023.

MURPHY-CHUTORIAN, E.; TRIVEDI, M. M. Head Pose Estimation in Computer Vision: A Survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 31, n. 4, p. 607–626, abr. 2009. Disponível em: <http://ieeexplore.ieee.org/document/4497208/>. Acesso em: 19 set. 2019.

NADA, H.; SINDAGI, V. A.; ZHANG, H.; PATEL, V. M. Pushing the Limits of Unconstrained Face Detection: a Challenge Dataset and Baseline Results. **2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems, BTAS 2018**, 26 abr. 2018. Disponível em: <https://arxiv.org/abs/1804.10275v3>. Acesso em: 7 set. 2023.

NOCK, R.; NIELSEN, F.; NOCK, R.; NIELSEN, F. A Real Generalization of Discrete AdaBoost. 2006. . Acesso em: 7 set. 2023.

PAPAZOV, C.; MARKS, T. K.; JONES, M. Real-Time 3D Head Pose and Facial Landmark Estimation From Depth Images Using Triangular Surface Patch Features. Em: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, [...]. 2015. p. 4722–4730.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <http://scikit-learn.sourceforge.net.>. Acesso em: 4 dez. 2019.

PEER, P.; EMERŠIČ, Z.; BULE, J.; ŽGANEC-GROS, J.; ŠTRUC, V. Strategies for exploiting independent cloud implementations of biometric experts in multibiometric scenarios. **Mathematical Problems in Engineering**, v. 2014, 2014.

PULLI, K.; BAKSHEEV, A.; KORNYAKOV, K.; ERUHIMOV, V. Real-Time Computer Vision with OpenCV. **Communications of the ACM**, v. 55, n. 6, p. 61–69, 2012. Disponível em: <queue.acm.org>.

RANJAN, R.; PATEL, V. M.; CHELLAPPA, R. A deep pyramid Deformable Part Model for face detection. Em: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2015, [...]. IEEE, 2015. p. 1–8.

RANJAN, R.; PATEL, V. M.; CHELLAPPA, R. HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 41, n. 1, p. 121–135, 1 jan. 2019a. . Acesso em: 5 set. 2023.

RANJAN, R.; PATEL, V. M.; CHELLAPPA, R. HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 41, n. 1, p. 121–135, 1 jan. 2019b. . Acesso em: 7 set. 2023.

RANJAN, R.; SANKARANARAYANAN, S.; CASTILLO, C. D.; CHELLAPPA, R. An All-In-One Convolutional Neural Network for Face Analysis. Em: 2017 12th IEEE International Conference on Automatic Face \& Gesture Recognition (FG 2017), 2017, [...]. IEEE, 2017. p. 17–24.

RAY, S. J.; TEIZER, J. Coarse head pose estimation of construction equipment operators to formulate dynamic blind spots. **Adv. Eng. Informatics**, v. 26, n. 1, p. 117–130, jan. 2012. . Acesso em: 7 set. 2023.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Em: Advances in neural information processing systems, 2015, [...]. 2015. p. 91--99.

RIEGLER, G.; FERSTL, D.; RÜTHER, M.; BISCHOF, H. Hough Networks for Head Pose Estimation and Facial Feature Localization. **Journal of Computer Vision**, v. 101, n. 3, p. 437--458, 2013. . Acesso em: 4 dez. 2019.

RUIZ, N.; CHONG, E.; REHG, J. M. Fine-Grained Head Pose Estimation Without Keypoints. Em: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, [...]. 2018. p. 2074--2083.

SANKARANARAYANAN, K.; CHANG, M. C.; KRAHNSTOEVER, N. Tracking gaze direction from far-field surveillance cameras. **2011 IEEE Workshop on Applications of Computer Vision, WACV 2011**, p. 519–526, 2011. . Acesso em: 8 set. 2023.

SAVCHENKO, A. V. **Video-Based Frame-Level Facial Analysis of Affective Behavior on Mobile Devices Using EfficientNets**2022. . Acesso em: 7 set. 2023.

SCHULZ, A.; DAMER, N.; FISCHER, M.; STIEFELHAGEN, R. Combined head localization and head pose estimation for video-based advanced driver assistance systems. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 6835 LNCS, p. 51–60, 2011. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-23123-0_6>. Acesso em: 7 set. 2023.

SHAHRAKI, A.; ABBASI, M.; HAUGEN, Ø. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. **Engineering Applications of Artificial Intelligence**, v. 94, 1 set. 2020. Disponível em: <https://www.researchgate.net/publication/342617769_Boosting_algorithms_for_network_intrusion_detection_A_comparative_evaluation_of_Real_AdaBoost_Gentle_AdaBoost_and_Modest_AdaBoost>. Acesso em: 7 set. 2023.

SHAO, X.; QIANG, Z.; LIN, H.; DONG, Y.; WANG, X. A survey of head pose estimation methods. Em: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), 2020, [...]. IEEE, 2020. p. 787–796.

SHARMA, S.; SHANMUGASUNDARAM, K.; RAMASAMY, S. K. FAREC - CNN based efficient face recognition technique using Dlib. **Proceedings of 2016**

International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2016, n. 978, p. 192–195, 2017.

SMITH, K.; BA, S. O.; ODOBEZ, J. M.; GATICA-PEREZ, D. Tracking the visual focus of attention for a varying number of wandering people. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 30, n. 7, p. 1212–1229, jul. 2008. . Acesso em: 8 set. 2023.

T, S.; RAJEST, S. S.; REGIN, R.; R, Steffi. A Review on Using Machine Learning to Conduct Facial Analysis in Real Time for Real-Time Profiling. **International Journal of Human Computing Studies**, v. 5, n. 2, p. 18–37, 23 fev. 2023. Disponível em: <https://researchparks.innovativeacademicjournals.com/index.php/IJHCS/article/view/6132>. Acesso em: 7 set. 2023.

TAIGMAN, Y.; YANG, M.; RANZATO, M.; WOLF, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Em: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, [...]. 2014. p. 1701--1708.

TENORIO, E. Z.; THOMAZ, C. E. Analise multilinear discriminante de formas frontais de imagens 2D de face. **Proceedings of the X simp{\'o}sio brasileiro de automa{\c{c}}{\~a}o inteligente SBAI**, p. 266–271, 2011.

THAI, C.; TRAN, V.; BUI, M.; NINH, H.; TRAN, H. An Effective Deep Network for Head Pose Estimation without Keypoints. **arXiv preprint arXiv:2210.13705**, 24 out. 2022. Disponível em: <http://arxiv.org/abs/2210.13705>.

THOMAZ, C. E.; GIRALDI, G. A. A new ranking method for principal components analysis and its application to face image analysis. **Image and Vision Computing**, v. 28, n. 6, p. 902–913, 2010. Disponível em: <https://www.researchgate.net/publication/220612060_A_new_ranking_method_for_Principal_Components_Analysis_and_its_application_to_face_image_analysis>. Acesso em: 8 set. 2023.

TRIANTAFYLLIDOU, D.; NOUSI, P.; TEFAS, A. Fast Deep Convolutional Face Detection in the Wild Exploiting Hard Sample Mining. **Big Data Research**, v. 11, p.

65–76, mar. 2018. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2214579617300096>. Acesso em: 20 jun. 2019.

TRUONG, V. T.; LAO, J. S.; HUANG, C. C. Multi-camera Marker-based Real-time Head Pose Estimation System. Em: 2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020, 2020, [...]. Institute of Electrical and Electronics Engineers Inc., 2020.

VALLE, R.; BUENAPOSADA, J. M.; BAUMELA, L. Multi-task head pose estimation in-the-wild. **IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE**, v. 43, n. 8, p. 2874–2881, 4 fev. 2020. Disponível em: <http://arxiv.org/abs/2202.02299>.

VALLE, R.; BUENAPOSADA, J. M.; VALDÉS, A.; BAUMELA, L. Head-Pose Estimation In-the-Wild Using a Random Forest. Em: International Conference on Articulated Motion and Deformable Objects, 2016, [...]. Springer, 2016. p. 24–33.

VAROQUAUX, G.; BUITINCK, L.; LOUPPE, G.; GRISEL, O.; PEDREGOSA, F.; MUELLER, A. Scikit-learn: Machine Learning Without Learning the Machinery. **GetMobile: Mobile Computing and Communications**, v. 19, n. 1, p. 29–33, 2015. Disponível em: <http://dl.acm.org/citation.cfm?id=2786984.2786995>.

VENTURELLI, M.; BORGHI, G.; VEZZANI, R.; CUCCHIARA, R. From Depth Data to Head Pose Estimation: a Siamese approach. **arXiv preprint arXiv:1703.03624**, 2017. Disponível em: <http://www.distraction.gov/index.html>. Acesso em: 4 dez. 2019.

VIOLA, P.; JONES, M. Fast Multi-view Face Detection. **Mitsubishi Electric Research Lab TR2000396**, v. 3, n. May, p. 2, 2003. Disponível em: <https://www.researchgate.net/publication/228362107>. Acesso em: 3 set. 2019.

VIOLA, P.; JONES, M. J. Rapid object detection using a boosted cascade of simple features. **CVPR**, v. 1, n. 3, p. 511–518, 2001. Disponível em: <https://www.researchgate.net/publication/3940582>. Acesso em: 27 mar. 2019.

WANG, X.; HAN, T. X.; YAN, S. An HOG-LBP human detector with partial occlusion handling. Em: 2009 IEEE 12th International Conference on Computer Vision, 2009, [...]. IEEE, 2009. p. 32–39.

WILLIAMS PONTIN, M. **Better Face-Recognition Software | MIT Technology Review**. Disponível em: <https://www.technologyreview.com/2007/05/30/225291/better-face-recognition-software/>. Acesso em: 7 set. 2023.

WOLLASTON, W. H. On the Apparent Direction of Eyes in a Portrait. **Philosophical Transactions of the Royal Society of London**, v. 114, n. 0, p. 247--256, 1 jan. 1824. Disponível em: <http://rspl.royalsocietypublishing.org/cgi/doi/10.1098/rspl.1815.0236>. Acesso em: 29 set. 2019.

XIA, J.; CAO, L.; ZHANG, G.; LIAO, J. Head Pose Estimation in the Wild Assisted by Facial Landmarks Based on Convolutional Neural Networks. **IEEE Access**, v. 7, p. 48470–48483, 2019.

XU, X.; LE, H. A.; DOU, P.; WU, Y.; KAKADIARIS, I. A. Evaluation of a 3D-aided pose invariant 2D face recognition system. **IEEE International Joint Conference on Biometrics, IJCB 2017**, v. 2018- Janua, p. 446–455, 2018.

YAN, J.; LEI, Z.; WEN, L.; LI, S. Z. The Fastest Deformable Part Model for Object Detection. Em: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, [...]. 2014. p. 2497–2504.

YAN, J.; ZHANG, X.; LEI, Z.; LI, S. Z. Real-time high performance deformable model for face detection in the wild. Em: 2013 International Conference on Biometrics (ICB), 2013, [...]. IEEE, 2013. p. 1–6.

YAN, Y.; RICCI, E.; SUBRAMANIAN, R.; LIU, G.; LANZ, O.; SEBE, N. A Multi-Task Learning Framework for Head Pose Estimation under Target Motion. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 38, n. 6, p. 1070–1083, 1 jun. 2016. Disponível em: <https://ieeexplore.ieee.org/document/7254213/>. Acesso em: 20 jun. 2019.

YANG, H.; MOU, W.; ZHANG, Y.; PATRAS, I.; GUNES, H.; ROBINSON, P. Face Alignment Assisted by Head Pose Estimation. **arXiv preprint arXiv:1507.03148**, 11 jul. 2015. Disponível em: <http://arxiv.org/abs/1507.03148>. Acesso em: 20 jun. 2019.

YIN, X.; YU, X.; SOHN, K.; LIU, X.; CHANDRAKER, M. **Towards Large-Pose Face Frontalization in the Wild**2017. Disponível em: <http://cvlab.cse.>. Acesso em: 8 set. 2023.

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. **Computer Vision and Image Understanding**, v. 138, p. 1–24, set. 2015. Disponível em: <http://dx.doi.org/10.1016/j.cviu.2015.03.015>. Acesso em: 20 jun. 2019.

ZELINSKY, A. Learning OpenCV---Computer Vision with the OpenCV Library (Bradski, GR et al.; 2008)[On the Shelf]. **IEEE Robotics \& Automation Magazine**, v. 16, n. 3, p. 100--100, 2009.

ZHANG, C.; ZHANG, Z. **A Survey of Recent Advances in Face Detection**. [s.l: s.n.]. Disponível em: <http://www.research.microsoft.com>. Acesso em: 28 ago. 2019.

ZHANG, L.; CHU, R.; XIANG, S.; LIAO, S.; LI, S. Z. Face Detection Based on Multi-Block LBP Representation. Em: International conference on biometrics, 2007, [...]. Springer, 2007. p. 11–18.

ZHANG, S.; ZHU, X.; LEI, Z.; SHI, H.; WANG, X.; LI, S. Z. FaceBoxes: A CPU real-time face detector with high accuracy. **IEEE International Joint Conference on Biometrics, IJCB 2017**, v. 2018- January, p. 1–9, 29 jan. 2018. . Acesso em: 7 set. 2023.

ZHANG, X.; SUGANO, Y.; FRITZ, M.; BULLING, A. Appearance-based gaze estimation in the wild. Em: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, [...]. 2015. p. 4511–4520.

ZHANG, Z.; LUO, P.; LOY, C. C.; TANG, X. Facial landmark detection by deep multi-task learning. Em: European conference on computer vision, 2014, [...]. Springer Verlag, 2014. p. 94–108.

ZHANG, Z. Q.; LI, M. J.; LI, S. Z.; ZHANG, H. J. Multi-view face detection with FloatBoost. **Proceedings of IEEE Workshop on Applications of Computer Vision**, v. 2002- January, p. 184–188, 2002. . Acesso em: 7 set. 2023.

ZHOU, Y.; GREGSON, J. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose. **arXiv preprint arXiv:2005.10353**, 20 maio 2020. Disponível em: <http://arxiv.org/abs/2005.10353>.

ZHU, X.; LEI, Z.; YAN, J.; YI, D.; LI, S. Z. **High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild**2015. Disponível em: <http://www.cbsr.ia.ac.>. Acesso em: 8 set. 2023.

ZHU, X.; RAMANAN, D. Face detection, pose estimation, and landmark localization in the wild. Em: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, [...]. IEEE, 2012. p. 2879–2886.