University of Salerno

Department of Computer Science

Dottorato di Ricerca in Informatica
Curriculum Computer Science and Information Technology
XXXV Ciclo

Tesi di Dottorato / Ph.D. Thesis

Behavioral Biometrics in the Era of Artificial Intelligence

Chiara PERO

Supervisor:	**Prof. Michele NAPPI**
PhD Program Director:	**Prof. Andrea DE LUCIA**

A.A 2021/2022

Dedicated to my parents,
for their endless love.
I am nothing without you.

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to my tutor, Prof. Michele Nappi, for his continuous support during my Ph.D. studies as well as for his patience and inspiration. His knowledge and experience helped me during my doctoral studies and in my everyday life. This undertaking would not have been possible without the invaluable supervision and tutelage of Prof. Aniello Castiglione, who was the first to believe in me. I am also extremely grateful to Prof. Andrea Abate, a truly exceptional person, for making these years so much more enjoyable and for being such a special mentor. I would like to extend my sincere thanks to my research group, Fabio Narducci, Carmen Bisogni, Lucia Cascone, and Lucia Cimmino, for the wonderful time spent together in the lab and because, after all, you will always represent a fundamental part of my life, no matter what happens. Last but not least, I would like to express all my affection to my parents and Francesco. Without their tremendous understanding and encouragement, it would have been impossible for me to go through this tortuous but beloved journey.

ABSTRACT

Biometric technologies have historically been explored as Pattern Recognition systems. Over the past three decades, biometric-based human recognition systems have significantly changed and improved. Applications in forensics, surveillance, healthcare, automotive, and human-computer interaction have benefited of such advancements and are currently globally used. However, the exclusive focus on pattern recognition may obscure or restrict the potential and capabilities of this discipline. Emerging biometric modalities have begun to impact the security of sensitive data, information, and systems. As the biometric challenges increase, the solution strategies shifted the attention on human behavior. Behavioral biometrics is the study of patterns in human activities that can be uniquely identified and measured. Recent technological advancements, especially in Artificial Intelligence, as well as hardware development, have increased the potential of biometric approaches and expanded their application fields. Beginning with the wide concept of behavioral biometrics, this thesis aims to advance the state-of-the-art in several applications, such as estimating an individual's head rotation to determine its intent or attention and analyzing facial expressions to detect human emotional state. Finally, behavioral biometric traits are investigated through users' touch interactions with modern mobile devices. For each of the presented methods, the complete processing pipeline is described, including data acquisition, feature extraction, experimental protocols, and decision-making, as well as a comparison to state-of-the-art methods to show advantages and discuss current challenges.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

HCI	Human-computer interaction
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
ANN	Artificial Neural Network
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
HPE	Head Pose Estimation
MAE	Mean Absolute Error
PIFS	Partitioned Iterated Function System
IFS	Iterated Function System
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
FER	Facial Expression Recognition
FACS	Facial Action Coding System
AUs	Action Units
TD	Touch Dynamics

1 INTRODUCTION

The idea behind biometrics dates back to the dawn of human perception. Identity verification systems refer to the art of developing authentication strategies with the aid of biometric features to automatically identify, measure, and validate a living human being. As a wide variety of applications require reliable verification schemes to confirm an individual's identity, biometrics is increasingly involved in everyday applications today. The need for more modern and sophisticated authentication systems is driving change in many organizations, exploring new key factors. The most promising of these is behavioral biometrics. In this work, we have presented a series of techniques developed during the three-year PhD course with a specific emphasis on behavioral biometrics. The next Chapters and Sections will show that human behavioral patterns are currently one of the most emerging biometric modalities, encompassing a broad range of application domains.

1.1 BIOMETRICS: DEFINITIONS AND APPLICATIONS

It is essential to first focus on the definition and history of biometrics to fully understand how contemporary biometric technology works. The study of the variability of characteristics between populations of living beings is known as biometrics. These characteristics can be defined by the meaning of the word "*Biometrics*", composed of bios, "*life*", and metron, "*measure*". The scientific assumptions of biometrics have led to the creation of biometric technology, which allows the identification and verification of an individual on the basis of physiological and behavioral traits. The former are related to the characteristics of the human body and include, for example, the iris, facial geometry, and fingerprints. Through a person's specific and repetitive behavior, behavioral traits develop. Examples include gait, typing on a keyboard, or a signature.

The first documented application of biometrics as a security measure dates back to the mid-1800s. Lawmakers originally kept records of photographs in loosely organized groups; this inefficient system made it difficult for police to identify repeat offenders on a regular basis. The police quickly adopted the anthropometric-based Bertillon identification system. This complex identification system, called Bertillonage, provided that the prisoner's personal details, his body measurements, any descriptions of physical anomalies, and photographs were reported on a single identification card (Figure 1.1). In the early 1900s, fingerprints overtook the Bertillon technique as the principal method of identifying criminals. Galton was interested in fingerprints primarily as a tool for detecting heredity and racial background. While he soon discovered that fingerprints did not provide conclusive evidence of an individual's intelligence or genetic background, he was able to scientifically prove what some

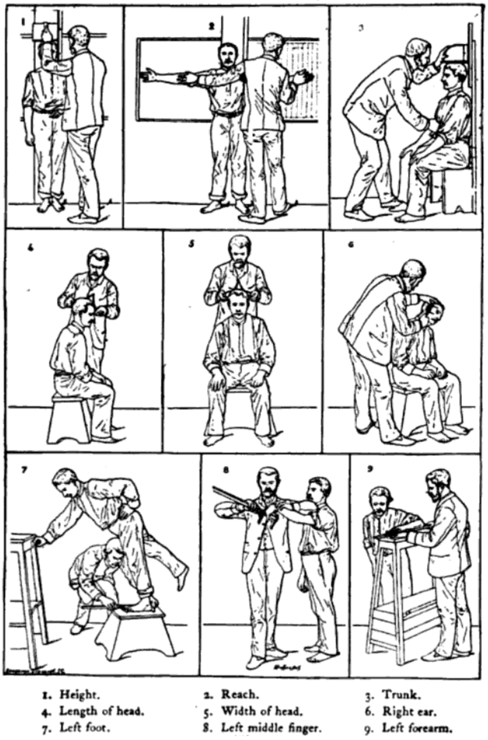


Figure 1.1: Anthropometric measurements in Bertillon's system of identification [22].

previous studies had hypothesized: fingerprints do not change over the course of a person's lifetime, and no two fingerprints are identical. Galton identified the characteristics used to identify fingerprints. These features (minutia) are essentially still in use today and are commonly known as Galton's Details (Figure 1.2). The creation of the Bertillon system and Sir John Galton's elementary fingerprint recognition system served as inspiration for the scientific community, which has dedicated its efforts to discovering numerous biometric modalities.

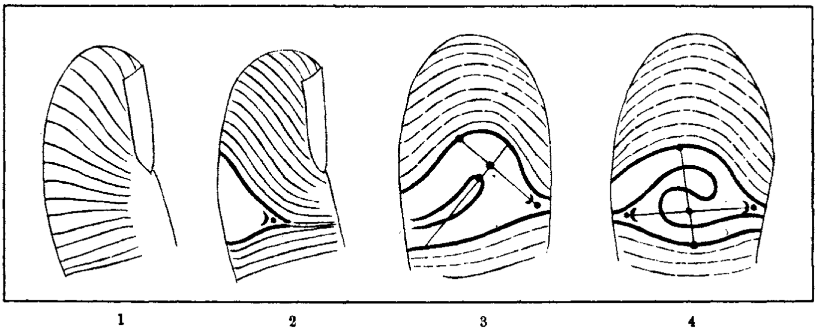


Figure 1.2: Francis Galton's diagram of primary patterns of the fingerprint [63].

Any physiological or behavioral attribute can be considered a biometric trait as long as it meets certain requirements [157], including: (*i*) *Universality*: possessed by all humans, (*ii*) *Distinctiveness*: discriminatory among the population, (*iii*) *Invariance*: the chosen biometric attribute must be time-invariant, (*iv*) *Collectability*: ease of collection in terms of acquisition, digitization, and extraction of features from the population, (*v*) *Performance*: refers to the availability of resources, imposing real constraints on data collection and ensuring high accuracy, (*vi*) *Acceptability*: the population's willingness to submit that characteristic to a recognition system. (*vii*) *Circumvention*: robust to imitation or mimicry in the case of fraudulent attacks on the recognition system.

A biometric system is a *Pattern Recognition* system that compares the discriminatory characteristics (probe) acquired during the enrollment/registration phase with the characteristics of a previous one (gallery). Each biometric system has the components indicated below. The Acquisition module collects the biometric trait and submits it to the system for processing. The Feature extraction module extracts the discriminatory features from the captured data to create a template. These templates compose the Database module, which includes digital representation of previously acquired samples. The Matcher module, as suggested by its name, matches the extracted features of the probe with those of the gallery to obtain a match score. Figure 1.3 shows the structure of a typical biometric system. A biometric system operates in one of the following modes: verification or identification. The technical framework for processing the feature models used in the comparison process is the cause. A 1:1 (one-to-one) comparison is used to verify identity. Using fingerprints as an example, a person's fingerprints are compared to a particular pattern that has already been registered on a specific medium. The identification is based on a 1:N (one-to-many) comparison, which means it tries to determine if any patterns in an existing fingerprint database match the one checked. While the second case aims at learning the identification, the first case seeks only to verify it. A biometric system's accuracy is typically measured in terms of performance errors. The False Accept Rate (FAR) is the percentage measurement of invalid matches. It measures how frequently the system identifies unauthorized users as genuine ones. This error must be as small as feasible in order to have a robust biometric system. The False Reject Rate (FRR) represents the percentage of times the system identifies an authorized user as an impostor. The point at which the FAR and the FRR are

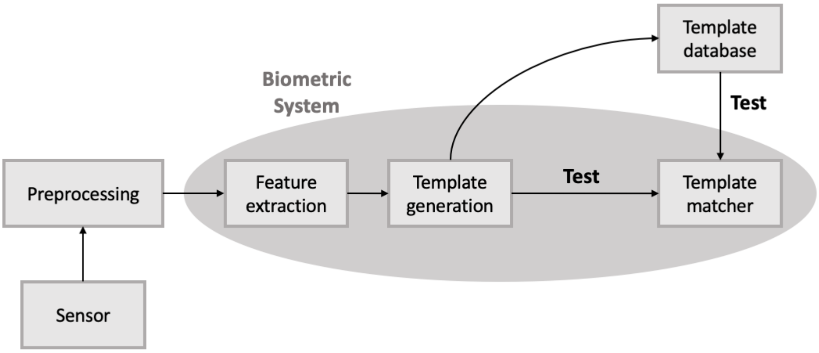


Figure 1.3: Composition of the biometric system.

equal to one another is referred to as the Equal Error Rate (EER). This is obtained through the Receiver Operating Characteristic (ROC), which involves a trade-off between the false acceptance and reject rates and a plot of FAR against the FRR. The ERR measures the accuracy of a biometric system.

The growing interest in biometric-based identity management systems by public and private entities is evidence of their superiority over knowledge-based and possession-based systems. Statistics on existing and projected revenues from the global biometric technology market are a source of proof. By 2030, the biometric systems market will be valued at $132.5 billion, up from $51.55 billion in 2022 (See Figure 1.4). A paradigm shift in business discourse towards greater privacy and fewer security risks is one of the major trends observed in the market for next-generation biometrics. Instead of relying on conventional techniques, end customers are continuously looking for integrated solutions. In industrialized nations, advanced biometric identification solutions are used to implement a variety of government programs, including the use of electronic passports, electronic driver's licenses, border control, and national identity documents. The low cost of biometric technology allows manufacturers to use it in a wider range of products, including biometric sensors that are becoming an increasingly common security feature in smartphones and other devices. According to a recent study by Visa [185], 86% of consumers expressed interest in using biometric data to verify identity or make payments; 70% of those who have used biometrics said it is easier; and 46% believe biometric technology is more secure than passwords or PINs. Therefore, the biometric market is set to grow in the coming years as biometric technology is progressively used in consumer electronics for authentication and identification purposes.

1.2 BEHAVIORAL BIOMETRICS: AN EMERGING TREND

Behavioral and biometrics are the origins of the term "*Behaviometrics*". Behavioral biometrics pertain to the personality and behavior of an individual. Behaviometrics, or behavioral biometrics, is a measurable behavior utilized to recognize or authenticate an individual. Humans learn primarily through their sensory

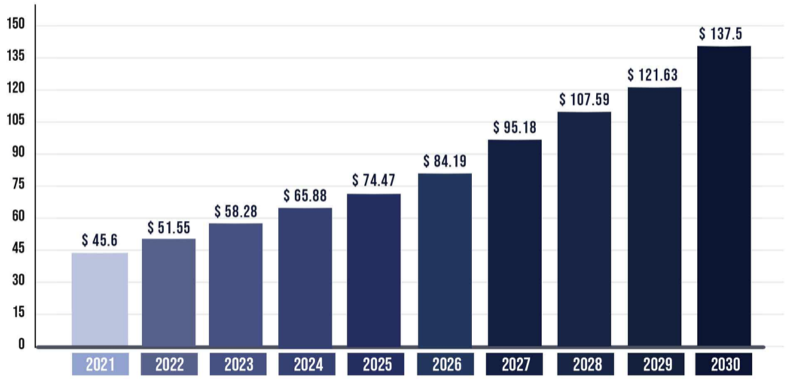


Figure 1.4: Biometric technology market size, 2021 to 2030 (USD Billion). Source: [112].

systems and direct experience. Humans' innate (inherited) and acquired (learned) behavioral skills have both similarities and differences. A behavior is a particular way of acting. Humans are born with such behaviors and instinctively know what to do in specific situations or conditions. In contrast, acquired or learned behaviors are not inherited but rather gained via experience. Behavioral biometric traits are completely reliant on the behavioral nature of human beings. It assesses human behavior, which does not directly focus on measurements of body parts. Few behavioral biometrics have such a direct correlation with human physiological traits. The connection between the majority of behavioral biometrics and human physiology is more complex and often indirect. For instance, keystroke dynamics, i.e., the way of interacting with a keyboard (typing rhythm), is connected not only to how we use our hands to type on a keyboard but also to the brain (learning and memorization) [153]. A human behavioral pattern is comprised of several different behaviors that are merged into a larger and more distinctive profile. Because a person's unique behavioral pattern is produced not just by biological characteristics but also by social and psychological factors, it is hard to imitate another person's behavior. Based on the type of information about the user being collected, behavioral biometrics can be classified into several categories [132]. The ability to use muscles is directly related to a person's motor skills. The definition of motor skill based on behavioral biometrics, as "kinetics", represents the user's distinctive and consistent muscular actions in carrying out a given task. Therefore, motor skills reflect the effectiveness of the functioning of such systems in an indirect way, allowing the verification of the person. Most motor skills are learned, not inherited. Authorship-based biometrics, for example, analyze a text or a drawing produced by a person, identifying the stylistic peculiarities of the author of the work in question. Last but not least, biometrics based on Human-computer interaction (HCI) cover events that can be collected by observing user behavior directly or indirectly with the aid of technological devices. In particular, indirect HCI-based biometrics encompass the events that can be collected by indirectly monitoring a user's behavior via observable low-level actions of device software. Direct HCI-based interaction is classified into two classes. Human interaction with input devices such as keyboards, touchscreen devices, etc. comprises the first category. The second category describes HCI-based behavioral biometrics, which evaluate sophisticated human behavior such as strategy, knowledge, or skill exhibited by the subject when interacting with various software applications.

Nowadays, behavioral biometrics are widely used in the context of information security to identify individuals using the distinctive characteristics of the activities they do, knowingly or unknowingly. Most of the existing behavioral traits involve voice and signature scanning, walking gait, keystroke and mouse dynamics, touch gestures, and, generally, people's movements (for example, the way a person is moving his or her head, facial parts, etc.). Researchers have recently developed approaches for speaker recognition by monitoring lip movements, performing biometric verification using finger motions, and extracting speech features for person identification. With the increase in mobile device usage, a new form of behavioral biometrics has recently been introduced. The usage data of a mobile device can be viewed as a unique profile due to the fact that people use their devices to engage with applications and digital services in a specific pattern. A user's behavioral profile can be constructed based on his interactions with either a network or a host. In the first scenario, user behavior is observed based on their patterns of connecting to Wi-Fi networks, service providers, etc., whereas in the second case, user behavior is monitored based on the manner in which applications are utilized at various locations and times [142]. Behavioral patterns do not disrupt normal workflow, unlike many popular biometric solutions that require the user to perform additional tasks. Numerous advantages related to applying behavioral biometrics are related to data acquisition, as it does not require specialized or dedicated hardware. Consequently, it is also considered cost-effective. The majority of behavioral features are acquired through simple, device-based interactions. Furthermore, the data acquisition is totally transparent for the user and does not involve delays in operations. For these reasons, biometric systems based on behavioral models are widely accepted in society. Although behavioral biometrics alone are not sufficiently unique to identify a person with high precision, they can obtain a high verification rate and enhance the recognition rate as part of a multimodal biometric system.

1.3 ARTIFICIAL INTELLIGENCE AND BIOMETRICS

Since John McCarthy originally introduced the term Artificial Intelligence (AI) at the Dartmouth Workshop in 1956, numerous definitions have been proposed. One of the most significant and comprehensive was actually devised by himself in 2004. He claims that Artificial Intelligence "*is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable*" [114]. In computer science, AI is defined as the study of "intelligent agents", which are devices that "perceive their environment and take actions to maximize their chance of success at some goal". Colloquially, the term AI is used when a machine imitates "cognitive" processes that people typically associate with other human minds, such as "learning" and "problem solving" [120]. Depending on the input data, AI systems can detect, reason, and act. ML and DL are two concepts closely related to the idea of Artificial Intelligence. These are subfields of AI (in fact, DL is itself a subfield of ML) and both rely on pattern extraction to predict or classify data. However, they show some differences in the learning operations of their algorithms [73].

As Arthur Samuel defined it, "*The field of study that gives computers the ability to learn without being explicitly programmed*" is Machine Learning (ML) [133]. ML is a branch of Artificial Intelligence that investigates the development of algorithms capable of learning and predicting data using a series of training examples. Nowadays, ML is a widely adopted technique in the field of data analysis, which allows researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and discover "hidden insights" by learning from historical relationships and trends in the data. In contrast, the concept of Deep Learning (DL), first introduced by Hinton [74], involves the use of Artificial Neural Network (ANN) models that mimic the structure and function of the human brain, proving to have strong abilities in detection, classification, segmentation, and key point estimation. It generally necessitates stacking multiple layers of learning algorithms to approximate highly nonlinear functions. This allows DL algorithms to learn hierarchical representations/features from the data. Lower-level layers learn simple characteristics, while higher-level layers learn increasingly complex features consisting of lower-level characteristics. This allows both local and global properties to be encoded in the final feature representation. In numerous fields, including computer vision and natural language processing, this feature learning has largely replaced hand-engineered features. The learned representations are distributed because a single factor can be explained by multiple neurons, and a single neuron can help explain multiple factors. This many-to-many interaction produces compact, dense representations that may generalize nonlocally [148]. Convolutional Neural Network (CNN) for automatic feature extraction and Recurrent Neural Network (RNN) for sequence estimation are two well-known categories of DL architectures.

Nonetheless, it is important to remember that DL is a subset of ML, and hence both can use labeled and unlabeled data. Consequently, the learning process depends on the specific problem to be solved and the data structure. Three groups can be distinguished [171]:

- **Supervised learning**: examples of the input vectors and associated target vectors are included in the training data. Learning activities are defined as classification or pattern recognition when the target vectors are categorical and regression when the target vectors are real values. Consequently, using the training data they have already learned, the algorithms are trained to classify or predict the outcome of new, unseen data.

- **Unsupervised learning**: the data is not labeled. Consequently, the algorithms must identify patterns in the hidden data and extract them, classifying the information according to the found characteristics. Since the input is unlabeled, the classification in this case depends on the structures identified by the model. Clustering algorithms are the most frequently used unsupervised learning methods.

- **Reinforcement learning**: this approach aims to create autonomous agents able to choose the actions to be taken to achieve certain objectives through their interaction with the environment in which they are immersed. Unlike the other two (i.e., supervised and unsupervised), this paradigm deals with sequential decision problems where the action to be performed depends on the current state of the system and determines its future state. The quality of an action is given by a numerical value of "reward" inspired by the concept of reinforcement, which aims to favor the correct behavior of the agent.

The application of AI techniques to explore large amounts of heterogeneous biometric data and provide user authentication and identification capabilities shows great potential. In this regard, since biometrics involves identifying individuals based on their features, it mostly requires supervised learning. In recent years, DL approaches for automatic feature extraction and description have become increasingly popular. DL methods have an edge over previous state-of-the-art techniques due to their ability to learn features from data. For example, biometric modalities such as face and voice require both local and global features and are suitable for hierarchical and compositional feature learning. Additionally, hand-crafting features for some modalities, such as behavioral biometrics, can obtain abstract and learning features from raw data that will be useful for such purposes. With a particular focus on the field of behavioral biometrics, due to growing security and privacy concerns, researchers are exploring user behavior-based implicit authentication. ML-based systems are, in fact, able to learn from human behavior, continuously generating unique user profiles and discovering behavioral patterns that persist over time - a fundamental approach to supporting continuous authentication systems [95].

Currently, Internet of Things (IoT) technology is able to permeate every area of our daily lives due to recent technological advances. Internet of Things (IoT) extends the Internet's capabilities to a vast range of devices that can be adopted in a variety of fields, including but not limited to smart homes, smart cities, environment, agriculture, smart grid, industry, healthcare, and transport [94]. However, the limits of low power and computational processing prevent the implementation of advanced security policies on IoT devices, particularly in terms of privacy, authentication, communication encryption, and data storage protection. In this regard, biometrics offers an intriguing opportunity to improve the usability and security of IoT, playing an essential role in safeguarding a large array of smart devices. Recently, there has been growing interest in Internet of Biometric Things (IoBT) applications based on DL. AI-related technology is advancing, which fosters the development of many disciplines. Nonetheless, the proliferation of smart devices around us presents a number of challenges for classical biometric techniques, demanding the development of new mechanisms to suit the dynamic nature of the smart environment [20].

1.4 CONTRIBUTIONS

Human behavior is the potential and expressed capacity (mentally, physically, and socially) of human individuals or groups to respond to internal and external stimuli throughout their live. The behavior of an individual is influenced by genetic and environmental factors, including thoughts and feelings, which provide insight into the individual psyche. Personality types differ from person to person, resulting in uniquely distinctive actions and behaviors. Recently, behavior-based facial biometrics, such as head dynamics and facial expressions, have gained a lot of interest in literature. The term Head Pose, which actually refers more specifically to face orientation in 3D space, represents a very popular key-word due to the vast number of related research topics and applications. It is a fact that determining how the human face is rotated with respect to an imaging sensor may provide crucial information for many tasks such as face recognition and analysis, body tracking, face frontalization, and the estimation of a subject's intention, just to name a few. The first studies were carried out under controlled conditions, e.g., faces were captured with uniform illumination, a frontal pose, and a neutral expression. However, to target real-world applications, the research has to face ever more challenging issues where the detection and recognition of a specific trait, in particular a behavioral characteristic, are usually affected by critical factors, like uneven illumination, natural and/or artificial occlusions, or self-occlusions. These factors play an especially important role when dealing with unattended acquisition. A relevant example is found in video surveillance, involving either partially or even totally unaware subjects. Not surprisingly, then, head pose estimation represents a hot research topic that has been approached by a large number of methods. Based on these assumptions, the methods presented here have shown promising results.

Human behavior is also greatly influenced by the emotional state of the subject. The recognition of facial expression and its intensity is a key component of behavioral biometrics. With the increased use of images over the last decade, automated facial analytics such as facial detection, recognition, and expression recognition have gained considerable importance. Facial expression recognition is being used in a variety of real-world applications in which a person's emotional state serves as a cue for the successful operation of these systems. Security and video surveillance systems, human-computer interface design, emotive marketing, and smart healthcare are only some of these. The criterion for the accuracy of intensity detection of the seven observed fundamental emotions is based on the analysis of facial behavior components pertinent to emotional intensity communication. This involves detecting the face and recognizing the intensity of emotion represented, both of which have been extensively investigated in literature. Currently, DL algorithms, particularly CNN architectures, are achieving promising results. Nonetheless, certain issues still persist. As an example, in real-world scenarios, the system should be able to recognize emotion from diverse facial angles. In addition, in unconstrained imaging environments, the images may suffer from various noise artifacts. So, accepting the current challenges and motivations behind this research, the framework presented yielded significant results.

Human behavior recognition has been considered a core method of user authentication. As smartphones have become an important part of our daily lives, the industry and academic communities have asserted that touch-based gestures can be used to uniquely identify a person. For this reason, in the context of human-mobile device interaction, user touch behavior patterns were investigated. Touch dynamics refers to the process of measuring and evaluating the characteristics of touch gestures that users perform on mobile devices such as tablets, touch panels, smartphones, etc. Like all behavioral biometrics, its use becomes particularly interesting when it is included in a multi-factor solution to increase the trust level of the system. The advantages of using touch dynamics as a biometric trait are numerous, one of which is certainly the non-intrusiveness, meaning that the user is free to approach the personal device in the most natural way possible while the biometric behavioral data is captured. Over recent years, the research efforts on the identification of so-called soft biometrics, such as age, gender, ethnicity, degree of confidence with a certain hardware, and so on, have paid off with interesting results and the definition of potential application fields. Some of the areas that could benefit from an ad-hoc analysis of behavioral data from mobile devices include targeted advertising based on the user, easier interaction with specific hardware, and the fusion of other information for better identification or verification of a subject. Based on the above, we have employed user touch-interaction behavior data for subjects' demographic classification based on soft biometric traits.

1.4.1 *Outline of the Thesis*

The thesis is structured as follows. The present Chapter introduces the principles of biometrics and our motivations and contributions. The next three Chapters define the core of the thesis:

- Chapter 2 is focused on Head Pose Estimation (HPE). We present different approaches to estimating the actual head orientation from 2D images by using fractal coding theory and, particularly, Partitioned Iterated Function Systems. We also develop a unified method for face recognition and HPE that uses the same fractal encoding features for both tasks.
- Chapter 3 is focused on Facial Expression Recognition (FER). Although FER uses multiple sensors, we limit our discussion to exclusively using static images, with a focus on recent DL-based FER systems. Therefore, we develop a CNN architecture to categorize the principal facial expressions. Here we also investigate the impact of facial expressions on HPE.
- Chapter 4 is focused on Touch Dynamics (TD)-based behavioral biometrics, that captures a person's typing patterns on mobile touchscreen devices. In particular, we explore the integration of soft biometric traits with touch-interaction behavior data. Our goal is to demonstrate how soft biometric analysis can be used to achieve lightweight continuous verification and improve the identification mechanism.

In the concluding Chapter (Chapter 5), we will draw our conclusions and also propose some ongoing research and future advances regarding the methodologies just presented.

2 HEAD POSE ESTIMATION: A MULTIFUNCTIONAL BIOMETRIC

The ability to determine the orientation of a person's head is called HPE. It can be used for a variety of purposes, such as a preprocessing phase to determine the optimal frame for face recognition in a video, a behavioral characteristic to determine the subject's intent, a descriptor to aid in face frontalization, and so on. The study of HPE represents a subset of the wider biometrics field. In order to develop a biometric system, the trait that should be used is mostly determined by the availability of processing resources and its applicability in terms of visible area. In this context, the advent of behavioral and soft biometrics has paved the way for the application of alternative biometrics. In fact, HPE can be applied to both behavioral and soft biometrics.

In this Chapter, we discuss the multitude of strategies and advancements in HPE as well as the impact of recent techniques such as ML. The candidate's research over the last three years has been strongly related to HPE, making a positive contribution to the state-of-the-art.

2.1 HPE DOMAIN

In terms of uniqueness, HPE shows a strong correlation to behavioral biometrics, exploring the distinctive properties of an individual's head movements. The head pose performs better in terms of measurability than the majority of behavioral biometric traits, where a subjective component may be seen. Each person has a head rotation that can be observed and is age-independent. Because they are the only identification traits involved, the acceptability of HPE can be compared to that of the face or, at most, the ear. However, depending on the specific characteristics of the technology being utilized, the shape of the head and face, which varies from person to person, can affect HPE and its applicability.

The main procedure to build a HPE technique, from the data to the evaluation of the errors, can be summed up as follows:

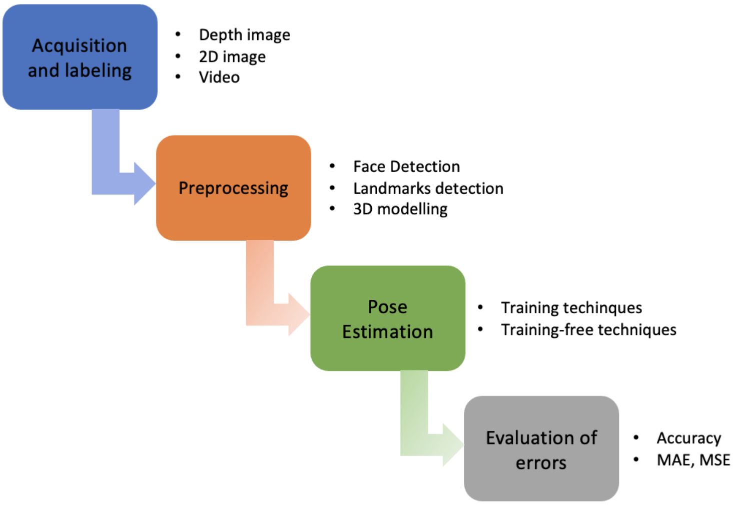


Figure 2.1: The main steps of an HPE framework.

- **Acquisition and Labeling**: The involved trait is acquired and categorized at this step. The same methods and tools used for face acquisition, such as cameras, depth cameras, near-infrared cameras, etc., are generally used. As a result, the acquisition process is relatively simple; however, labeling is a more difficult operation. In addition, since rotation angles are not human observable if they are small, it is challenging to estimate them manually in the absence of depth data. Therefore, it is not advised for humans to participate in this task.

- **Preprocessing**: The data must be normalized at this phase in order to be ready for the HPE model. It differs significantly between various architectures.

- **Pose estimation**: This is the fundamental component in which the image, or the preprocessed data, is used to estimate the head pose. The final result will always be the rotation in terms of angles, regardless of the input.

- **Evaluation of errors**: The accuracy is typically shown in this field as the angular difference between the true label and the estimate.

The procedures just described are summarized in Figure 2.1. Rotation angles are used to measure the variation in head pose. The $O(0,0,0)$ center point of the rotation is the center of the head, or, in the absence of 3D data, the nose. The head is a three-dimensional entity, so there are three possible angles of rotation. The Motion Imagery Standards Board (MISB) [108] traditionally refers to the axes as *pitch*, *yaw* and *roll*. The head's degrees of freedom are illustrated in Figure 2.2. Using the frontal view as a reference point, the majority of individuals can turn their heads $\pm 90^\circ$ in yaw, $\pm 45^\circ$ in roll, and $\pm 30^\circ$ in pitch. Since facial recognition is often the primary focus of HPE, severe head postures produced by body motions are rarely considered. Even though there are numerous ways to describe a 3D rotation, the Euler angles, the rotation matrix, and the quaternions are the most popular representations in HPE databases and algorithms.

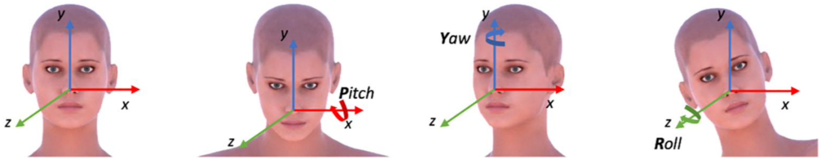


Figure 2.2: The Pitch, Yaw and Roll axes.

Leonhard Euler devised two types of Euler angles - proper Euler angles and Tait-Bryan angles - to characterize the orientation of a rigid body in space. As previously introduced, the MISB rules are followed during the head pose rotation. The Tait-Bryan angles are assumed to describe the rotations. After the rotation, X, Y, and Z are the axes, whereas x, y, and z are the original axes. The intersection of plans xy and YZ defines the line of nodes N. These conventions allow the following formulation of Euler angles:

- ϕ the rotation angle between x and N, covering a range of 2π.

- θ the rotation angle between z and Z, covering a range of π.
- ψ the rotation angle between N and X, covering a range of 2π.

Using the rotation matrix, which can define three rotation matrices, one for each axis, one rotation angle θ can be used to calculate the rotation with respect to the axis. If we define the rotations in yaw, pitch, and roll as α, β, and γ, respectively, the final rotation matrix will be:

$$\begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\gamma & \cos\beta\cos\gamma \end{bmatrix} \tag{2.1}$$

Since using this representation might not be convenient, the Rodrigues' formula can be used to turn the rotation matrix into a rotation vector:

$$v_{rot}(\theta) = v\cos\theta + (k \times v)\sin\theta + k(k \cdot v)(1 - \cos\theta) \tag{2.2}$$

Where v is a 3-D vector, k is a unit-vector indicating the axis around which v rotates by an angle θ according to the right hand rule, $k \times v$ is a cross product, and $k \cdot v$ is the scalar product.

Last but not least, William Rowan Hamilton introduced the quaternions, also referred to as versors, which are currently very popular among game creators. We can introduce the generic form for expressing quaternions based on the concepts of complex number and complex plane.

$$q = s + xi + yj + zk \tag{2.3}$$

where $s, x, y, z \in \mathbb{R}$ and i, j, z are imaginary number that follow the rules:

$$i^2 = j^2 = k^2 = ijk = -1 \tag{2.4}$$

and

$$ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j \tag{2.5}$$

The Euler angles are more "self-explanatory", but they can lead to ambiguity issues. This implies that for the same rotations that result in an obviously problematic estimate problem, we will have an endless number of pose estimation solutions. Quaternions, on the other hand, are easier to construct and do not have this issue. Their representation is more compact than the rotation matrix. Since different datasets may have different annotations for the angles, the testing methods often select a representation and normalize the dataset's label in accordance using transformation formulas. In contrast, there is consistency in the evaluation of errors. The angular values of the differences between the estimated pitch, yaw, and roll and the true pitch, yaw, and roll for each head in the data represent the error predictions. To achieve valuable results, the algorithms measure these three errors, and then they compute the Mean Absolute Error (MAE) for each axis as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |\theta_j - \hat{\theta}_j| \tag{2.6}$$

where θ_j is the ground truth, i.e. the true angular value and $\hat{\theta}_j$ is the prediction, i.e. the predicted angular value. An overall MAE of the error along the three axes is also computed.

2.2 DATASETS WITH HP ANNOTATIONS

To perform HPE, there are primarily three types of input data: *depth images*, *2D images*, and *video*. Depth image datasets include both RGB and depth information from the same image. Many datasets have been proposed in the last decade that are helpful for evaluating the applicability of HPE techniques. The BIWI Kinect Head Pose Database (BIWI) [59] is without a doubt one of the most well-known depth datasets. BIWI includes 24 sequences of 20 individuals, totaling over 15 K images captured by a Kinect 1. The yaw and pitch variations of the head pose are ±75° and

±60°, respectively.The head postures have been annotated using Faceshift. ICT-3DHP [11] is a database collected with the Kinect and is made up of 10 RGB-D videos for a total of roughly 1,400 frames. A Polhemus FASTRACK was used to obtain the labels. With the new Kinect 2, data from the SASE benchmark [106] was gathered. There are 50 total subjects, and each subject has an average of 600 frames. They use five blue stickers applied to the face to provide pitch, yaw, and roll. SASE has a yaw range of ±75° and a pitch range of ±45°. Over 10 K images and 20 subjects are included in the ETH Face Pose Range Image Dataset [28]. The 3D nose tip coordinates and the coordinates of a vector pointing in the face direction serve as the ground truth. ETH has a pitch range of ±45° and a yaw range of ±90°. The Kinect 1 was used to capture images for the Pandora dataset [27]. The set of 22 subjects contains over 250 K frames, with each subject having 5 recordings.

Despite having reliable labels, depth datasets are not the preferred input for developing and testing HPE techniques. This is due to the fact that depth images require controlled environments, whereas HPE approaches employing only 2D RGB images are designed and evolved to address in-the-wild problems. The datasets without depth information and how they were labelled are presented below. A stereo camera approach is used to estimate the head pose on the RGB images in the CMU-MultiPIE dataset [69]. With more than 750 K images of 337 subjects in 13 positions, 4 recording sessions, and 6 facial emotions, this dataset is an expansion of the earlier CMU-PIE [140]. The PRIMA Lab created the Pointing'04 Head Pose Image Database [68] with 15 subjects. There are two series of 93 pictures for each subject, for a total of 2,790 images. The roll angle rotation is not included. The dataset contains only a few poses (9 for pitch and 13 for yaw) and their combinations between ±90°. The authors asked people to stare at the 93 post-it notes without moving their eyes in order to capture positions with known labels. The Annotated Facial Landmark in the Wild (AFLW) dataset [87], which consists of roughly 25 K images collected from the web, has a wide range of poses, expressions, ages, genders, and ethnicities. AFLW2000, which includes the first 2000 pictures of AFLW annotated using a 3DMM fitting and can be obtained at [60], is a more accurate version of AFLW. The Flicks images that make up the Annotated Face in the Wild (AFW) dataset [182] were gathered. There are just 205 images totaling roughly 468 faces in the collection, which is extremely little. The 300W_lp Dataset [184] is an expansion of 300W, which contains 68 landmark localizations. The datasets that 300W_lp gathers include AFW, LFPW, HELEN, IBUG, and XM2VTS. There are 61,225 images total, including 17,860 from IBUG, 5,207 from AFW, 16,556 from LFPW, and 37,676 from HELEN. The CAS-PEAL database [64] contains 99,594 images of 1040 different subjects. In all, 27 distinct poses are present in a controlled environment. 3425 YouTube videos of 1595 participants are included in the Youtube Faces database [164]; over 600 K extracted and annotated frames are available. Faces are identified with the Viola-Jones method, which we will discuss in Section 2.3. The McGill real-world face video database [43] is a collection of video sequences for investigating the problem of unconstrained face classification. This database comprises 18,000 frames from 60 video sequences, each of which was captured from a different subject (31 females and 29 males). The participants' movements were completely free, resulting in arbitrary face scales, facial expressions, head poses (in yaw, pitch, and roll), motion blur, and local or global occlusions. Finally, GOTCHA-I [16] is a recent multiview video dataset containing video from cooperative and non-cooperative subjects. GOTCHA-I consists of 682 recordings of individuals walking in different areas made by 62 subjects in 11 different locations. There are 137,826 labeled frames with 2223 head poses per subject. In Table 2.1, an overview of depth and 2D datasets with head pose annotations and key information is provided.

The application of HPE to videos has the goal of using several frames to comprehend the user's behavior. Actually, in the HPE domain, video datasets have not gained a lot of popularity. The majority of the benchmarks described below are used for HPE tracking, which necessitates a few key properties. There are 120 videos of 10 different subjects in the UPNA Head Pose Database [9]. The authors gathered six guided-motion sequences and six free-motion sequences because this dataset was designed for head tracking and pose estimation. Each video has 300 frames. They identified the head position by using the frontal face's

Table 2.1: Depth and 2D RGB datasets that contain pose annotation.

Dataset	**Year**	**Type**	**#Subj**	**#Frames**	**Limit**
ETH	2008	Depth+RGB	20	+10 K	\
ICT-3DHP	2012	Depth+RGB	10	1400	\
BIWI	2013	Depth+RGB	20	+15 K	\
SASE	2016	Depth+RGB	50	+30 K	\
Pandora	2017	Depth+RGB	22	+250 K	\
Pointing'04	2004	RGB	15	2790	No roll
CAS-PEAL	2008	RGB	1040	+99 K	27 poses
AFLW	2011	RGB	20	25K	\
Youtube Faces	2011	RGB	1595	+600 K	\
AFW	2012	RGB	nd	205	No pitch
McGill	2013	RGB	60	+18 K	No pitch No roll
CMU MultiPIE	2013	RGB	337	+750 K	15 poses
300W_lp	2016	RGB	nd	+61 K	
AFLW2000	2018	RGB	nd	2000	\
Gotcha-I	2020	Video	62	+137 K	\

nd = "not declared".

initial frame as a reference point. The Boston University Head Pose Database [88] is composed of 45 video sequences in which 5 subjects were instructed to perform 9 different head movements under uniform illumination in a standard office setting. The head is constantly visible, and there are no occlusions except for small self-occlusions. The Head Pose and Eye Gaze Dataset (HPEG) [10] is a collection of 20 videos with 10 people that were produced in lab conditions. Each video contains roughly 400 frames. Three LEDs are used to track their positions in each frame in order to gather the ground truth. Only yaw and pitch angles are accessible in this dataset. The EYEDIAP Database [62] was created with the intention of tracking gaze. They captured 94 sessions with 16 participants while simultaneously using the Kinect and an HD camera that was positioned as close to the Kinect as possible. 4860 frames or so make up each video. Also, the labels for the UBIPose dataset [118] were obtained using a Kinect; only 22 of the 32 videos include annotations. There are roughly 10 K frames available for the head pose. Finally, the second Strategic Highway Research Program (SHRP2) [30] is a video dataset made up of subjects driving. The database is fairly large, containing more than 3100 videos of the same number of participants shot over the course of two years. Nevertheless, only 63 K out of 41 frames include head pose annotations. As previously introduced, the labelled frames are approximately 1537 per video; however, they are each about 15 minutes long and captured at 15 frames per second.

2.3 PREPROCESSING TECHNIQUES

The head region or some keypoint on the face is typically detected using several preprocessing techniques. These methods can be divided into three major categories: *face detection*; *landmark detection*; *3D head modeling*.

Most HPE algorithms use face detection as a preliminary step to exclude from analysis other parts of the body or the scene. The Viola-Jones face detection method has been a prominent technique for a long time [160]. However, this approach would fail to locate the face if the head posture was extreme (greater than 60°). For this reason, the Histogram of Oriented Gradients (HOG) and linear Support Vector Machine (SVM) combination presented by Pang et al. [122] was favored first, followed by learning methods such as the training of random forests [105] or deep architectures like ArcFace [45]. In recent years, the effect of such existing detectors in the case of facial masks has also been of major relevance [81]. Face detection can be used in conjunction with or as a substitute for landmark detection. Facial keypoints are the positions of certain particular facial characteristics, such as the mouth, eyes, and nose. Some authors developed adaptive boosting methods to locate these keypoints, while others used SVM or the aforementioned random forest classifier. The application of DL has gained popularity in this field [42]. Landmark detection produces on results similar to the ones in Figure 2.3. The data is expressed in terms of spatial coordinates. For an in-depth investigation of landmark detection, refer to Wu and Ji [166]. Finally, 3D modeling is a technique that aims to build a 3D face model in order to determine the position of the head. In this situation, having access to a depth image may be essential to building a realistic model [39, 125, 138].

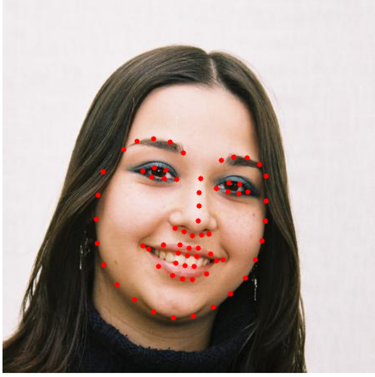


Figure 2.3: 68 landmarks detected on a 2D RGB image [84].

2.4 METHODS

The most difficult and dynamic field of HPE application is 2D RGB images. This is why it is important to distinguish the different methodologies used given the numerous works that have been published in the last five years. The use of training techniques is the primary distinction that can be made [4].

2D RGB HPE USING TRAINING-FREE TECHNIQUES The procedures utilizing training are superior to the techniques using training-free methods. In training techniques, a part of the test data from the same dataset is further altered to produce a trained neural network, whereas training-free methods use certain images as just a reference for features. The quad-tree based technique is the core of HPE in [17]. After face detection, the method continues to work only on landmarks. The same procedure is carried out for images of a reference synthetic model, and pitch and yaw angles are detected by comparing the trees in binary vector form. This work has been improved in [1], adopting a more accurate reference synthetic model. Additionally, it presented a YouTube video's best frame selection experiments for locating non-frontal faces. In [13] the authors first detect the positions of 68 well-known facial landmarks and, after that, apply a web-shaped model over the detected landmarks to associate each of them with a specific face sector. AFLW2000, BIWI, and Pointing'04 have all been used to test this strategy. Only three keypoints are detected in [119], and the emphasis is on mobile device applications. The experiments, however, were conducted using a few smartphones and a homemade dataset, so they cannot be compared to state-of-the-art methods. Peng et al. [124] tackle the 3D HPE problem by taking advantage of a characteristic of 3D spheres during rotation. A technique known as "Homeomorphic Manifold Analysis" serves as the foundation for the 3D sphere's function as a model of the head's potential rotation. Proença et al. [125] do something similar by estimating the head pose using geometric properties on a synthetic model. Projective geometry is used in the synthetic model to connect the images 2D points. The set of landmarks in the model that produce results that are closest to the input is chosen using convex energy minimization techniques. The work in [131] represents a methodology using feature-based techniques. In this study, a similarity kernel that is identity-invariant is learned by utilizing the feature correspondences of Geometric Blur. The difficulty of identifying hidden head positions in [134] enhances the difference with a classical HPE using training. Here, the posture angle of a facial image is represented by a multivariate label distribution.

2D RGB HPE USING TRAINING TECHNIQUES Recent literature has, as stated, focused more on techniques based on training. In order to enhance performance over the BIWI, AFLW2000, and Pointing'04 datasets, various regression techniques are evaluated. Following this direction, the features retrieved using the method [13] are combined with regression in [2]. The method in [49] formulates the head pose as a high-dimensional to low-dimensional mixture of linear regression problem. Liu et al. [101] propose a multi-level structured hybrid forest (MSHF) for joint head detection and pose estimation. Regression is used in a completely different way in Cao et al. [34], where three vectors in a rotation matrix as the representation in HPE are used and develop a new neural network based on the characteristics of such a representation. When the issue involves low-resolution images, regression seems to work especially well. The HOG features are merged with non-linear regression in Chen et al. [40], just like in the earlier technique. Here, the Support Vector Regression (SVR) is specifically trained using incredibly low-resolution images. The authors also offer enhancements based on depth information. This is why BIWI was used as the dataset; HOG was also used by Diaz-Chito et al. [48]. This approach, which combines continuous local regression, generalized discriminative common vectors, and HOG, is based on manifold learning-based methods. In [50] a mixture of linear regressions with partially-latent output is introduced. This regression method learns to map high-dimensional feature vectors (extracted from bounding boxes of faces) onto the joint space of head-pose angles and bounding-box shifts. In Alioua et al. [6], a novel descriptor resulting from the fusion of four most relevant orientation-based head descriptors is proposed. [161] is another tree-based algorithm that employs HOG. Presented here is a system known as Stacked Auto Encoder with Extreme Gradient Boosting (SAE-XGB). In contrast to other studies that extract features through a separate procedure, Liu et al. regression approach [96] is backed by a synthetic model. There are 37 head models in use, and there are 74K frames that represent the poses. A mixture of linear inverse regression is utilized by Lathuiliére et al. [89]. In particular, they propose the coupling of a Gaussian mixture of linear inverse regressions with a CNN (ConvNet), describing the methodological foundations and the associated algorithm to jointly train the deep network and the regression function. Multi-regression is also the strategy of Hsu et al. [75]. Yang et al. [173] propose another regression-based technique that is very recent. The method combines feature aggregation methods and soft stagewise regression. In [128], the image intensity is used in a multi-loss network with a classification and regression component, using a different loss function for each angle. In contrast to the preceding approach, [71] and [33] analyze the relationship between keypoints to increase accuracy. The work in [126] uses the method of locating faces and landmarks as additional features. This approach uses CNNs to simultaneously perform face identification, landmark localization, pose estimation, and gender recognition. In Xia et al. [167], the landmarks detection is likewise offered as a support for CNNs.

While using face landmarks in the HPE process could help increase accuracy on the one hand, it can also have a negative influence on the computational time required. The goal of [162] is to obtain a very quick technique because the use of HPE is concentrated on video. According to the experiments conducted, the method only needs 21.8 ms to obtain the HPE on a standard device. The 3D mean of landmark placement, specifically with 10 landmarks, is used to estimate the head posture. Shao et al. [136] investigate proper face localization and its implications for HPE by proposing a new loss approach in CNN. The application of ResNet is also discussed by Rieger et al. [127]. In Li et al. [91], a completely different strategy is suggested. The proposed method integrates the deep Task-Simplification oriented Image Regularization (TSIR) module with the Anchor-Guided Pose Estimation (AGPE) module, and formulate the HPE problem into a unified end-to-end learning framework. The driver's attention is highlighted in [143]. However, this time around, an object detection algorithm SSD which has inherent capabilities of simultaneous classify and regress, is used to create a lightweight network. The usage of conditional random fields (CRF) in [86] represents the method's fundamental component. The model trained assigns probabilities to each segmented face portion in order to classify each image in input. The goal of [115] is to make the approach insensitive to external factors. Here, the images are transformed into one-dimensional vectors as time series using the Peano–Hilbert space-filling curve. The multivariate label technique was changed in [169] to address issues resulting from its use in unconstrained environments. To prevent overfitting, they added regularization terms to the loss function using the weighted Jeffreys divergence. In [100], the authors propose a deep convolutional neural forests to handle occlusions and low image quality (D-CNF). In [102], the shortage of training data for many poses is tackled as a label distribution learning problem. They consider each face image as an example associated with a Gaussian label distribution, as opposed to a single label, and trained a CNN with a multi-loss function to predict facial poses directly from color images. Zhang et al. [175] proposed a three-branch network architecture, termed as Feature Decoupling Network (FDN), a powerful architecture for landmark-free head pose estimation from a single RGB image. Finally, Valle et al. [158] suggested a network architecture (an encoder-decoder CNN with residual blocks and lateral skip connections) and training strategy that harness the strong dependencies among facial posture, alignment and visibility, to produce a top performing model for all three tasks.

2.5 CONTRIBUTION TO THE LITERATURE

Estimating the actual head orientation from 2D images with regard to its three degrees of freedom is a well-known problem that is highly significant for a large number of applications involving head pose knowledge. This explains the growing interest in literature, as detailed in Section 2.4. In the last three years, we have developed different approaches to this topic by using Fractal Coding theory and particularly Partitioned Iterated Function Systems. Initially, these methods were devised as training-free techniques, capable of providing similar if not better performance. Recent advances in ML algorithms and the use of various regression techniques to improve HPE performance have meant that the characteristics extracted from the proposed methods can also be used in combination with different regression models.

2.5.1 *HP^{2}IFS: PIFS Fractal Encoding approach*

Based on Partitioned Iterated Function System (PIFS) and fractal image coding, HP^{2}IFS [24] is a novel method for HPE. PIFS, which was originally used as a lossy image compression algorithm, is exploited as a means to encode auto-similarities within the face image. Benoit B. Mandelbrot pioneered fractal theory in the early 1980s. He starts from the observation that there are self-similar structures in nature, called *fractals*, which have almost identical features at any level of detail they are enlarged.

The Iterated Function System, Fixed-Point Theory, and Collage Theorem form the fractal theoretical basis of image coding. Fractal coding is based on the idea that an image can be represented by a contractive transform whose fixed point is close to the image. Banach's Fixed-Point Theorem, also known as the Contraction Theorem, guarantees that, in a complete metric space, the fixed point of such a transform may be recovered by iterated application thereof to an arbitrary initial element of that space. The Collage Theorem establishes a distance bound between the image to be encoded and the fixed point of a transform, in terms of the distance between the transform of the image and the image itself. In the early 1980s, Barnsley applied his fractal and mathematical knowledge to image compression, noting that this transform was formed of the union of a series of affine mappings on the entire image - an Iterated Function System (IFS). By making the partitioned-IFS, which is different from an IFS in that each mapping works on a subset of the picture instead of the whole picture, Jacquin made it possible to use fractal compression in real life. Each PIFS includes a complete metric space (X, d) and a set of contraction mappings $w_i : M \rightarrow M$ as defined in this space. Fractal image coding just uses a PIFS to represent an original image so that the image after iterative decoding closely approximates the attractor of this PIFS as well as the original image. The coefficients of contraction mapping constitute the fractal coding of the original image.

The fractal image encoding algorithm works, in principle, as follows:

- The image to be encoded is partitioned into R_i, non overlapping Range blocks.

- The image is then partitioned into larger non overlapping blocks D_j called Domain blocks.

- For every Range block R_i, a Domain block D_{R_i} is found such that a contractive affine transformation w, transforms this Domain block to a good approximation of the Range block.

The contracted D block is extended by eight isometric transformations: identity, rotation through 90°, rotation through 180°, rotation through 270°, reflection about the middle vertical axis, reflection about the middle horizontal axis, reflection about the first diagonal, and reflection about the second diagonal. The extended domain pool used to generate a codebook is denoted as $\{\tilde{D}\}$. W_i contraction mapping transformation is defined as:

$$W_i = \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix} \tag{2.7}$$

Where x and y are spatial coordinates, z is pixel value; a_i, b_i, c_i, and d_i define one of the eight isometric transformations; s_i is a brightness adjustment factor with an absolute value less than 1, and o_i is a brightness offset factor. This operation finds the best matching block from the extended domain pool $\{\tilde{D}\}$ for each R block while minimizing distortion error $E(R, \tilde{D})$, which is defined as follows:

$$E(R, \tilde{D}) = ||R - (s \cdot \tilde{D} + o \cdot I)|| \tag{2.8}$$

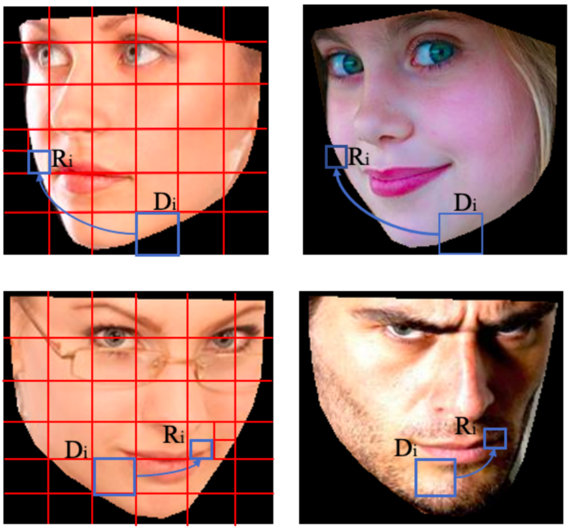


Figure 2.4: The transformations between the domain blocks (D) and the range blocks (R) on image pairs featuring similar angular values..

After fractal parameters of all D_{R_i} blocks are stored as the result of compression, the total fractal encoding process is completed. In Figure 2.4, we can see the transformation of the block of the domain into the block of the range. Images shown are from the dataset AFLW2000, to which we have applied a mask. The two images are labeled with the same head pose: 5° Pitch, 30° Yaw and $-5°$ Roll. We obtain that the same blocks of domain will go in the same blocks of range for both images within an acceptable margin of error due to the self-similarity induced by the fractal codec (Figure 2.5).

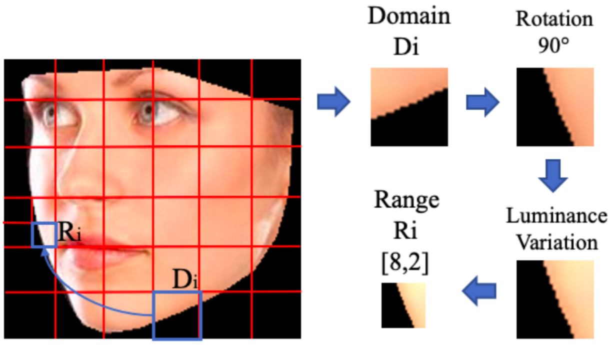


Figure 2.5: A detail of the fractal coding process, in particular the rotation and the lighting variation leading to the range located in row 8 and column 2.

HP^{2}IFS method is composed of three phases, which are summarized below and illustrated in Figure 2.6:

1. Face detection and landmark prediction;
2. Fractal image coding algorithm to generate a matrix containing fractal codes;
3. Pose estimation, transforming the fractal parameters into an array and comparing it to the angular array references obtained in the same way through the Hamming distance metric.

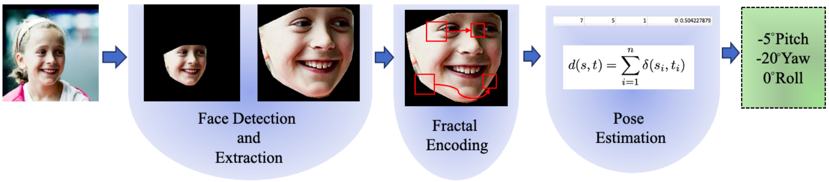


Figure 2.6: The HP^{2}IFS workflow.

As a result of detection and landmark prediction, the input image is used to generate a facial mask based on the boundary landmarks detected. The image is then scaled to 256×256 pixels and encoded with an 8×8 domain and 4×4 range. The resulting codec matrix is of 256 rows and 6 columns where each row defines a block. The first two columns are the block coordinates;

the third is the affine value of the inversion; the fourth is the affine value of the rotation; the final two columns are, respectively, brightness and contrast. In order to perform the comparisons, the matrix is transformed into a 1536-element array. The latter are carried out using the Hamming distance metric, which is ideal for comparing two data strings of the same length and is defined as the number of bit positions in which the two bits differ, as follows:

$$d(s,t) = \sum_{i=1}^{n} \delta(s_i, t_i) \tag{2.9}$$

where s and t are the strings to compare having length n and $\delta(s_i, t_i)$ is the following function:

$$\delta(s_i, t_i) = \begin{cases} 1, & \text{if } s_i \neq t_i \\ 0 & \text{if } s_i = t_i \end{cases} \tag{2.10}$$

This metric is renowned for being straightforward to implement and fast to compute, as its time complexity is proportional to the length of the string. The minimal result obtained by calculating the Hamming Distance between the model arrays and the input array provides the most similar head pose and its pitch, yaw, and roll values. The experiments are conducted on BIWI and AFLW2000 datasets. Figure 2.7 shows some image samples of these datasets. The BIWI database provides numerous images per identity. Because of this, the model is created using the *one-left-out* strategy, in which only one individual is used as a tester and the others as a model for carrying out the comparisons. For the AFLW200 dataset, in order to obtain the subdivision necessary for HP^{2}IFS, about 80% of the dataset images randomly selected are used to build the model and the remaining to perform the tests. The results can be seen in Table 2.2.

Figure 2.8 shows the overall trend of errors as a percentage of images tested. For BIWI, we can observe the same trend anticipated by the numerical results. In terms of pitch, around 35% of images have no error, 77% have an error equal to or less than $5°$, 95% have an error equal to or less than $10°$, and almost all

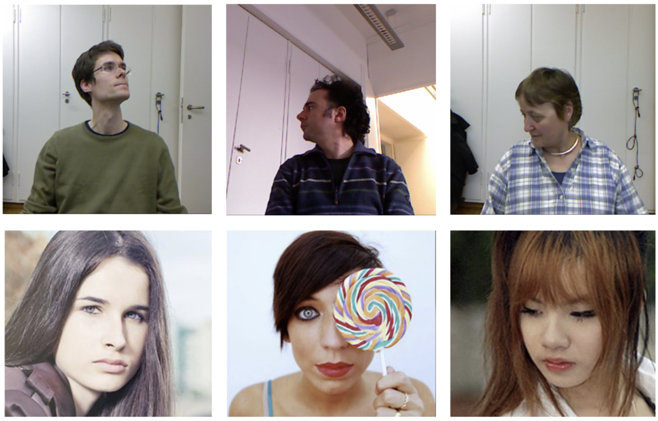


Figure 2.7: Samples from BIWI and AFLW2000 databases with different head-poses.

Table 2.2: HP^{2}IFS: Mean Absolute Error of yaw, pitch and roll angles over BIWI and AFLW2000 datasets.

Dataset	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
BIWI	4.05	6.23	3.30	4.52
AFLW2000	6.28	7.46	5.53	6.42

images have an error less than 15°. In terms of yaw, the results are more promising. Approximately 55% of images are error-free, 90% of images have an error of 5° or less, 97% have an error of 10° or less, and practically all images have an error of less than 15°. Roll achieved the best results, as approximately 71% of images are error-free, 97% of images have an error of 5° or less, and nearly all images have an error of 10° or less. For AFLW2000, the behavior of the error along the three axes is comparable to that of BIWI. Particularly for pitch, yaw, and roll, around 30% of images are error-free. Also in this case, roll errors are the better results: 72% of images have an error equal to or less than 5°, 92% have an error equal to or less than 10°, 98% have an error equal to or less than 15°, and almost all images have an error less than 20°.

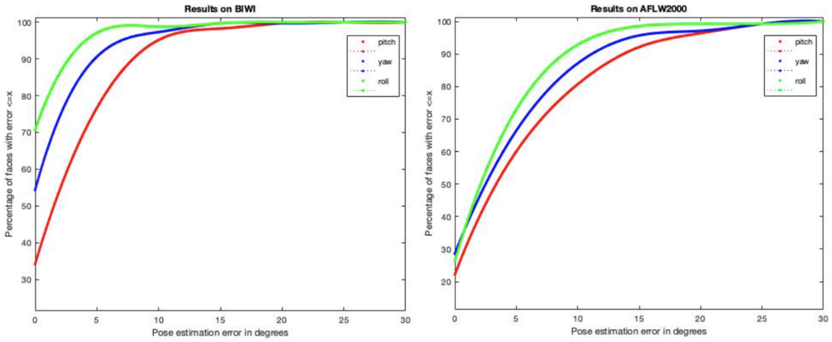


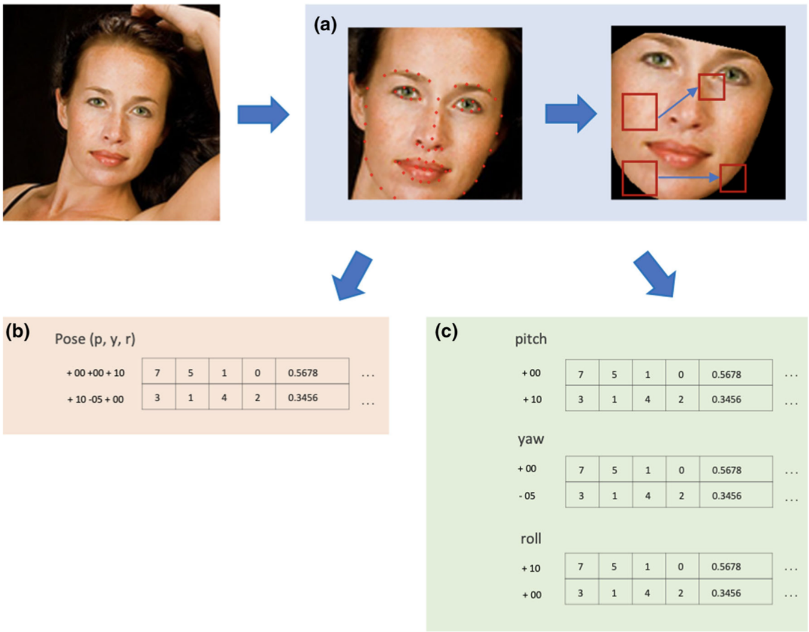
Figure 2.8: Errors on BIWI and AFLW2000 datasets in terms of percentage of tested images.

2.5.1.1 *PIFS by Regression models*

Starting from HP^{2}IFS algorithm to identify the pose, the method presented in [3] adopted different regression models to predict the angular value errors. This approach combines the fractal image compression characteristics, such as self-similar structures in order to identify similar head rotations, with regression analysis prediction. The procedure is illustrated in Figure 2.9.

The classification approach employed in HP^{2}IFS, shown in Figure 2.9-(b), and the regression method are compared to identify the pose in the extracted array. In order to find the most similar one, the pose feature array from the classification method is compared with prototypical vectors extracted using the same method from samples whose poses are known. A pose's encoding in terms of pitch, yaw, and roll is represented by each of the arrays. The pose classification is carried out in [24] by comparing the extracted pose feature vector with those stored in the dataset. The result is the pose whose reference vector has the lowest distance (Hamming) from the extracted vector of the input image. Regression is used to get findings that outperform those obtained using the same methodology. So, for each experiment, three different regression models are built: for pitch, yaw and roll (Figure 2.9-(c)).

To evaluate the predictive power of regression analysis, four different models are involved in the experiments as follows: Linear regression (HP^{2}IFS-LR), Bayesian Ridge regression (HP^{2}IFS-



Pose (p, y, r)

+ 00 +00 + 10	7	5	1	0	0.5678	...
+ 10 -05 + 00	3	1	4	2	0.3456	...

pitch

+ 00	7	5	1	0	0.5678	...
+ 10	3	1	4	2	0.3456	...

yaw

+ 00	7	5	1	0	0.5678	...
- 05	3	1	4	2	0.3456	...

roll

+ 10	7	5	1	0	0.5678	...
+ 00	3	1	4	2	0.3456	...

Figure 2.9: The HP^{2}IFS Regression workflow: a) HP^{2}IFS approach; b) Classification; c) Regression.

BRR), Lasso regression (HP^{2}IFS-LsR) and, finally, Logistic regression (HP^{2}IFS-LgR). The datasets used for experimentation and comparison are BIWI and AFLW2000. The same data splitting protocol adopted in HP^{2}IFS (BIWI: one-left-out technique, AFLW2000: 80/20 ratio) to perform the experiments is applied. Tables 2.3 and 2.4 show, respectively, the results obtained on BIWI and AFLW2000 datasets and compared with the classic HP^{2}IFS approach. Analyzing the Bayesian Ridge regression model for BIWI dataset (Table 2.3), it is possible to note that the roll angular error and the overall MAE are comparable to the HP^{2}IFS classification method, and the pitch angular error is improved over the traditional approach.

The comparison findings from the AFLW2000 database are shown in Table 2.4. The Lasso regression model delivers the lowest MAE value, including pitch and roll angular errors. Additionally, it should be noticed that the yaw angular error value in HP^{2}IFS-LsR and HP^{2}IFS are very close.

Table 2.3: HP^{2}IFS Regression: Mean Absolute Error of yaw, pitch and roll angles on BIWI dataset.

Method	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
HP^{2}IFS	**4.05**	6.23	**3.30**	**4.52**
HP^{2}IFS-LR	6.57	5.47	3.80	5.28
HP^{2}IFS-BRR	6.59	5.46	3.80	5.28
HP^{2}IFS-LgR	9.73	5.82	6.22	7.86
HP^{2}IFS-LsR	6.58	**5.29**	3.80	5.28

Table 2.4: HP^{2}IFS Regression: Mean Absolute Error of yaw, pitch and roll angles on AFLW2000 dataset.

Method	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
HP^{2}IFS	**6.28**	7.46	5.53	6.42
HP^{2}IFS-LR	6.71	6.90	4.48	6.03
HP^{2}IFS-BRR	6.59	7	5.19	6.26
HP^{2}IFS-LgR	8.16	7.71	5.86	7.24
HP^{2}IFS-LsR	6.70	**6.90**	**4.48**	**6.02**

2.5.2 *FASHE: Optimized Fractal Encoding algorithm*

A new fractal-based HPE approach called FASHE has been developed in [23] that exploits a single frontal face image as the reference to build only once the domain blocks required for the fractal encoding algorithm, thus increasing both accuracy and efficiency of pose estimation. The adoption of the classic fractal encoding algorithm in fact foresees that the domain and range blocks belong to the same image. In FASHE, the same reference image is used to build the domain blocks regardless of which image we want to estimate the head pose from. For this purpose, the best candidate image should have a neutral frontal head pose, that is, $0°$ Pitch, $0°$ Yaw and $0°$ Roll. Also note that the reference image does not necessarily have to come from the same subject whose pose you want to estimate; rather, the domain block is built only once during the entire process using a generic front image of a random subject. A detail of the optimized fractal encoding process is show in Figure 2.10. The following steps, described extensively in Section 2.5.1, are the most representative of the overall method.

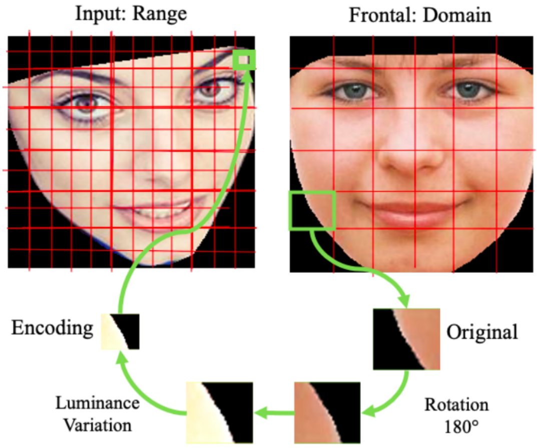


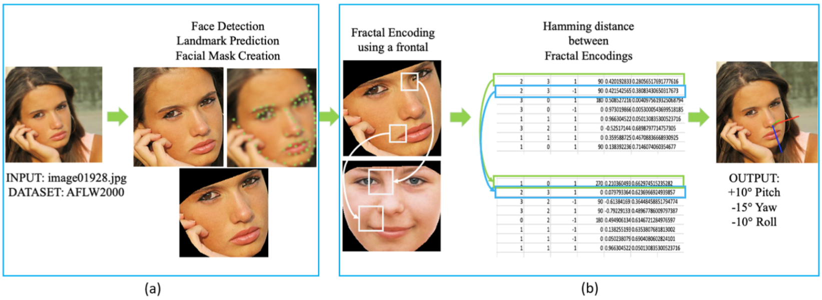
Figure 2.10: FASHE: a detail of the optimized fractal encoding process.

- Step 1: Face detection;
- Step 2: Landmark prediction;
- Step 3: Facial Mask creation;
- Step 4: Fractal codec array;
- Step 5: Hamming distance.

The overall framework of this method can be found in Figure 2.11. It is crucial to note that a filter is necessary before carrying out the aforementioned steps if the input image contains a background with a lot of detail. To determine whether the filter is required, the overall entropy of the image is calculated as follows:

$$E = -\sum_{i} p_i \log_2(p_i) \tag{2.11}$$

Where p_i is the probability of the i-th gray level. When the entropy is higher than the mean entropy of the images stored in the model, it can be assumed that there is a lot of information in the image that can result in excessive differences between two



2	3	1	90	0.420192833	0.28056517691777616
2	3	-1	90	0.421542565	0.38083430650317673
3	0	1	180	0.508527216	0.0040975619325068794
3	0	-1	0	0.973019866	0.0053000543699518185
1	1	1	0	0.966304522	0.050130835300523716
3	2	1	0	-0.52517144	0.6898797714757305
1	1	1	0	0.359588725	0.46708836668930925
1	0	1	90	0.138392236	0.7146074060354677

1	0	1	270	0.210360493	0.662974515235282
2	3	1	0	0.079793364	0.6236966924939857
3	2	-1	90	-0.61384169	0.36448458851794774
3	2	1	90	-0.79229133	0.48967786009797387
0	2	-1	180	0.494906134	0.6146721284976597
1	1	-1	0	0.138255193	0.6353807681813002
1	1	-1	0	0.050238079	0.6904080602824101
1	1	1	0	0.966304522	0.050130835300523716

Figure 2.11: The FASHE workflow.

images, even when they have the same head pose. In this scenario, a Gaussian filter will be applied to the original image. Figure 2.12 illustrates an example of the action taken in this instance. This step will now be added to the modules (a) and (b) of the workflow shown in Figure 2.11.

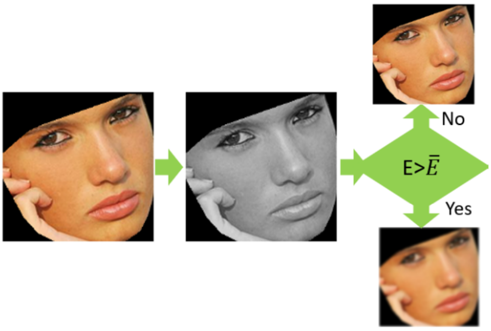


Figure 2.12: FASHE: the additional step for chaotic images.

In order to test and evaluate the performance of the proposed scheme, the experiments are conducted on images taken from BIWI, AFLW2000 and Pointing'04 databases, respectively. In Figure 2.13 we can see some images from Pointing'04 dataset.

FASHE approach likewise uses the configuration of 4 as Range and 8 as Domain, as indicated in Section 2.5.1. By using BIWI to test various range and domain configurations, we are able to demonstrate why $(4, 8)$ is the optimal choice. First of all, since compression is not important to our scope, we can simply set the domain to be twice the range. Next, we can see that both range and domain values must precisely divide the image's dimension,

Figure 2.13: Images from Pointing'04 database.

in this case 256. From those observations, we can analyze the following pairs of ranges and domains: $(2,4)$, $(4,8)$, $(8,16)$, $(16,32)$, $(32,64)$, $(64,128)$ and $(128,256)$. We excluded from our analysis the first pair, because it required 2.53 seconds per image to be computed, resulting in a non-real-time application. Additionally, we do not compare the last two couples because they do not leave enough blocks to compare. By examining the remaining pairs, we are able to determine the Biwi angular values errors and computational times that are shown in Table 2.5. These findings allow us to identify the ideal configuration in $(4,8)$ that we used in our experiments. Finally, some experiments using the Gotcha Video Dataset, a multiview video database that includes video from both cooperative and non-cooperative subjects, are depicted in Figure 2.14. This kind of data, in particular the non-cooperative subjects, is highly helpful to assess the efficacy of a head pose estimate approach to locate the (most) frontal face in a video-sequence acquired in the wild. The experiment findings showed that FASHE was able to identify the closest match to a frontal face pose among the frames provided for each person in the database.

Table 2.5: FASHE: different errors and computational time in seconds, in different range and domain configurations.

Configuration	E_yaw	E_pitch	E_roll	Comp_time
(4,8)	3.13	4.61	2.74	0.1553
(8,16)	3.41	4.81	3.22	0.013
(16,32)	4.13	5.51	4.09	0.0015
(32,64)	9.83	10.08	7.26	0.0002

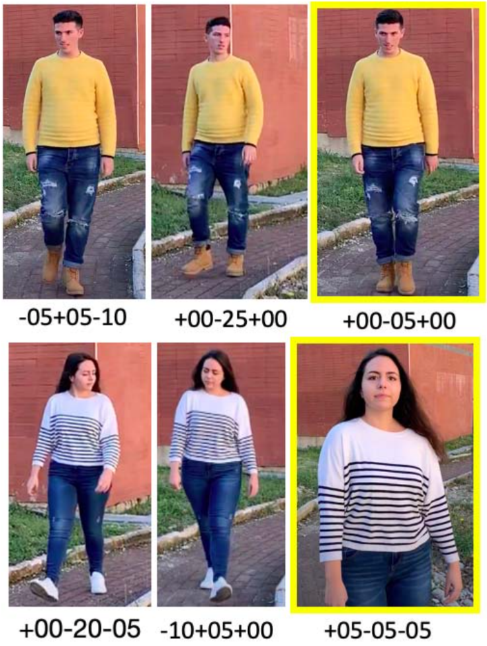


Figure 2.14: Gotcha video sequence: the most frontal pose detected by FASHE algorithm is in yellow.

Table 2.6: FASHE: Mean Absolute Error of yaw, pitch and roll angles over BIWI, AFLW2000 and Pointing'04 datasets.

Dataset	E_yaw	E_pitch	E_roll	MAE
BIWI	3.13	4.61	2.74	3.50
AFLW2000	4.54	6.42	3.71	4.89
Pointing'04	6.6	9	\	7.8

Table 2.6 shows the results of FASHE method over the three datasets. Note that Pointing'04 database does not contain roll information. The percentage of images in BIWI, AFLW2000, and Pointing'04 datasets with errors below a given angle are illustrated in Figure 2.15. More than half of the images on BIWI have an error of 0°, while less than 5% of images have an error greater than 5°. On AFLW2000, approximately 30% of images have an error of 0°, approximately 80% of images have an error of 5° or less, and less than 5% of images have an error greater than 5°. On Pointing'04, the distinction between pitch and yaw is more

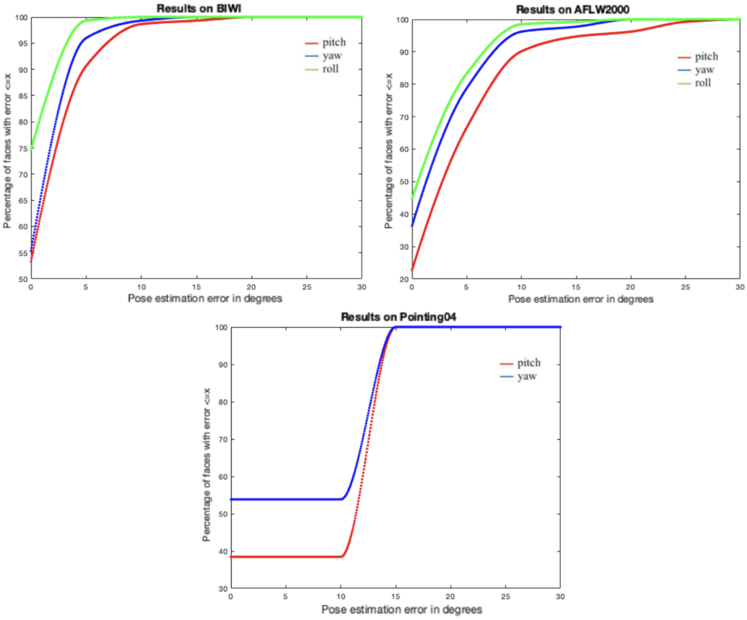


Figure 2.15: Errors on BIWI, AFLW2000 and Pointing'04 datasets in terms of percentage of tested images.

marked. Approximately 40% of images for pitch and 60% of images for yaw have an error of approximately 0°, and there are no images with an error greater than 10°.

2.5.2.1 *FASHE: Gradient boosting regression model*

This work [18] focuses on improving the efficiency of fractal encoding in the context of HPE, aiming at real-time operation while further enhancing the accuracy of the pose estimation. For this purpose, we combine an efficient application of the optimized fractal encoding algorithm, i.e. FASHE Method, with the Gradient Boosting Regressor model (GBR) and the Extreme Gradient Boosting model (XGBoost), yielding a significant improvement in terms of both prediction accuracy and computational efficiency. Figure 2.16 shows an overview of our HPE system.

In FASHE, we modify the classic fractal image compression algorithm and adopt an optimization technique to speed up the blocks search (Figure 2.16-(a)). The neutral frontal head pose, that is 0° pitch, 0° yaw, and 0° roll, is selected as the best candidate

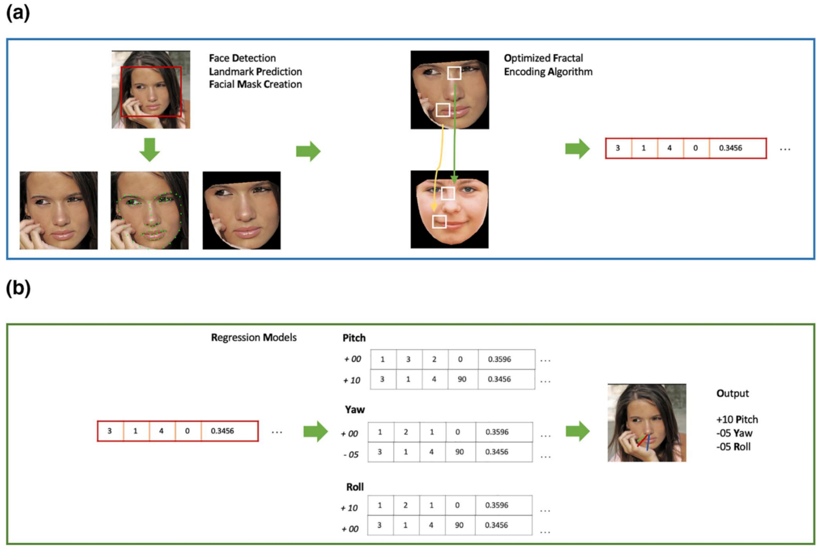


Figure 2.16: The FASHE Regression workflow: a) FASHE approach; b) GBR and XGBoost regression models.

by using the same reference image as when building the domain blocks. In fact, the adoption of the classic fractal algorithm foresees that domain and range blocks belong to the same image. Using the Hamming distance, the experimental results from the different encoded poses are performed. Starting from this strategy, we outperform the results using GBR and XGBoost, two powerful approaches for building supervised regression models (Figure 2.16-(b)). Gradient boosted machines (GBMs) are a boosting machine learning model that adopts a series of "weak" learners in order to create an arbitrarily accurate strong learner. GBMs are additive gradient-based learning models in which a new weak learner is added and trained to reduce the overall error of the entire model while not modifying the weak learners present in the model. Although XGBoost and GBR both adhere to the gradient boosting principle, XGBoost utilizes a more regularized model formalization to avoid overfitting and to better utilize the computational resources. To do this, the objective functions can be made simpler while still allowing an optimal computation speed. The proposed optimised fractal coding combined with GBMs regression models is evaluated on BIWI and AFLW2000 datasets. The errors are listed in Tables 2.7 and 2.8, showing an improved estimation accuracy on the yaw and roll axes and on MAE as well, depending on the regression model used.

Table 2.7: FASHE Regression: Mean Absolute Error of yaw, pitch and roll angles on BIWI dataset.

Method	E_yaw	E_pitch	E_roll	MAE
FASHE	3.13	4.61	2.74	3.50
FASHE-GBR	3.83	4.60	3.49	3.97
FASHE-XGBoost	3.25	**4.01**	3.15	**3.47**

Table 2.8: FASHE Regression: Mean Absolute Error of yaw, pitch and roll angles on AFLW2000 dataset.

Method	E_yaw	E_pitch	E_roll	MAE
FASHE	4.54	6.42	3.71	4.89
FASHE-GBR	4.93	5.97	**3.54**	**4.81**
FASHE-XGBoost	5.19	**5.91**	3.81	4.97

Finally, Table 2.9 reports the comparison with FASHE approach regarding the computing time, including the hardware configuration on which the experimental evaluation was carried out. Both approaches were performed on a MacBook Pro Intel i7 @ 2.6 GHz CPU. As we can see in Table 2.9, our method takes only 0.006 seconds, showing a reduction in computing time of about three orders of magnitude, also highlighting its suitability for real-time operation.

Table 2.9: FASHE-GBMs vs. FASHE: computational time.

Method	Hardware	Total time (in seconds)
FASHE-GBMs	i7 @ 2.6 GHz CPU	**0.006**
FASHE	i7 @ 2.6 GHz CPU	6.604

2.5.3 *HPE Comparisons*

As stated in Section 2.4, there are more training techniques than training-free approaches. The reason is the recent popularity of DL algorithms that result in greater precision. In Tables 2.10–2.12, the results produced by the methods described on the most popular datasets, BIWI, AFLW, and Pointing'04, are shown. Because AFLW2000 is a subset of AFLW and exhibits the same heterogeneity and environment characteristics, we include both AFLW and AFLW2000 in the AFLW table. In the tables below, the best results in terms of angular error for yaw, pitch, roll, and MAE are presented. The protocol chosen throughout our experimental phase calls for the use of the same dataset to perform both training and testing. As a result, in order to conduct a fair comparison, we only report works that used the same strategy. The minimum error on BIWI is approximately 2.5°, as shown in Table 2.10. For both yaw and roll axes, the best results are comparable to those of PIFS-based approaches. From the point of view of the representation used for the angles, the only method using quaternions is [75]; Euler angles are employed by the other BIWI techniques. In this instance, there is no significant difference between the performances obtained using either representation.

Table 2.11 shows that the best result for AFLW is around 1.5°. All methods use the representation of the Euler angle; the only exception is in Xia et al. [167], which adopt a matrix representation, achieving the best results. However, as these are not in line with other state-of-the-art methods, it can be assumed that the representation used is not indicative of the performance of the algorithm.

In Pointing'04, the best result is around 1°. It is obtained in correspondence with a technique that performs the testing using a cross-fold validation approach. To estimate the angles, all of the approaches reported prefer the Euler representation. In general, most of the methods mentioned are not able to achieve considerable results. In this sense, we have to underline that this dataset was collected using 15° as a step instead of 5°. This means that all the approaches are under the level of sensitivity of the dataset, which makes them acceptable.

Table 2.10: The errors in degree for the methods using 2D RGB images of the BIWI dataset.

RGB on BIWI	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
Drouard et al. [49] (2015)	4.9	5.9	4.7	5.17
Chen et al. [40] (2016)	9.9	12.9	6.9	9.90
Lathuiliére et al. [89] (2017)	**3.12**	4.68	3.07	3.62
Drouard et al. [50] (2017)	6.06	7.65	5.62	6.44
Hsu et al. [75] (2018)	4.01	5.49	2.93	4.14
Gupta et al. [71] (2019)	3.46	3.49	2.74	3.23
Abate et al. [2] (2020)	**3.12**	**2.31**	**1.88**	2.43
HP^{2}IFS*	4.05	6.23	3.30	4.52
HP^{2}IFS-LsR	6.58	5.29	3.80	5.28
FASHE*	3.13	4.61	2.74	3.50
FASHE-XGBoost	3.25	4.01	3.15	3.47

Table 2.11: The errors in degree for the methods using 2D RGB images of the AFLW/AFLW2000 dataset.

RGB on AFLW	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
Ranjan et al. [126] (2017)	6.24	5.33	3.29	4.95
Cao et al. [33] (2018)	7.04	7.14	3.86	6.01
Rieger et al. [127] (2019)	8.5	6.5	3.9	6.30
Xia et al. [167] (2019)	**0.63**	**2.05**	**1.70**	**1.46**
Gupta et al. [71] (2019)	5.22	4.43	2.53	4.06
Khan et al. [86] (2020)	4.25	4.89	3.20	4.11
Abate et al. [2] (2020)	4.31	5.43	2.62	4.09
HP^{2}IFS*	6.28	7.46	5.53	6.42
HP^{2}IFS-LsR	6.70	6.90	4.48	6.02
FASHE*	4.54	6.42	3.71	4.89
FASHE-GBR	4.93	5.97	3.54	4.81

All of those results should be taken into account in light of the specific data and framework employed in the methodologies. Several methods in the tables involve manually annotated land-

Table 2.12: The errors in degree for the methods using 2D RGB images of Ponting'04 dataset.

RGB on Pointing'04	**E_yaw**	**E_pitch**	**MAE**
Drouard et al. [49] (2015)	7.5	7.3	7.40
Alioua et al. [6] (2016)	6.1	4.6	5.35
Liu et al. [101] (2017)	nd	nd	6.6
Drouard et al. [50] (2017)	7.93	8.47	8.20
Diaz-Chito et al. [48] (2018)	8.1	9.6	8.85
Xu et al. [169] (2019)	3.92	\	3.92
Vo et al. [161] (2019)	7.17	6.16	6.67
Bounoua et al. [115] (2020)	**1.78**	**0.82**	**1.30**
Khan et al. [86] (2020)	2.68	1.32	2.00
Abate et al. [2] (2020)	4.44	7.55	5.99
FASHE*	6.6	9	7.8

marks, facial annotations, etc. Furthermore, approaches marked with an asterisk (*) do not use training techniques. The methods described lead us to the conclusion that approaches that appear to produce the worst results in terms of angular error frequently have a low computational time requirement because they prioritize speed above accuracy. As a final point, approaches that adopt a previously existing network (such as VGG16, GoogLeNet, or ResNet50) inherit the starting weight resulting from a lengthy training phase. This could have a significant advantage in terms of training time, compensating for the very small datasets that are currently accessible.

2.5.4 *SHEEF: PIFS Scheme for HPE aimed at a faster Face recognition*

In the context of one of the most common computer vision tasks, face recognition, where head orientation (especially in the case of large angles) represents a challenging intra-class variation, the information offered by HPE may be significant. In this work [25], head pose estimation is viewed as a synergistic component of a fast face recognition method in which three crucial biometric pipeline steps - feature extraction, pose estimation, and feature vector matching - exploit the PIFS and fractal encoding to achieve high efficiency and accuracy. The goal of this study is to dramatically increase the efficiency of face recognition by utilizing the same type of fractal features for both tasks, building on the HPE approach introduced in Section 2.5.1. The framework that results will be defined as SHEEF: partitioned iterated function systems Scheme for HEad Pose Estimation aimed at a faster Face recognition. SHEEF claims that when the face in the probe image has been detected and extracted, face features are encoded using PIFS, and HPE is then carried out by comparing the fractal code that results to a set of pre-computed codes in a reference template in order to determine the minimum distance. By finding the most similar poses (encoded in the related feature vectors) out of all those available for each subject in the gallery, the resulting pose angles are used to speed up the future matching stage. Finally, the probe's identity is determined by matching it to the gallery's template with similar poses through a metric distance. The main contributions of this work can be summarized as follows:

- a synergistic combination of HPE and face recognition achieving high processing efficiency in one-to-many matching scenarios and real-time applications;
- the same fractal code is conveniently used for both tasks, achieving an high accuracy for both pose estimation and face recognition;
- newly recognized probes of the same subject can be used to extend his gallery template, thus increasing future chance to find a match for a greater range of presentations.

Figure 2.17 illustrates the SHEEF workflow. The main steps can be summarized as follows: face detection using HOG+SVM; face encoding using partitioned iterated functions; face extraction using regression trees; HPE, using the distances between encodings; searching for similar poses through the estimated head pose; identity estimation using the distances between encodings among similar poses; and finally, extension of the template by possibly adding the new head pose fractal encoding of the subject in his/her template.

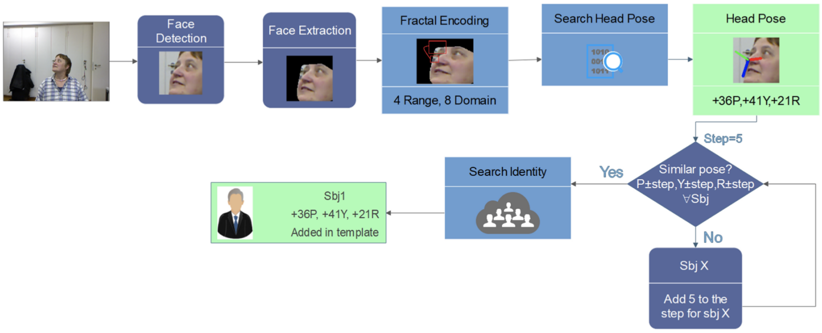


Figure 2.17: The SHEEF workflow.

Searching for similar poses in the template can dramatically increase speed without significantly decreasing recognition accuracy, as we will demonstrate in the results. For this reason, in this step, we have to choose the strategy with which to perform the comparisons. The predicted Yaw, Pitch and Roll angles are defined as Y, P and R, respectively. The technique will look for encodings whose poses fall within the range specified by a step of $\pm 5^\circ$, e.g., $Y \pm 5^\circ$, $P \pm 5^\circ$, $R \pm 5^\circ$. If a subject's poses are missing from the template, the research for this subject will be expanded by a further step of $\pm 5^\circ$. This procedure is repeated until, for each individual, at least one encoding can be extracted in order to perform the comparisons. It is clear that, if we proceed in this manner, we will have the set of encodes for each subject with the most similar head pose to the input image.

The use of the same encoding for head pose estimation and identity identification is one aspect of this approach that deserves to be highlighted. It follows that changing the distance metric is sufficient to draw attention to various parts of the fractal encoding that are useful for HPE or identity identification, rather than extracting different features to do those two tasks. This has a double advantage. On the one hand, we create a fast approach by performing the feature extraction task just once for both purposes. On the other hand, we only need to keep one template with two types of labels (subject and head pose), which greatly reduces the amount of storage space needed. The fact that the framework structure allows for template updating at any time - either by adding new head poses to existing subjects or by introducing new subjects to be recognized - is another significant component of the framework. To ensure that a new subject will be included in the comparisons, just one additional frame of that subject is required. Our technique is significantly different from a traditional neural network architecture in that it may be totally re-adapted at any moment without the need for further training stages or modifications to previously learned structures or information.

COMPARISON OF DIFFERENT DISTANCE METRICS After computing the fractal encoding, we have to use a metric that can compare the fractal codes and provide the correct label. Our goal in this instance is to experiment with different metrics that emphasize the characteristics of fractal encoding for both tasks. Several distances, including Hamming, Canberra, Jaccard, Cityblock, Correlation, Chebyschev, Euclidean, and Braycurtis, were used in our tests. It is evident that each of them can accomplish the given tasks - HPE and Face Recognition - more or less effectively. For this reason, we tested the two datasets introduced in Section 2.5.1 using all of these metrics. Tables 2.13 and 2.14 show that Canberra is the best distance for solving the HPE problem. Canberra proves to be equally as effective as Hamming, which was utilized in [24]. Table 2.15 reveals that Canberra, once more, is the best distance for resolving the Face Recognition task. Here, we used 80% of the faces in the template and the remaining 20% for testing. This configuration will be known as *SHEEF (step=inf)*. The time complexity of this metric is linear with respect to the length of the strings being compared, making it very simple to implement.

OVERALL RESULTS OF HPE We investigated widely used image compression methods such as the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT) in order to highlight the potential of the fractal encoding approach and, therefore, the characteristics obtained from fractal objects self-similarity. The results of utilizing DCT and DWT image compression techniques on both datasets are reported in Tables 2.16 and 2.17, respectively, along with a comparison with the proposed fractal encoding algorithm. We used the same configurations

Table 2.13: The distances evaluated on Biwi HPE problem.

HPE - BIWI Dataset				
Distance	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
Hamming	4.05	6.23	3.3	4.52
Canberra	**3.18**	**4.63**	**2.84**	**3.55**
Jaccard	3.75	5.39	3.33	4.15
Cityblock	10.31	8.55	6.56	8.47
Correlation	9.26	8.34	7.04	8.21
Chebyschev	18.51	13.77	12.66	14.98
Euclidean	15.26	10.49	8.49	11.41
Braycurtis	6.97	7.62	5.76	6.78

Table 2.14: The distances evaluated on AFLW2000 HPE problem.

HPE - AFLW2000 Dataset				
Distance	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
Hamming	6.28	7.46	5.53	6.42
Canberra	**6.49**	**7.28**	**4.54**	**6.10**
Jaccard	7.62	8.61	5.59	7.27
Cityblock	11.04	10	6.18	9.07
Correlation	10.7	9.26	6.32	8.76
Chebyschev	12.31	12.57	6.38	10.42
Euclidean	12.51	9.94	6.77	9.74
Braycurtis	9.88	8.89	5.56	8.11

in all three cases: face detector and face mask with Dlib [84]; encoding of the face mask with the corresponding algorithm; Canberra distance to perform comparisons. Finally, the same data splitting protocol adopted in all HPE experiments is applied (BIWI: one-left-out technique, AFLW2000: 80/20 ratio).

FACE RECOGNITION Fractal encoding has been used to perform facial recognition in several works [52, 116, 150]. However, none of these consider HPE from the perspective of enhancing

Table 2.15: The distances evaluated on BIWI Face Recognition problem.

Face Recognition - Biwi Dataset	
Distance	**Accuracy (%)**
Hamming	92.02
Jaccard	92.49
Canberra	**94.69**
Cityblock	89.91
Correlation	76.46
Chebyschev	17.27
Euclidian	74.82
Braycurtis	90.14

Table 2.16: Comparative results with DCT and DWT on BIWI dataset.

Method	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
DCT	20.37	9.87	12.63	14.29
DWT	15.18	9.81	11.72	12.23
PIFS	**3.18**	**4.63**	**2.84**	**3.55**

Table 2.17: Comparative results with DCT and DWT on AFLW2000 dataset.

Method	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
DCT	15.81	10	8.16	11.32
DWT	12.88	11.12	8.30	10.76
PIFS	**6.42**	**7.28**	**4.54**	**6.103**

the framework. In fact, since the faces stored in the template are ideally frontal, a non-frontal head pose is sometimes viewed as a disadvantage in face recognition. On the other hand, in our study, we want to exploit the head pose information to our advantage when employing fractal encoding for recognition. There isn't yet a database created expressly to study HPE and face recognition together. Due to this, we used the HPE dataset, which includes subjects' identity information (i.e., BIWI). In fact, AFLW has no label for the subject identities, despite being more competitive for HPE. To the best of our knowledge, there are no face recognition datasets that have the head pose labeled along the three axes. Following the procedure in Figure 2.17 on BIWI (Step = 5), the results obtained in recognition reached an accuracy of 94.84% and a mean processing time per image of 0.003s.

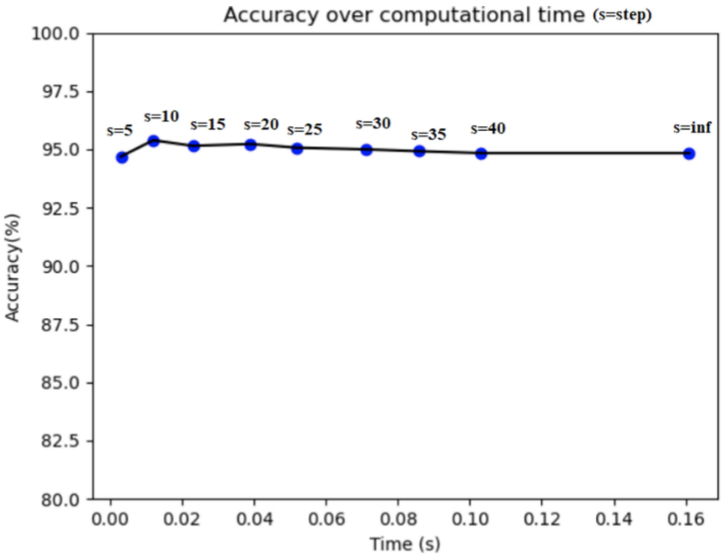


Figure 2.18: Accuracy over time, an initial-step evaluation.

To enhance the contributions of the HPE in order to considerably accelerate face recognition in a real-time application situation, we investigated the dependencies between accuracy and time. We increased the initial step to determine whether the selective HPE search's improved speed is associated with a loss of precision. Figure 2.18 illustrates the obtained results. As can be observed, the overall variation in accuracy is only 0.5% across the initial steps. In contrast, the processing time required is vastly different, ranging from 0.003 seconds for the initial step 5 to 0.161 seconds for the exhaustive search, e.g., without knowing HPE. This means that checking for similar poses requires only 1.86% of the time of an exhaustive search. As stated in previous works, HPE is a real-time process; thus, it is convenient to search for similar poses in the dataset. This result is particularly observable for fractal encoding. This is due to the fact that fractal encoding is the most expensive step in the process and only needs to be performed once, regardless of whether the HPE improvement is utilized or not. We must emphasize that similar poses are always considered, and the pool of search results is only changed around those poses.

Table 2.18: Performance comparisons with state-of-the-art Deep Neural Networks on BIWI dataset.

Method	**Time (s)**	**Accuracy (%)**
FaceNet [135]	0.09	91.22
Facenet512 [135]	0.12	96.49
OpenFace [12]	0.11	56.14
DeepFace [149]	0.48	70.17
DeepID [165]	0.12	72.80
ArcFace [45]	0.14	94.73
VGGFace [123]	0.114	93.85
VGGFace2 [32]	0.09	94.73
SHEEF (step=inf)	0.161	**96.99**
SHEEF (step=5)	0.003	94.84
SHEEF (step=10)	0.012	95.39

Since we used BIWI for face recognition for the first time, we fine-tuned the most popular networks in literature with available code to obtain their performances in accuracy and efficiency. FaceNet and FaceNet512 [135], OpenFace [12], DeepFace [149], DeepID [165], ArcFace [45], VGGFace [123] and VGGFace2 [32] are the architectures that we tested. Table 2.18 displays the resulting accuracy and time in comparison to our results. The same device was used to measure all computational times. We have listed three SHEEF versions in this table. We indicate the exhaustive search along the template with Step = inf. The framework shown in Figure 2.17 starts with Step = 5, which requires the lower computational time, and, finally, the result of Step = 10 is shown. Even if the dataset structure affects the best-performance step, there are few oscillations and only a small 0.5% difference between it and the standard step 5. We can see that SHEEF surpasses the other ML methods when performing exhaustive comparisons. However, with only a slight loss in accuracy and a significant time difference between SHEEF and the fastest of the reported techniques, our method is able to determine the identification in the shortest time (FaceNet and VGGFace2). If we consider that SHEEF has further room for improvement when the image per subject grows and that it may be used without retraining on new subjects, this is a remarkable advantage. In conclusion, SHEEF accuracy and computational efficiency are comparable to those of ML methods.

IMPACT OF THE FACE DETECTOR Dlib was the face detector utilized in previous experiments. The most notable Dlib advantage is its computational speed. To examine the impact of face detector performance on the final results, we conducted further experiments with RetinaFace [44]. RetinaFace is a recent regression-based face landmarks detector. Table 2.19 presents the results achieved for HPE and face recognition using both face detectors on the identical image sets.

Table 2.19: The overall accuracy compared using different Face Detectors.

HPE-AFLW				
Detector	**E_Yaw**	**E_Pitch**	**E_Roll**	**MAE**
Dlib	6.49	7.28	4.54	**6.10**
RetinaFace	12.63	8.35	5.2	8.72
HPE-BIWI				
Detector	**Err_Pitch**	**Err_Yaw**	**Err_Roll**	**MAE**
Dlib	3.18	4.63	2.84	**3.55**
RetinaFace	3.67	4.5	3.11	3.76
FACE RECOGNITION-BIWI				
Detector	**Accuracy (%)**			
Dlib	94.69			
RetinaFace	**96.99**			

As can be seen, the accuracy of the detectors is similar on Biwi, however, RetinaFace is superior on the face recognition task and Dlib on the AFLW database. Here, we ensure that the errors obtained are evaluated on the same set of images. However, we want to highlight that RetinaFace is able to detect more faces than Dlib, despite having very similar performance to that shown in this table. Thus, we can deduce that RetinaFace could only be used if Dlib failed to detect the face.

2.6 CONCLUSIONS

In order to considerably reduce computing time, we proposed various fractal encoding-based HPE methods that initially explore training-free techniques before combining the features extracted with well-known regression models. The core of the proposed approaches is considerably different from current research relying on CNNs methods, as they are based on Partitioned Iterated Function Systems to represent the self-similarity characteristics of two images exhibiting similar head rotation. In general, the main problem in IFS-based encoding resides in finding the domain that can be best transformed into a given range. This process, which must be carried out for each range-domain couple, is very expensive computationally. To overcome these limitations, we subsequently adopted an optimization algorithm to speed up the search of the blocks while also modifying the classic fractal image compression technique. Extensive experiments conducted on challenging databases show competitive accuracy with current state-of-the-art approaches, highlighting a significant gain with regard to the time required to estimate the pose, thus making the method fully suited to real-time applications. We also developed a unified method for face recognition and HPE that uses the same fractal encoding features for both tasks. An increase in face recognition efficiency is made possible by precise pose estimation. Furthermore, the advantage of our framework in terms of matching time and overall time-to-recognition may be especially helpful in light of the fact that the almost universal availability of imaging devices capable of capturing video is changing the operational horizon for face recognition applications. On the other hand, currently, one of the main limitations is undoubtedly the need to work with multiple frames for each subject. In fact, in the event that few frames are available or are very different from those inserted as input, the system performance could drop drastically.

Looking at the new challenges, we can conclude that HPE has the potential to be effective in more emerging fields than have been presented so far. With a booming literature, HPE is an interesting technique to apply in support of biometric frameworks. Due to the growing interest in driver applications, we will carry over our approaches to depth imaging in the near future. In fact, in addition to improved best frame selection and facial recognition, HPE has proven its worth in assessing the subject's attentional state. This confirms the great horizontal expansion that HPE approaches can achieve in the future, supporting the growing number of methods that have already emerged in this field.

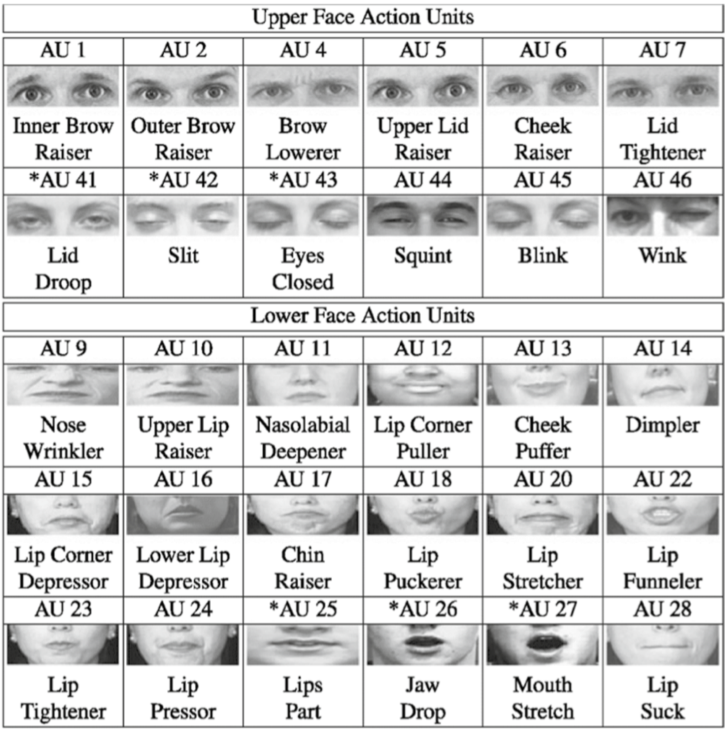
3 FACIAL EXPRESSION RECOGNITION AS BEHAVIORAL BIOMETRICS

Facial expression is one of the most powerful, natural, and immediate means for human beings to communicate their emotions and intentions. Since emotion state is involved in activating the facial muscles movements, FER can be classified as behavioral biometrics. In the field of computer vision and pattern recognition, significant progress has been made in the development of computer systems that can interpret and utilize this natural form of human communication. In this Chapter, we explore the main existing models in literature for quantifying affective facial behaviors, with a special emphasis on the categorical model. Although FER uses multiple sensors, we limit our discussion to exclusively using static images, with a focus on recent DL-based FER systems. Therefore, motivated by these studies, we developed a CNN-based approach to categorize the principal facial expressions. As a further contribution, we also investigate the impact of facial expressions on HPE, specifically the axis most affected by the error when a specific facial expression is visible.

3.1 BACKGROUND

The human face is an important interface for conveying nonverbal emotional information. Individuals' facial expressions reflect their reactions to personal thoughts or external stimuli. These can offer valuable biometric data to automated human recognition systems. The study of FER has received extensive attention in the fields of psychology, computer vision, and pattern recognition. FER has broad applications in multiple domains, including virtual and augmented reality, human-computer interaction, advanced driver assistance systems, education, entertainment, and healthcare [19, 76]. Facial behaviors can be thought of as observable patterns that explain how facial expressions change over time. Different facial expressions, as well as their frequency, duration, and intervals between expressions, all contribute to the subject's facial behavior. Although typical human emotions are nearly universal, facial behavior reveals their temporal and interdependent relationships. FER is based on face muscle activation analysis obtained from its deformation or movement. Since the emotional state is related to the activation of the facial muscles, facial expression analysis can be classified as behavioral biometrics [55]. FER has gained popularity in affective computing due to the premise that the human face carries more information through non-verbal communication channels than speech and body movement. Previous emotional events and physical factors can both have an indirect effect on an individual's emotional state, biasing various facial expressions. However, as with any other behavioral biometric, it is possible to make some safe assumptions when, in a normal and consistent situation, an emotional state is considered.

In literature, there are several models for quantifying affective facial behaviors: 1) *categorical model*, in which the emotion or affect belongs to a list of affective-related categories such as the six basic emotions identified by Ekman et al. [53], 2) *dimensional model* [129], where a value is selected from a continuous emotional scale, such as valence and arousal, and 3) *Facial Action Coding System (FACS) model* [54], in which all possible facial actions are defined in terms of Action Units (AUs). The FACS model describes facial movements but does not directly reflect affective states. There are various methods for converting AUs to affect space (e.g., EMFACS [61] states that the occurrence of AU6 and AU12 is a sign of happiness). In 1971, Ekman and Friesen defined six essential emotions based on cross-cultural research suggesting that humans perceive certain basic emotions equally regardless of culture [53]. These facial expressions are anger, disgust, fear, happiness, sadness, and surprise (as well as neutral). Contempt was later identified as one of the fundamental emotions [113]. Recent neuroscientific research has argued that the six major emotional paradigms are culture-specific and not universal [78]. Some years later, inspired by cognitive and psychological studies, Ekman and Friesen developed the FACS for describing facial expressions using AUs. They focused on physical cues and anatomical knowledge of face behavior to guide facial expression recognition. FACS measures and classifies facial behavior by correlating momentary changes in appearance with muscle action. It can also describe emotion intensities and compound emotions, as well as distinguishing between fake and authentic emotional expressions. The FACS employs AUs, which reflect the muscular activities required to define and analyze facial expressions. A single muscle is used by the majority of AUs. However, in certain cases, two or more AUs are involved to represent the relatively independent actions of different parts of a single muscle. Currently, FACS is composed of a total of 46 AUs. Figure 3.1 illustrates some sample images for specific AUs.



Upper Face Action Units

AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink

Lower Face Action Units

AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 3.1: Facial Action Coding System: Action Units [54, 180].

In the categorical model, it is impossible to translate complex emotions into a small number of words [51]. To circumvent this issue, some studies sought to define numerous unique compound emotion categories (e.g., happily surprised, sadly fearful). The set is still limited, and the categorical model cannot characterize the strength or intensity of the emotions. On the other hand, the dimensional model can differentiate between subtly different affect displays and capture minor changes in the intensity of each emotion, such as valence and arousal, on a continuous scale. Valence relates to whether an event is positive or negative, whereas arousal indicates if an event is exciting, agitating, or soothing. As shown in Figure 3.2, there are several affect types and minute variations within the same emotion that cannot be adequately mapped into the limited vocabulary of the categorical model. The dimensional model incorporates both intensity and distinct emotion categories in the continuous domain. Despite this, there have been few studies focusing on the development of automated methods for evaluating affect using the continuous-dimensional model.

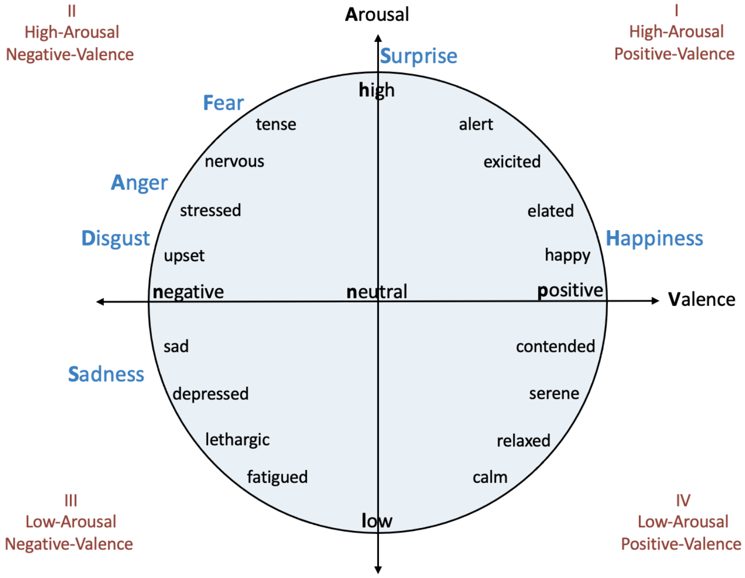


Figure 3.2: The 2D valence-arousal emotion space.

Even though the affective model related to basic emotions is limited in its ability to represent the complexity and subtlety of our daily affective manifestations and other models of describing emotions, such as the FACS and the continuous model that uses the affective dimensions, represent a wider range of emotions [70], the categorical model is still the most popular perspective for FER due to its pioneering studies and direct and intuitive definition of facial expressions. In this Chapter, we will focus on FER using the categorical model. Figure 3.3 illustrates the seven basic facial expressions.

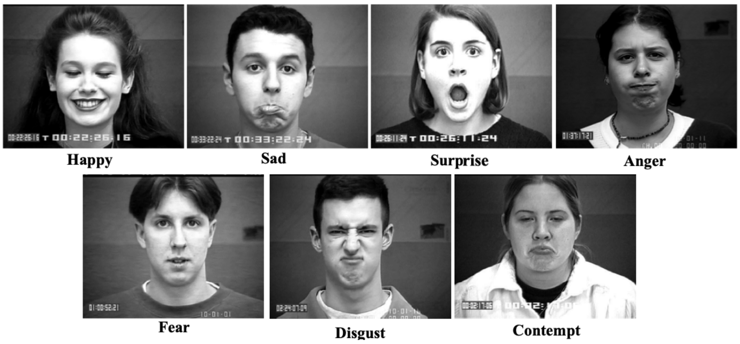


Figure 3.3: Basic facial expressions (sample image from the Extended Cohn–Kanade (CK+) database [103]).

FER systems can be divided into two main categories based on their feature representations: static images and dynamic sequences. Static-based methods encode the feature representation using only spatial information from the current image, whereas dynamic-based methods consider the temporal relationship between consecutive frames in the input facial expression sequence. Before researchers can select the most suitable feature extraction and classification method for the target dataset, the image must be preprocessed. Figure 3.4 depicts the FER system architecture, which consists of three basic steps: image preprocessing, feature extraction, and expression categorization. The performance of feature extraction and expression classification is directly affected by image preprocessing. Noise reduction, face detection, image normalization, and histogram equalization are examples of typical image processing techniques. Most traditional FER systems depended mainly on laboratory-controlled databases and employed hand-crafted features or shallow learning. Due to the significantly increased chip processing capabilities (e.g., GPU units), well-designed network architectures, and more effective training data, numerous DL techniques have been developed in literature to deal with expression information in facial representations. These approaches have improved recognition accuracy and achieved higher performance levels than conventional methods.

Recently, the research community has focused its attention on in-the-wild settings, implicitly promoting the transition of FER from lab-controlled to real-world scenarios.

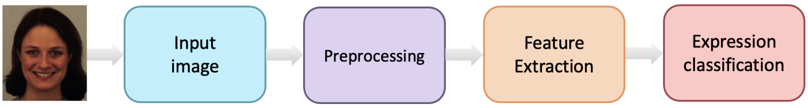


Figure 3.4: FER system architecture.

FACIAL EXPRESSION DATABASES Researchers have access to numerous databases in order to build FER systems capable of producing results comparable to related work. The following is an overview of the most well-known 2D datasets that contain basic expressions and are widely used in literature. Table 3.1 provides a summary of these datasets, such as the number of participants, images or video samples collected, the collection environment, and the distribution of the expressions.

The Extended Cohn–Kanade (CK+) database [103] is the most extensively used laboratory-controlled dataset. The CK+ database was released in 2010 as an extension of the Cohn–Kanade (CK) database. Specifically, the number of emotion states in CK+ was increased to eight, and all emotion labels were modified and validated to improve database performance. Meanwhile, the complexity of recognizing expressions has substantially increased. The CK+ database contains 593 image sequences from 123 participants as well as eight core emotion categories: anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral. It should be noted that image sequences range in length from 10 to 60 frames, with the same criteria that show a shift from neutral expression to peak expression. The MMI Facial Expression database [159] is also laboratory-controlled. It contains more than 2900 videos and high-resolution images of 75 subjects. Each AU in the videos is exhaustively annotated. In contrast to CK+, MMI sequences are classified onset-apex-offset, i.e., the sequence begins with a neutral expression, reaches a peak in the middle, and then returns to a neutral expression. The Karolinska Directed Emotional Faces (KDEF) database [104] involves 4900 images of human facial expressions. The dataset includes 70 people (35 females and 35 males) who portray the six different basic emotions as well as the neutral expression. Each expression was captured in two sessions and from five distinct angles. The Oulu-CASIA dataset [178] includes 2,880 image sequences from 80 subjects, each with six basic expressions. Every video was captured using one of two imaging systems, near-infrared (NIR) or visible light (VIS), under three different lighting conditions. Subjects were instructed to produce a facial expression based on an example provided in picture sequences. The first frame, similar to CK+, is neutral, while the last frame has the peak expression. The Japanese Female Facial Expression (JAFFE) database [107] is composed of 213 samples of posed expression from 10 Japanese female models. Each subject has 34 images of each of the six fundamental facial emotions and one neutral expression. The Acted Facial Expressions in the Wild (AFEW) dataset [47] is a dynamic temporal facial expression database comprised of video clips belonging to various films with spontaneous facial expressions, diverse head poses, occlusions, and illuminations. The AFEW is divided into three data partitions based on subject and movie or TV source, ensuring that the data in the three sets pertains to mutually exclusive films and actors. The Static Facial Expressions in the Wild (SFEW) dataset [46] was produced by selecting frames from the AFEW database. The Facial Expression Recognition 2013 (FER2013) database [67] is a large-scale and unconstrained database automatically collected using the Google image search API. The dataset consists of 35,887 images resized to 48x48 pixels, including 28,709 training sets, 3,589 verification sets, and 589 test sets. These samples are classified into seven categories: angry, disgusting, fearful, happy, neutral, sad, and amazed. FER2013 contains a substantial diversity in age, gender, ethnicity, and pose, simulating the real world. The EmotioNet [58] is an extensive database containing one million images of facial expressions gathered from the Internet. The automatic action unit detection model in [58] labeled 950,000 images, while the remaining 25,000 were manually annotated with 11 AUs. The EmotioNet Challenge's second track [21] includes six basic expressions and ten compound expressions, as well as 2,478 images with expression labels. The Expression in-the-Wild Database (ExpW) [176] and the Real-world Affective Face Database (RAF-

DB) [93] are additional real-world datasets containing a variety of facial images collected from the Internet and Google image search, respectively. The ExpW database includes 91,793 manually annotated faces belonging to one of the seven expressions. RAF-DB is made up of 29,672 highly different facial images that have been manually labeled by about 40 annotators for a total of seven basic and twelve compound emotion labels. Finally, the AffectNet database [117] comprises over one million images acquired from the Internet by querying multiple search engines with emotion-related tags. It is by far the largest dataset with facial expressions in two different emotion models (categorical model and dimensional model). With eight basic expressions, 450,000 frames were manually labeled.

Table 3.1: Summary of FER-related databases [92].

Dataset	**Samples**	**#Subj**	**Source**	**Elicit.**	**Expressions**
CK+	593 image sequences	123	Lab	P & S	7 basic expr. plus contempt
MMI	740 images and 2,900 videos	25	Lab	P	7 basic expr.
KDEF	4,900 images	70	Lab	P	7 basic expr.
Oulu-CASIA	2,880 image sequences	80	Lab	P	6 basic expr.
JAFFE	213 images	10	Lab	P	7 basic expr.
AFEW	1,809 videos	nd	Movie	P & S	7 basic expr.
SFEW	1,766 images	nd	Movie	P & S	7 basic expr.
FER2013	35,887 images	nd	Web	P & S	7 basic expr.
EmotioNet	1,000,000 images	nd	Internet	P & S	23 basic expr. or compound expr.
ExpW	91,793 images	nd	Internet	P & S	7 basic expr.
RAF-DB	29,672 images	nd	Internet	P & S	7 basic expr. and 12 compound expr.
AffectNet	450,000 images (labeled)	nd	Internet	P & S	7 basic expr.

nd = "not declared", expr. = "expressions", Elicit. = "elicitation method", P = "posed", S = "spontaneous".

3.2 LITERATURE REVIEW

In traditional FER systems, hand-crafted appearance and/or geometric features are employed to recognize basic facial expressions. Geometric features define the face shape and its components, whereas appearance features represent information about the image texture using gray values of pixels along with their neighbors. There has been a lot of interest in using DL in recent years [92]. However, handcrafted features are still used. The authors in [155] classified FER features into three groups: geometric, appearance, and deep features. The recognition of facial expressions using geometric features has been adopted in [170]. The authors have extracted motion-based characteristics via the facial landmarks to encode the expression intensity. Kas et al. [83] combine geometric and appearance feature methods. Although the results of both approaches are encouraging, they were not assessed under uncontrolled conditions. Geometric feature extraction typically requires a precise face and landmark detection procedure; however, in appearance feature-based methods, the feature vector is frequently produced by convolution of the facial image using hand-crafted filters. In this regard, the Gabor filter is a popular and commonly employed FER method. Local Binary Patterns (LBP) are another image texture descriptor approach that has been used in this field, and its variants have been utilized to extract facial appearance features. The authors in [111] proposed a novel edge-based descriptor, Local Prominent Directional Pattern (LPDP), which used pixel neighborhood information to encode more meaningful and reliable information than existing descriptors. Several histogram-based image descriptors have also been presented for FER problem [155].

In the past few years, several DL-based algorithms for FER have been presented. Yang et al. [172] introduced a novel feature separation model exchange-Generative Adversarial Network (GAN), which divides expression-related features and expression-independent features with great precision. The authors in [97] suggested a facial expression recognition scheme via a generation scheme termed Identity-Disentangled Facial Expression Recognition Machine (IDFERM) that separates identity factors from other facial expression-causing factors. They claimed that the identity-preserving neutral face image generation is efficient for hard negative mining, requiring fewer similarity comparisons. Li and Xu [90] presented a framework based on reinforcement learning for pre-selecting useful images(RLPS) for emotion classification in the wild. In [168], a feature sparseness-based L2-norm regularization that learns deep features with better generalization capability is proposed. In [146], the authors proposed a self-adaptive approach to learn and extract active features based on a priori knowledge. Ji et al. [82] introduced, respectively, an intra-category common feature representation channel and an inter-category distinction feature representation channel, finally combining the learned features of the two channels in cross databases. Another study [144] presented an eleven-layered CNN with visual attention. In a video-based FER method [99], the authors advised utilizing CNN to extract spatio-temporal features. It is important to underline that video data gives more information than static images; nevertheless, videos are not always accessible and are more computationally complex. Furthermore, the proposed FER system, which is detailed below, is designed for single frames (static). Consequently, this comprehensive review is focused only on 2D images.

Sun et al. [145] designed a deep model for fine-tuning a pre-specified deep CNN. To increase the facial expressions, they designed a novel data augmentation strategy called artificial face. The authors in [179] proposed a novel selective feature-sharing method, and establish a multi-task CNN network for facial expression synthesis and recognition. In [130], a deep histogram metric learning in a CNN was presented. FER is currently working on different attention approaches using deep networks for detecting salient and facial features. In [65], for example, the authors developed an attention based architecture devoid of external sources like landmark detector. Recently, Liu et al. [98] utilized three parallel multi-channel CNN to learn fused global and local features from different facial regions. DML-Net, a Dynamic Multi-channel metric Learning network for pose-aware and identity-invariant representations of facial expressions, was proposed in order to render the system invariant to these two crucial challenges in FER.

3.3 DEEP FACIAL EXPRESSION RECOGNITION: PROPOSED APPROACH

Inspired by the above discussions, we designed a novel DL-based facial expression recognition system [156]. The proposed framework is divided into three main components: (a) face detection, (b) feature learning using a CNN architecture, and (c) prediction of facial expressions. Some data augmentation techniques were applied to strengthen the model's training capabilities, including the fine-tuning of hyperparameters to improve the FER system's performance. Further details are shown below.

FACE DETECTION Various challenging issues could affect the robustness of the FER approaches, such as illumination changes, pose variations, occlusions, and individual differences. Compared with other nuisance factors such as illumination, occlusion, and individual difference, pose variation has a greater impact on FER performance, according to conventional FER literature. For this purpose, the Tree-Structured Part Model (TSPM) technique [183] is employed. The TSPM operating principle is based on the combination of trees, with each tree T representing a node with two elements, namely the facial landmarks as parts V and the connection between those parts, i.e., E. As a result, $T = (V, E)$ for each tree. The TSPM performs a global composition of capturing topological changes owing to multiple viewpoints of the facial region, utilizing a separate template that is a mixture of each tree, i.e., $T_m = (V_m, E_m)$, where m indicates a mixture and $V_m \subseteq V$. It computes the Histogram of orientated Gradient (HoG) descriptors for each template and applies the Tree-structured component model to those descriptors to find 68 landmark points for the frontal face region and 39 landmark points for the profile face region. Figure 3.5 shows the TPSM-based face detection process and some examples of the detected faces in different poses and expressions used in the proposed system.

IMAGE AUGMENTATION AND DEEP CNN Training a CNN on limited datasets makes it prone to overfitting, which hinders its ability to generalize to unseen invariant data. One of the potential solutions is image augmentation [139], a regularization

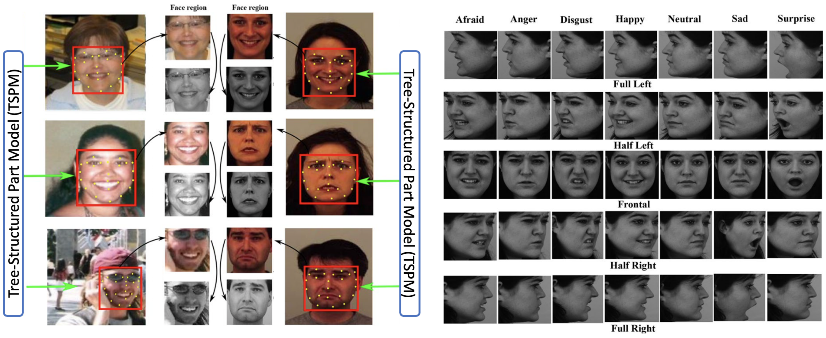


Figure 3.5: On the left, the TPSM-based face detection. On the right, some samples of the detected faces in different poses and expressions.

strategy in DL models that artificially adds new samples to the dataset via label-preserving modifications. This method, which is used to enhance performance, generates similar new samples based on the original one, allowing the model to learn from further examples. Classic image augmentation techniques mostly include geometric transformations and other image processing functions. Image sharpening, image smoothing, and affine transformations (rotation, flipping, reflection, shearing, and scaling) are the augmentation techniques applied to enhance our CNN model. These transformations encode many of the previously discussed invariances that present challenges for facial expression classification and, more generally, in image recognition tasks. Edge enhancement techniques such as bilateral filtering, unsharp mask filtering, and image sharpening not only preserve edge information but also increase tone mapping and contrast stretching during multiscale image decomposition for feature extraction. Image noise that reduces texture information is suppressed using the image smoothing technique. Image rotation and flipping increase the number of training examples, improving the classification mode's robustness and effectiveness for unknown test samples. Finally, image scaling, zooming, and shear mapping all contribute to the image resolution by expanding or reducing the image size with more or fewer pixels. Figure 3.6 demonstrates the effects of some geometric transformations applied to the facial expression images.

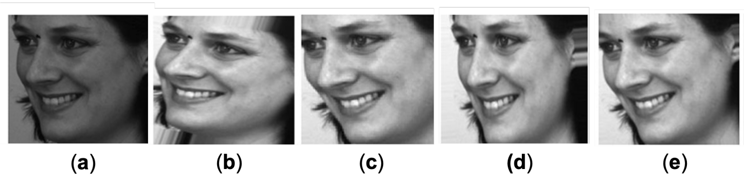


Figure 3.6: Some geometric transformations example: (a) the original face image, (b) result of image rotation, (c) result of image zooming, (d) result of image scaling, (e) result of image shearing.

CNN is a type of DL model for processing data inspired by the organization of the animal visual cortex [85] and designed to automatically and adaptively learn spatial hierarchies of features, from low-level to high-level patterns. A CNN is typically made up of three layers (or building blocks): convolution, pooling, and fully connected layers. The CNN convolution layer extracts features by combining linear and nonlinear operations, particularly the convolution operation and the activation function. A nonlinear activation function is applied to the outputs of a linear operation (such as convolution). The most common nonlinear activation function is the Rectified Linear Unit (ReLU). A pooling layer implements a standard downsampling procedure that reduces the in-plane dimensionality of the feature maps and decreases the number of subsequent learnable parameters. Max pooling is a widely used pooling operation that takes patches from the input feature maps, outputs the maximum value in each patch, and discards all other values. The output feature maps from the final convolution or pooling layer are usually flattened - that is, made into a one-dimensional array of numbers - and connected to one or more dense layers, also known as fully connected layers, where each input is connected to each output by a learnable weight. A subset of fully connected layers transfers the features to the network's final outputs, such as the probabilities for each class in classification tasks, after the features have been recovered by the convolution layers and downsampled by the pooling layers. There are typically the same number of output nodes as classes in the final fully connected layer. A nonlinear function such as ReLU follows each fully connected layer. Figure 3.7 shows the general architecture of the proposed CNN. As can be seen, there are six blocks (with each block containing convolution layers, batch normalization, activation, maxpooling, and dropout layers), two fully connected layers, and three dense layers, the last of which provides the probability scores for the seven facial expression classes (and two facial expression types for the GENKI-4K database). Batch normalization is a type of supplementary layer that adaptively normalizes the input values of the following layer, thereby reducing the risk of overfitting and boosting gradient flow through the network, enabling higher learning rates. Dropout is a recently introduced regularization method that makes the model less sensitive to individual network weights by randomly setting activations to 0 during training. The activation function applied to the last fully connected layer differs from that of the preceding ones. For each task, the proper activation function must be determined. In our scenario, the activation function employed for the facial expression multiclass classification task is a softmax function, which transforms output real values from the last fully connected layer into target class probabilities. Our CNN model is described in detail in Table 3.2.

EXPERIMENTAL RESULTS The experiments are conducted on two well-known FER databases: CK+ and KDEF (see Table 3.1). Further experiments are also carried out on the GENKI-4K dataset [163]. The GENKI-4K contains 4000 labelled images of human faces covering a wide range of subjects, facial appearances, lighting, geographic locations, imaging settings, and camera models. 2162 are identified as smiling or happy, while 1838 are classified as non-smiling or non-happy. The images were collected from the Internet in various real-world scenarios, making detection more difficult. Currently, the GENKI-4K database has become the standard dataset for evaluating smile recognition algorithms in the wild. Figure 3.8 shows some image samples of these datasets. Each database is randomly partitioned, with 50% of the data for the training set and the remaining 50% for the test set. To obtain fair evaluations of the proposed method, we use a subject-independent methodology, k-fold cross-validation ($K = 10$), and report the average results. This technique enhances classifier generalizability (the test set does not include training

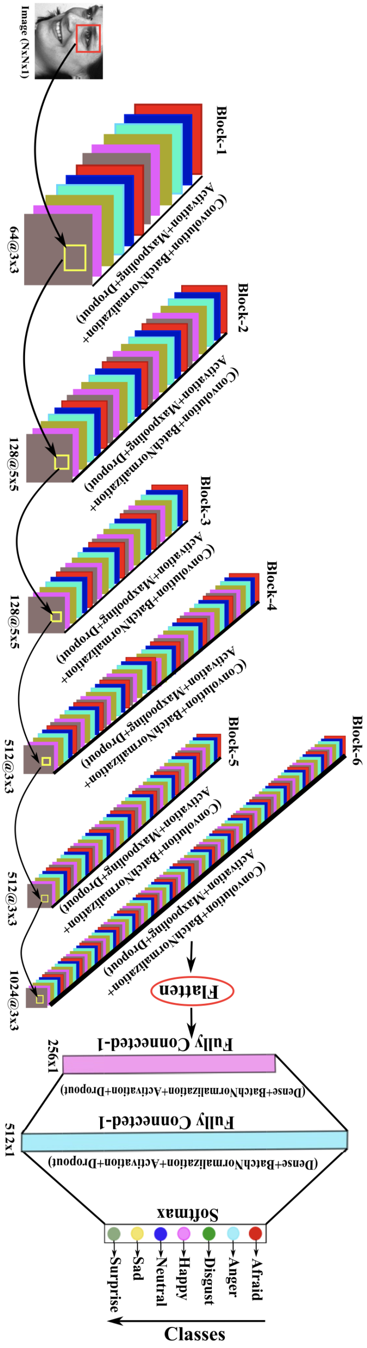


Image (NxNx1)
Block-1
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
64@3x3
Block-2
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
128@5x5
Block-3
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
128@5x5
Block-4
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
512@3x3
Block-5
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
512@3x3
Block-6
(Convolution+BatchNormalization+Activation+Maxpooling+Dropout)
1024@3x3
Flatten
Fully Connected-1
256x1
(Dense+BatchNormalization+Activation+Dropout)
Fully Connected-1
512x1
(Dense+BatchNormalization+Activation+Dropout)
Softmax
Afraid
Anger
Disgust
Happy
Neutral
Sad
Surprise
Classes

Figure 3.7: Architecture of the proposed CNN model.

Table 3.2: Overview of all layers and model parameters of the proposed CNN architecture.

Layer	Output Shape	Image Size	Parameters	Layer	Output Shape	Image Size	Parameters
Block-1				**Block-4**			
Convolution2D (3x3)@64	(n,n,64)	(128,128,64)	((3x3)+1)x64 =640	Convolution2D (3x3)x512	$(n_3, n_3, 512)$	(16,16,512)	((3x3x128)+1) x 512 =590336
Batch Normalization	(n,n,64)	(128,128,64)	4x64=256	Batch Normalization	$(n_3, n_3, 512)$	(16,16,512)	4x512=2048
Activation Relu	(n,n,64)	(128,128,64)	0	Activation Relu	$(n_3, n_3, 512)$	(16,16,512)	0
Maxpooling2D (2x2)	$(n_1, n_1, 64)$ $n_1 = n/2$	(64,64,64)	0	Maxpooling2D (2x2)	$(n_4, n_4, 512)$ $n_4 = n_3/2$	(8,8,512)	0
Dropout	$(n_1, n_1, 64)$	(64,64,64)	0	Dropout	$(n_4, n_4, 512)$	(8,8,512)	0
Block-2				**Block-5**			
Convolution2D (5x5)@128	$(n_1, n_1, 128)$	(64,64,128)	((5x5x64)+1) x 128 =204928	Convolution2D (3x3)@512	$(n_4, n_4, 512)$	(8,8,512)	((3x3x512)+1) x 512= 2359808
Batch Normalization	$(n_1, n_1, 128)$	(64,64,128)	4x128=512	Batch Normalization	$(n_4, n_4, 512)$	(8,8,512)	2048
Activation Relu	$(n_1, n_1, 128)$	$(64, 64, 128)$	0	Activation Relu	$(n_4, n_4.512)$	(8,8,512)	0
Maxpooling2D (2x2)	$(n_2, n_2, 128)$ $n_2 = n_1/2$	$(32, 32, 128)$	0	Maxpooling2D (2x2)	$(n_5, n_5, 512)$ $n_5 = n_4/2$	(4,4,512)	0
Dropout	$(n_2, n_2, 128)$	$(32, 32, 128)$	0	Dropout	$(n_5, n_5, 512)$	(4,4,512)	0
Block-3				**Block-6**			
Convolution2D (5x5)@128	$(n_2, n_2, 128)$	(32,32,128)	((5x5x128)+1) x 128 =409728	Convolution2D (3x3)@1024	$(n_5, n_5, 1024)$	(4,4,1024)	((3x3x512)+1) x 1024= 47,19,616
Batch Normalization	$(n_2, n_2, 128)$	(32,32,128)	4x128=512	Batch Normalization	$(n_5, n_5, 1024)$	(4,4,1024)	4,096
Activation Relu	$(n_2, n_2, 128)$	(32,32,128)	0	Activation Relu	$(n_5, n_5, 1024)$	(4,4,1024)	0
Maxpooling2D (2x2)	$(n_3, n_3, 128)$ $n_3 = n_2/2$	(16,16,128)	0	Maxpooling2D (2x2)	$(n_6, n_6, 1024)$ $n_6 = n_5/2$	(2,2,1024)	0
Dropout	$(n_3, n_3, 128)$	(16,16,128)	0	Dropout	$(n_6, n_6, 1024)$	(2,2,1024)	0

Layer	Output Shape	Image Size	Parameter
Flatten	$(1, n_6 \times n_6 \times 1024)$	(1,4096)	0
Dense	(1,256)	(1,256)	(4096+1)x256= 1048832
Batch Normalization	(1,256)	(1,256)	1024
Activation Relu	(1,256)	(1,256)	0
Dropout	(1,256)	(1,256)	0
Dense	(1,512)	(1,512)	(256+1)x512= 131584
Batch Normalization	(1,512)	(1,512)	2048
Activation Relu	(1,512)	(1,512)	0
Dropout	(1,512)	(1,512)	0
Dense	(1,7)	(1,7)	(512+1)x7= 3591
Total Parameters for Image ($\mathcal{I}$) size (128 x 128)			**9481607**

subjects) [66]. It is widely acknowledged that evaluations without overlapping subjects are more standardized and equitable. The objective of the first experiment is to evaluate the effectiveness of multiresolution and multiscaling images of various sizes, including 48×48, 64×64, 96×96, and 128×128. In ML, an epoch is

Figure 3.8: Some examples from the FER datasets. From top to bottom: KDEF,GENKI-4K, CK+.

defined as the number of times an algorithm "sees" a dataset. In other words, it specifies the number of epochs, or full passes, of the entire training dataset through the algorithm's learning procedure. Many hyperparameters have to be tuned for a robust CNN that can properly classify facial expressions. One of the most important is the batch size [77], which is the number of images utilized to train the network during each epoch. Setting this hyperparameter too high can cause the network to take too long to converge, while setting it too low can cause the network to oscillate without achieving acceptable performance. For this purpose, the Mini-Batch Gradient Descent technique was utilized with varying batch sizes $(20, 30, 40)$ and number of epochs $(50, 100, 200, 500)$. Figure 3.9 illustrates the performance of the proposed FER system (with different image sizes) based on the trade-off between batch sizes and the number of epochs, without data-augmented training images.

A further experiment is conducted without applying the image augmentation techniques, with the aim of demonstrating their effectiveness. Artificially augmenting training samples has been observed to improve performance by approximately 10% across

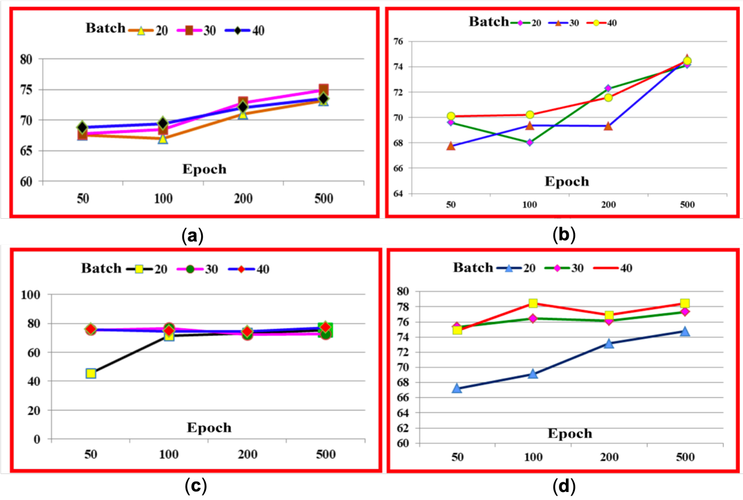


Figure 3.9: Performance comparison of the proposed CNN model's trade-off between number of epochs and batch size.

all experimental datasets. The final results obtained in terms of accuracy (measured by the proportion of facial expressions correctly classified in the test phase) on the KDEF, GENKI-4K, and CK+ databases are presented in Table 3.3. Specifically, the proposed FER system achieved outstanding performance with an image size of 128×128.

Table 3.3: Performance of the proposed FER system in accuracy (%) due to varying image sizes.

Image size	KDEF	GENKI-4K	CK+
48x48	73.67	80.34	87.23
64x64	75.89	84.78	91.87
96x96	78.92	89.45	94.35
128x128	**82.79**	**94.33**	**97.69**

According to a recent comprehensive survey of Deep FER networks [92], our results are highly competitive with state-of-the-art methods. In particular, on CK+ database, our model achieved 97.8%, a result very close to the work of Zhang et al. [177], which reported 98.9% accuracy (but on six facial expressions). In Table 3.3, it is evident that the performance of KDEF dataset is lower than that of the other databases involved in the experimentation. This is due to the following reasons: although KDEF is a lab-controlled database, it includes many visual facial expression patterns from different viewpoints, resulting in an overall recognition effect that is not ideal. Consequently, some literary works employ only frontal images. Nonetheless, the accuracy obtained is competitive with several recent studies [121, 147].

3.4 IFEPE: ON THE IMPACT OF FACIAL EXPRESSION IN HPE

HPE is applied in a wide range of application fields, from surveillance to user authentication, from autonomous systems to human-robot interactions (Chapter 2). Facial expressions are known to impact HPE evaluation errors, particularly if faces are captured in real-world scenarios. Based on this premise, in [26] we investigated the correlations between facial expressions, head pose errors, and facial keypoint distances. In particular, the aim of our study is to highlight the quantitative relationships between HPE errors and facial expressions, identifying and relating the axis most affected by the error when a specific type of facial expression is observable. To the best of our knowledge, this study is the first of its kind.

To extract pertinent information for our research, we have chosen two distinct techniques. Our HP^{2}IFS method [24] is used to perform HPE. As extensively described in Section 2.5.1, this approach is based on the concept of PIFS and, consequently, fractal compression. PIFS allows to reconstruct an image using self-similarities in the image itself. So, fractal image compression is adopted to evaluate the self-similarity of two pose images that result in a similar head rotation. In contrast to the HP^{2}IFS technique, we prefer an DL-based framework for FER, namely Facial Motion Prior Networks (FMPN), presented in [41] (because it produces faster results than training-free algorithms). The FMPN comprises three networks, as shown in Figure 3.10: *Facial-Motion Mask Generator (FMG)*, *Prior Fusion Net (PFN)*, and *Classification Net (CN)*. The goal of FMG is to develop a mask, specifically a facial-motion mask, that emphasizes the moving areas of a grayscale expressive face. PFN combines the original input image with the FMG-generated facial-motion mask to bring domain knowledge to the entire framework. CN is a common CNN for feature extraction and classification.

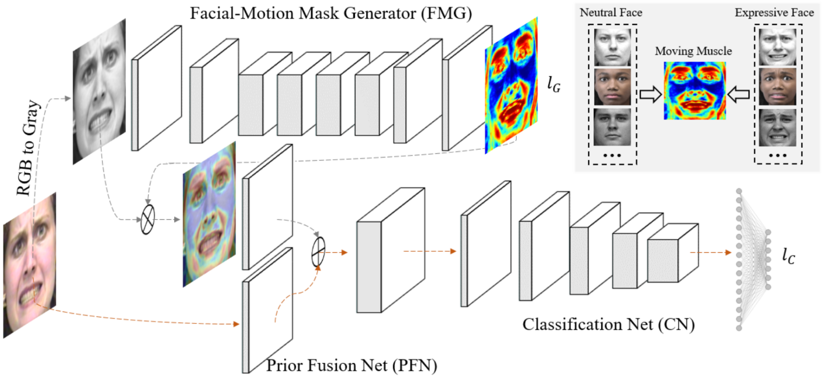


Figure 3.10: The architecture of the FMPN framework [41].

As previously discussed, facial expressions are produced by the contraction of facial muscles, and individuals with the same expression share a similar pattern. Thus, in FMG, for a certain type of facial expression, muscle-moving areas are modeled as the difference between an expressive face and its corresponding neutral face. On the other hand, the similarity characteristic is modeled by averaging the above differences of all training examples in the same facial expression category. The ground truth mask $I_m(k)$ is built for a k-th type of facial expression as follows:

$$I_m^{(k)} = f(\frac{1}{N_k}\sum_{i}^{N_k}|g(R_{e,i}^k) - g(R_{n,i}^k)|) \tag{3.1}$$

where $R_e^{(k)}$ denotes the raw face with the k-th type of facial expressions, $R_n^{(k)}$ is the corresponding neutral face, N_k represents the number of faces in the k-th expression category, and $g(*)$ and $f(*)$ define the pre-processing and post-processing, respectively. The Mean Square Error (MSE) is used for the training objective function of FMG.

Prior Fusion Net (PFN) fuses the original face input with the mask learned using FMG. Specifically, PFN produces a fused output I_s by a weighted sum, defines as:

$$I_s = w_1 * I_{e'} + w_2 * (I_e \otimes f_G(I_e))) \tag{3.2}$$

where $I_{e'}$ is the RGB image version of I_e, $\otimes$ is the element-wise multiplication between the face and its corresponding mask, and w_1 and w_2 are weights of convolutional layers updated during the training. After PFN, the fused output I_s will then be fed into a CNN-based classification network. Three well-known popular databases were used to train and evaluate the FMPN framework, namely CK+, MMI, and AffectNet (Table 3.1) , achieving considerable performance compared to state-of-the-art methods.

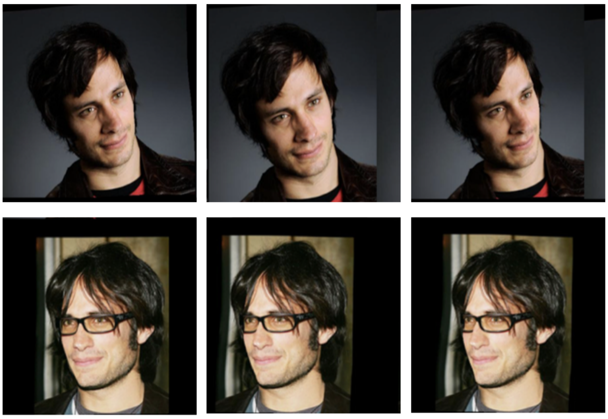


Figure 3.11: Images from 300W-LP 300W_lp dataset.

Our experiments are conducted on the 300W_lp dataset [184]. The 300W_lp dataset is an extension of the 300W, which standardizes 68 landmark points for several face alignment benchmarks, such as AFW, LFPW, HELEN, and IBUG. There are a total of 61,225 images, consisting of 17,860 from IBUG, 5,207 from AFW, 16,556 from LFPW, and 37,675 from HELEN. As illustrated in Figure 3.11, digitally generated rotations do not respect all of the relationships between facial keypoints, resulting in higher HPE errors along the three axes than those reported in [24].

To evaluate the correlations between the HPE and facial expression results, we examine the variations in facial point distances during expressions. Due to the large diversity in expression per subject, we analyze samples from the dataset described in [51]. The classification of the seven expressions, including neutral, by the FMPN algorithm is illustrated in Figure 3.12-(a). Then, using the same method as HP^{2}IFS, facial landmarks are detected (Figure 3.12-(b)). Since the landmark detector is equivalent, HPE will verify the same scenario regardless of the presence of wrongly positioned landmarks.

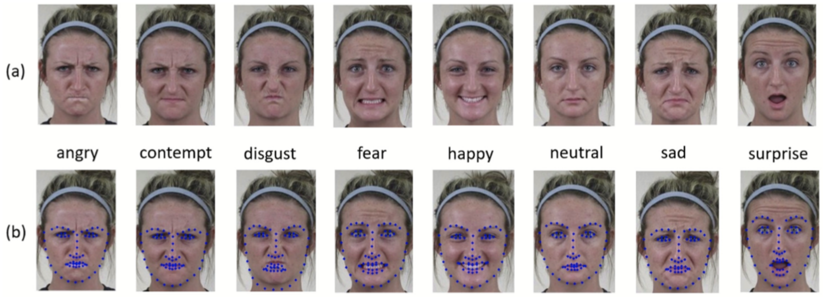


Figure 3.12: Facial expressions considered for the distances analysis (a) and their landmark locations (b). Sample images from [51].

Each landmark point is numbered and occupies a fixed location within the array; for instance, the nose is always in the 33rd array position. On the basis of these arrays, we established some distances that are crucial for facial recognition due to their changes in facial expressions:

- *Eye_l*: the left eye opening, i.e., the distance between the landmark corresponding to the eyelid and the landmark related to the lower part of the eye.
- *Eye_r*: the right eye opening, computed identically to EL for the right eye.
- *H_Mouth*: the horizontal opening of the mouth, measured at its farthest horizontal points.
- *V_Mouth*: the vertical opening of the mouth, measured as the distance between the mouth's highest and lowest points.
- *Eyeb_l*: The distance between the left eye and the left eyebrow (in the center).
- *Eyeb_r*: the distance between the right eye and the right eyebrow (in the center)

- *Chin_Mouth*: the distance between the chin and the mouth's lower landmark.

After computing these distances, we assessed their variation using a neutral expression as a reference. Table 3.4 presents the percentages of increasing or decreasing distances in relation to the facial expression.

Table 3.4: The percentages of variation of relative distances for different facial expressions.

Expr/Dist %	Eye_l	Eye_r	H_Mouth	V_Mouth	Eyeb_l	Eyeb_r	Chin_Mouth
Angry	-24.55	-25.58	-7.22	-10.97	-37.19	-36.75	11.60
Contempt	-25.58	-24.55	-2.46	-10.97	-36.26	-44.25	19.23
Disgust	-25.58	-37.98	-4.84	21.98	-12.28	-22.14	34.67
Fear	-0.77	0.00	23.76	98.78	-12.50	-16.61	-3.85
Happy	-37.98	-25.58	30.83	76.69	0.20	-5.70	3.92
Neutral	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sad	-25.58	-24.55	-0.08	-0.61	-12.28	-22.34	15.45
Surprise	24.65	24.65	-11.97	220.25	25.00	10.94	-11.54

Figure 3.13 shows a histogram to better visualize those values. First of all, the V_Mouth distance is greatest for surprise, followed by fear and happiness. Except in cases of surprise, the Eye_l and Eye_r decrease in all cases. The behavior of anger and contempt is mostly similar. In disgust, the Eye_r distance changes (relative to angry and contemptuous expressions) and the V_Mouth increases. Sadness emerges as the emotion with the fewest noticeable changes. At this point, we evaluated the angular error values produced by the HPE method. Specifically, each of the 300W-LP, AFW, HELEN, IBUG, and LFPW datasets was sequentially chosen as the method reference model. For the remaining images, we extracted the angular error along the three axes. The same images are used in FMPN to classify the subject's facial expressions. The result is a set of relative errors in facial expressions, divided by the provided reference model. So, we estimated the percentage increases or decreases of the errors relative to the neutral image. However, the test data obtained is not uniform in terms of the image numbers for each facial expression. As a result, we evaluated the weighted mean relative errors, shown in Figure 3.14.

As regards the rotation axes, it can be stated that:

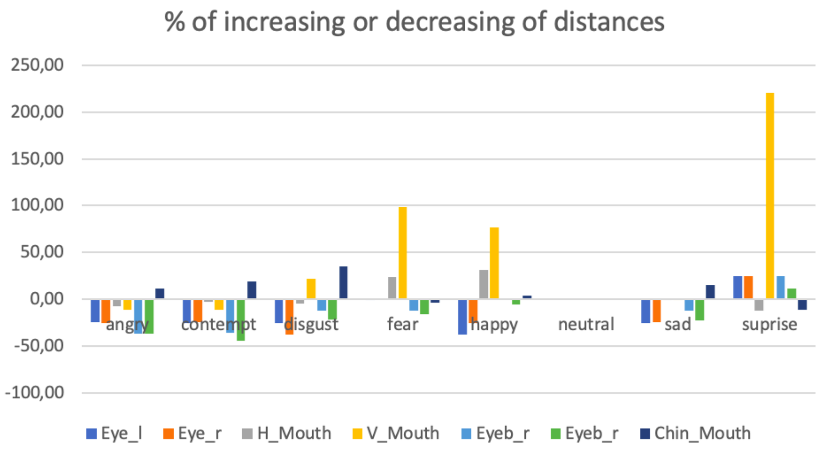


Figure 3.13: Histogram of increasing or decreasing distances in different facial expressions.

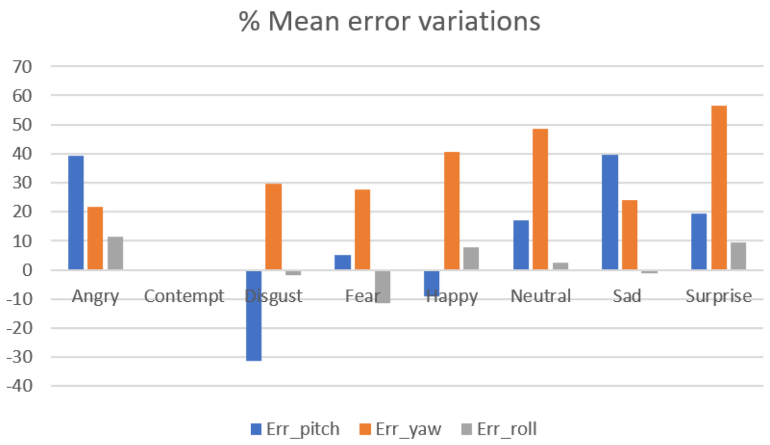


Figure 3.14: Histogram of the mean weighted error percentage increasing or decreasing with respect to the facial expressions.

- *Pitch error* for the expressions contempt, fear, happiness, and, in particular, disgust is decreasing. The same is increasing for anger, sadness, and surprise.

- *Yaw error* is decreasing for angry, disgusted, fearful, happy, sad, and especially contempt. Surprise is the only expression that is increasing.

- *Roll error* for the expressions contempt, disgust, sadness, and, in particular, fear is decreasing. The same is increasing for anger, happiness, and surprise.

In the final step of our analysis, we show the direct relationships between facial point distances and relative errors. The relationships were defined as being *directly proportional* when an increase in distances leads to a rise in errors and vice versa. On the other hand, we evaluated that the relationships are *inversely proportional* when an increase in distances implies a decrease in errors and vice versa. After establishing our criteria, the final considerations are as follows:

- The distances between eyes and eyebrows (Eyeb_l and Eyeb_r) are directly proportional to yaw errors.

- The vertical mouth opening (V_Mouth) is directly proportional to the yaw errors.

- The horizontal mouth opening (H_Mouth) is inversely proportional to roll errors.

- Mouth variations (V_Mouth and H_Mouth) in the absence of eye variation (Eye_l, Eyeb_r, Eyeb_l, and Eyeb_r) are inversely proportional to roll errors.

- The distance between the mouth and chin (Chin_Mouth) seems to be related to pitch.

Due to the mismatched values obtained in the expressions of anger and contempt, we cannot say with certainty whether the relationship between Chin_Mouth and pitch errors is direct or inverse. This can be related to the HPE method chosen to perform the error evaluations. The HP^{2}IFS technique, in its original form, presents the larger error values precisely on pitch. In addition, the error is less homogeneous along this axis, since having some outliers may result in a larger error than yaw and roll. Despite this, it is obvious that facial point distances and HPE method errors are directly related, and facial expression estimation may be able to derive this relationship efficiently.

To further demonstrate this claim, we selected the facial expression with the minimum mean error in HPE method along

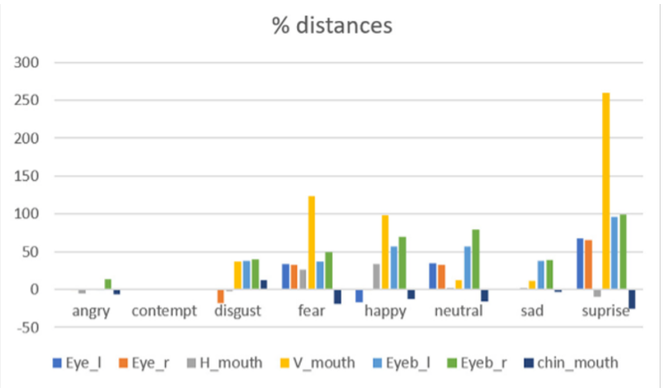


Figure 3.15: Histogram of the percentage of distances increasing considering the expression "contempt" as a reference.

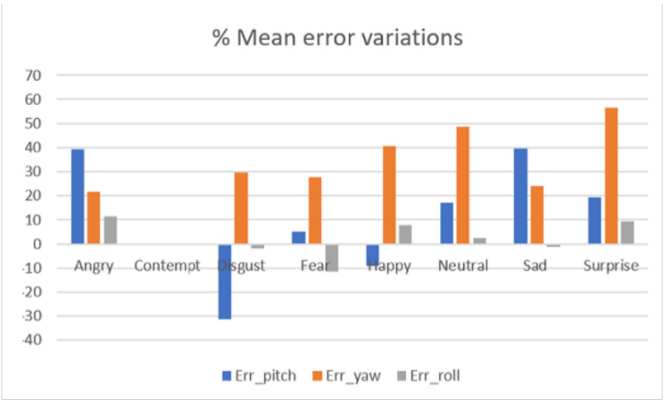


Figure 3.16: Histogram of the percentage of mean error increasing considering the expression "contempt" as a reference.

the entire dataset, i.e., the expression of contempt. We show the histograms of Figures 3.13 and 3.14, respectively, in Figures 3.15 and 3.16, using the expression "contempt" as a reference. As can be seen, all the distances are increasing, as are the errors. The disgust expression is the only relevant exception, with overall distance increasing by small values and eye distances decreasing, with a relevant pitch error decreasing.

3.5 CONCLUSIONS

FER is an important research problem in the AI domain due to its widespread applications in both academia and industry. While FER can be accomplished using a variety of sensors, research indicates that the use of images and/or video is superior, as visual expressions convey significant emotional information. Recently, DL approaches have been increasingly implemented to deal with the challenging factors for emotion recognition in nature. Motivated by this, we designed a CNN-based model with the aim of recognizing expression types in images. We focused the most on sophisticated data augmentation techniques as well as the fine-tuning of hyperparameters to solve the emotion recognition task. The obtained results were very encouraging, with the proposed CNN architecture achieving nearly 98% accuracy on the CK+ database, one of the most widely used test beds for the development and evaluation of FER algorithms. Although facial expression recognition based on 2D images might achieve promising results, facial expression is simply a component of human behavior. For expression classification, we only investigated seven distinct and unique emotion classes. Furthermore, most of the datasets used in the experimental phase and training procedure contain frames acquired in the laboratory. Consequently, in the future, we intend to analyze classes of compound emotions while also examining images from real-world scenarios.

Despite the powerful feature learning ability of deep learning, there are still difficulties when applied to FER. Deep neural networks require massive amounts of training data to avoid overfitting. Existing facial expression databases, with a few exceptions, are insufficient to train the well-known neural network with deep architecture. Due to changes in personal traits, posture, illumination, and occlusion, there are also significant intersubject discrepancies when facial expressions are acquired in unconstrained scenarios. Another major issue that requires consideration is that, due to the ease of data processing and their availability, a significant number of studies conducted expression recognition tasks utilizing 2D static images without addressing temporal information. In this regard, 3D FER with 3D face shape models and depth information can capture subtle facial deformations that are naturally resistant to pose and lighting variations. In light of the numerous open problems in this field that may inspire new techniques to improve FER systems in the future, we can conclude that FER will play an increasingly vital part in our daily lives. Future human-machine environment applications will be multimodal, combining additional information from dynamic behavior, voice, text, audio, and image, so as to make machines' use as intuitive and natural as feasible.

4 TOUCH DYNAMICS IN MOBILE DEVICES

TD, one of the most powerful behavioral biometrics, captures a person's typing patterns on mobile touchscreen devices. When an individual interacts with such devices, a digital signature is generated that is highly discriminatory and unique. Since most touchscreen devices already include sensors, this technology can be widely adopted to implement continuous authentication methods, making the system even more secure and reliable to prevent access by impostors.

This Chapter explores the integration of soft biometric traits with TD-based behavioral biometrics. The goal is to analyze users' typing patterns for demographic classification in age, gender, and user experience. Using traditional lightweight ML classification algorithms, it is possible to achieve effective demographic analysis as well as contribute to improving the identification mechanism.

4.1 MODELLING TOUCH AND TYPING BEHAVIOR

Biometric technology is becoming increasingly widespread and accepted by society, primarily because of its success on mobile devices. Traditional biometric modalities used in modern smartphones include fingerprints, iris, and face recognition. The research community has explored novel identifying systems based on behavioral biometric traits such as gesture, keystroke, and gait as a result of the weakness of conventional authentication mechanisms. It has been demonstrated that behavioral biometrics provide higher security than physiological features and can be used in a mobile multimodal authentication system. This is because mobile devices have multiple sensors that are capable of simultaneously acquiring a vast amount of behavioral biometric data. This data can also reveal a significant amount of information about the user [141].

Nowadays, increasingly difficult online operations are performed via mobile devices due to their portability and ease of use compared to large desktop systems, as well as the availability of fast wireless Internet connections. These activities often require the submission of sensitive and valuable data, such as personal identifiers, passwords, bank accounts, credit card information, and so on. As a result, serious issues arise with the security and privacy of such data. In order to control production costs, considerable attention is paid to the creation of solutions that do not use dedicated hardware. In this circumstance, using TD as a biometric identifier on a mobile device seems like a natural choice. Based on interactions such as typing rhythm, finger-swiping speed, and device-holding posture, TD-based biometric approaches generate a behavioral model of a user. Touchscreen gestures provide user discrimination since they represent an indication of muscle behavior [57]. The built-in sensors on the mobile device record touch data such as timestamps, finger pressure, and finger area in contact. This raw data contributes to creating a user model for enrollment, which is then used for identification and verification purposes. There are many research papers that deal with the study of how people interact with smartphones and, specifically, how each user performs the screen pressure phase. The way in which each human being interacts with the touchscreen (and, in turn, with the device) can be seen as a kind of digital signature closely related to the individual interacting with the system. Such a signature is considered a distinctive element that uniquely identifies a particular person. During the Second World War, it was not uncommon for telegraphists to state with certainty whether or not a message was typed by the same operator. Since 1980, experimental studies have confirmed the existence of discriminative features in each subject's typing behavior. These are typically considered to be the first documented cases of identifying a human being by looking at how it interacts with a keyboard [110]. Numerous efforts in this field have been made since then, in particular with physical keyboards. Clearly, the situation just described is from the last century, but thanks to modern technologies and the increase in the use of smartphones, performing the same recognition is becoming very simple. As already mentioned, recording these particular traits does not require the use of any additional hardware because a person's regular typing rhythm may be obtained by utilizing a simple keystroke logging software to capture the timings of key-related interactions. Furthermore, due to the presence of multiple embedded sensors, some not necessarily designed for biometric identification, modern smartphones constitute a particularly suitable environment to perform TD-based recognition. In the context of mobile applications, the user's touch pattern has evolved into a non-intrusive biometrics model that can be implicitly captured. Additionally, it provides a better balance between security and usability because the user's touch-interaction behaviors are not considered private information. Active and continuous authentication, offered by touch biometrics, can be summarized as the continuous confirmation of the identity of a person based on specific features of their behavior when interacting with a computing device. TD-based authentication maintains a constant state during the entire time the user interacts with the device, thanks to periodic, transparent re-authentication tasks. Without interfering with the user's activities, the entire process can be run in the background. Over the years, many existing classification techniques have been utilized in touch biometrics research. Classical statistical methods as well as advanced ML methodologies were applied. K-Nearest Neighbor classifiers, K-mean methods, Bayesian classifiers, Fuzzy Logic, Boost learning, Random Forests and Support Vector Machines are some of the most commonly used ML approaches. Several metrics, including Euclidean distance, Mahalanobis distance, and Manhattan distance, were also used. However, because the research used diverse datasets and evaluation criteria, a valid comparison of numerous approaches is not possible [137].

Although a user's typing is highly unique and governed by a person's neurophysiological path, it can also be influenced by their psychological state [72]. It should be noted that human touch interactions show high instability due to numerous transient stimuli such as emotions, stress, etc. External factors, such as the input keyboard device used, which could have a different layout of the keys, also have an impact [181]. The plethora of features offered by mobile devices allows for the capture and storage of a wide variety of behavioral and physiological attributes.

Recently, several studies have suggested that by combining different biometric traits to form a multimodal model, the accuracy of the recognition system could be further improved. For example, the front camera enables users to interact with the device screen even with eye movements, thus opening up a new dimension of data collection in real-world conditions. In light of this, it may be easy to consider the gaze as an additional biometric aspect [35, 36]. On the assumption that each mobile device is used by a single user, behavioral profiling-based active authentication mechanisms have also been developed. Behavioral profiling describes the user's interactions with mobile sensors and services [109]. Mobile devices involve multiple sensors (camera, gyroscope, magnetometer, accelerometer, GPS location, touchscreen, etc.) and full connectivity (e.g., Bluetooth, WiFi, 4G, app usage). All of this information is generated by the user's normal smartphone usage, and it has been proven that it can be exploited for person identification under specific conditions [5]. In the context of mobile devices, Figure 4.1 illustrates the sensors and services that can be utilized to obtain behavioral biometric data.

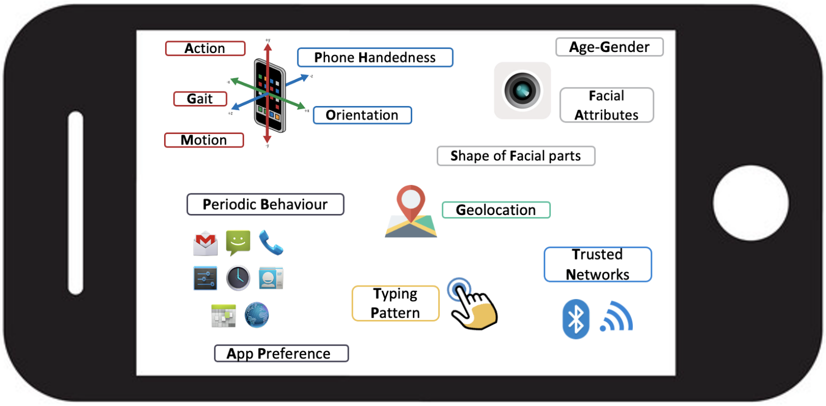


Figure 4.1: Sensors and accessories available in a mobile device.

4.2 TOUCH DYNAMICS SYSTEM: OVERVIEW

Touch biometrics-based continuous user authentication is similar to a classic biometric recognition system in that it involves an enrollment phase and a verification or identification phase. During the enrollment phase, the scheme collects the touch features from the user's touch gestures, generates a template for each individual, and stores it as a user profile in a database. During the recognition phase, the user's touch data is recorded in order to extract the relevant key information. Using a classification algorithm, the system compares these features with the user's profile. The current user is identified as legitimate if the classifier scores are higher than the predefined threshold. Otherwise, the user will be labeled as an illegitimate user. The main modules involved in the enrollment and recognition phases are described below.

DATA ACQUISITION The choice of the device for data acquisition is critical. Smartphones are used as data collection devices for the vast majority of research work in the field of touch dynamics. This is partly caused by the fact that more people use smartphones than, for example, digital tablets. Modern devices typically have sensors with higher resolution and accuracy, capable of recording higher-quality features. In addition, modern devices have more computing power and resources, which makes it easier to apply more sophisticated algorithms and powerful sensors. The proposed development platform associated with a mobile operating system is an additional criterion for device selection. Additionally, modern devices have more computing power and resources, which makes it easier to apply more sophisticated algorithms and powerful sensors. To collect touch dynamics data, a toolkit must be used. According to a literature review [151], Android is the most widely used development platform for acquiring touch dynamics data, followed by iOS and Windows. Unlike its rivals, Android offers open-source library functions that enable developers to modify the application framework, providing them with more freedom in designing and customizing their applications.

Data collection can take place in two possible contexts: under strict supervision or not monitored at all. The vast majority of studies published in literature have been conducted in monitored and controlled situations. Controlling an experiment is done primarily to reduce the degree of variance in touch dynamics patterns induced by external variables such as cognitive load, distractions, and so on. The main experimental variables (for example, the discriminative abilities of the characteristics or the accuracy of the performance of classifiers) can be evaluated more correctly when an experiment is tightly controlled with a rigorous methodology. However, compared to the results found in uncontrolled settings, the performance obtained in supervised settings is overly optimistic. Two distinct scenarios, *Fixed text* and *Free text*, can be considered for touch data acquisition on both desktop and mobile frameworks. Fixed text uses personal identification numbers (PINs), passwords, or passphrases. The PINs used to log in to mobile phones are normally four digits long, but longer number sequences can also be used depending on the applications evaluated and the desired level of security. The sequences of alphabetic or numeric characters make up the passwords. The standard password length is between six and fifteen characters, including special characters, which are typically used to increase security. Last but not least are passphrases, in which users type sentences or paragraphs that are reproduced during the recognition phase after being preset during enrollment. The length of the inputs in this situation can be in the order of several dozens of characters. Unlike in previous cases, users can enter any free text-related information, regardless of what was previously collected during the enrollment phase. Typing free text is obviously much more challenging in the context of recognition. Since they can only be used once before the sequence becomes unusable, one-time passwords can be seen as a specific example of free text input.

FEATURE EXTRACTION Various signals can be employed to characterize users by analyzing the raw data collected from a subject. *Timing*, *spatial*, and *motion* are the three categories of common features described in literature [151]. Most of the methods that use TD for biometric recognition rely on timing data to distinguish the users. The availability of timestamp information, which records the precise moments when certain keys are pressed and released, is typically provided by mobile devices. Several measures can then be computed from such information. The first one is called the *Dwell Time (DT)*, and it describes how long a touch event lasts when the same key is pressed. In literature, it is often referred to as "interval," "press," or "hold time." This value is calculated by subtracting the key release timestamp value from the key press timestamp value. *Flight Time (FT)* describes the time period between the touch events of two successive keys. It is also known as latency. There are four different FT variations, as shown in Figure 4.2. In more detail, given an input string and any two distinct characters within it, the extracted main features represent the timing differences between two events, i.e., press/press, release/release, press/release, and release/press. Different feature lengths can be extracted to derive a timing feature. The timing feature that is obtained by taking the touch event timestamp values of the same key is referred to as "*uni-graph*". Similarly, the terms "*di-graph*" and "*n-graph*", respectively, refer to the timing features that can be derived from two or more keys. Figure 4.3 represents the different n-graph lengths. The typical number of words the considering user may type in a minute is one metric that can be used to estimate the user's typing speed from the TD data collected. Other metrics, for example, the adjusted words per minute that take typing errors into account or the keystrokes per second, which include the backspace key, are also considered.

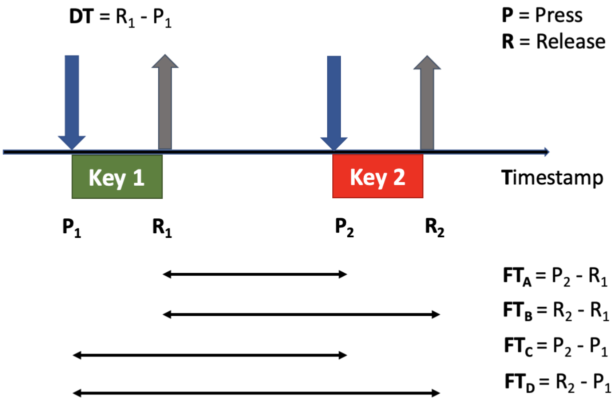


Figure 4.2: Touch Dynamics features: Dwell Time and Flight Time.

A touch event can result in the acquisition of a spatial feature, which is a property connected to physical interactions between a fingertip and a device's touchscreen surface. Touch size, pressure, and position are the three spatial characteristics that are most frequently mentioned in literature. The touch size is related to the screen area touched during a touch event. The size of the subject's fingertip influences the user's "touch size" value. According to several studies, adult male subjects often produce larger touch size values than do children or female subjects. A touch pressure value represents the approximate force applied to the screen upon each touch event. It is important to note that each individual's particular finger muscle is connected to a touch pressure value. As a result, it is challenging for a user to replicate another subject's touch. Finally, the touch position, which captures the location where a fingertip lands on a device screen, is a two-dimensional matrix feature (or key). A pixel-based x and y coordinate can be assigned to each touch event. The size of the fingertip and an individual's cognitive preference influence where a user should touch a key. Thus, it is possible to identify a subject by using the touch position as a discriminating feature. The touch position can be described in one of two ways: as an offset from the key's center or as the absolute coordinates of a touch event relative to the full screen. Some mathematical manipulations can also be used to derive new features, i.e., the distance or angle between two touch events. This coordinate representation does have a drawback, though, in that the screen's coordinate system is device-dependent. The values of the collected touch positions are inconsistent between devices. Therefore, touch posi-

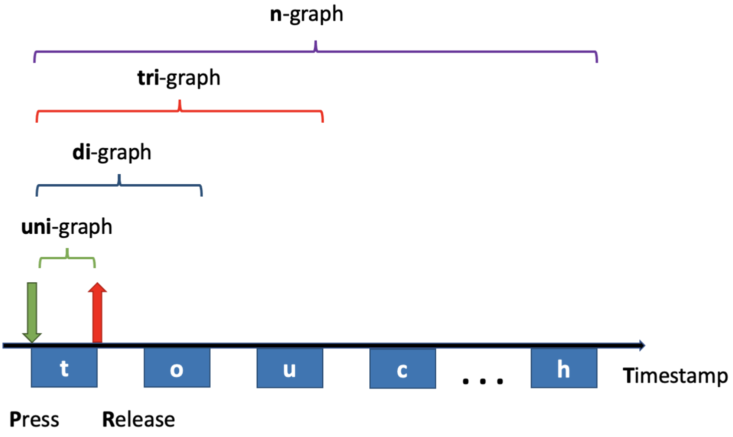


Figure 4.3: The different timing feature lengths: n-graph.

tion values should be normalized unless data collection is carried out on similar devices.

The accelerometer and gyroscope, two hardware motion sensors, are built into most modern mobile devices. Typically, each touch event causes the device to move or rotate slightly. The accelerometer sensor measures the rate of linear acceleration applied to a device over time. It is designed to track motion along the x, y, and z axes in both positive and negative directions. The gyroscope sensor, on the other hand, measures the speed of rotation that a device undergoes in relation to the three axes: forward and backward tilt (pitch), side-to-side rotation (roll), and vertical-to-horizontal rotation (yaw). Figure 4.4 shows the various motions detected by both sensors.

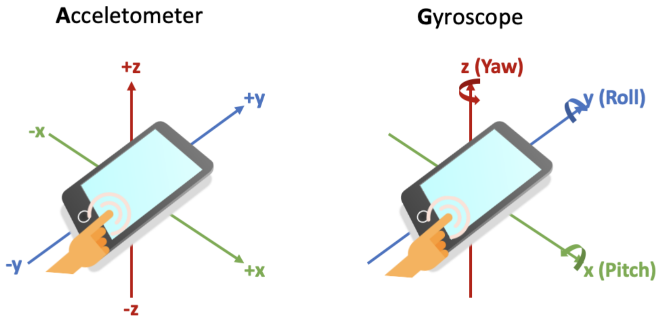


Figure 4.4: The different motion data recorded by the mobile sensors.

CLASSIFICATION/MATCHING During the enrollment phase, methods such as distance-based metrics and ML algorithms are utilized to generate a *user profile*. The authentication task can be viewed as a two-class classification problem (legitimate users vs. impostors), in which the classifiers analyze touch behavior data and distinguish between legitimate users and impostors. If the module adopts a one-class classification problem (as opposed to conventional binary classification), the information is only available to a single class, i.e., legitimate users. Each classifier has a training phase in which a set of feature vectors is used to build a model of the user's touch behavior, followed by a test phase where each new test vector is assigned a classification score. In the identification scenarios, the recognition phase is carried out by selecting probe samples from the considered users and making a decision about the identity of their owners by exploiting trained ML approaches to perform one-to-many comparisons [174].

4.3 SOFT BIOMETRIC TRAITS THROUGH TYPING PATTERNS

In literature, soft biometric traits are defined as "*characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*" [79]. Automatic recognition of common personal traits, such as age range and gender, is a current research topic. There are a number of possible applications for estimating a user's soft biometric attributes while using a mobile device. By estimating the age, it is possible to prevent access to particular multimedia content by people belonging to certain age groups. Conversely, gender detection can be used for marketing purposes to promote personalized products, while touch-experience detection would improve the usability of the device. These features could also be used to improve the performance of the authentication systems by combining their contributions with other reliable biometric traits (such as fingerprints, faces, eyes, and the like). Pupil size [31, 37], gait [14], and the 2D human skeleton [15], among others, can be exploited to extract such soft information with convincing performance. Concerning the use of smartphones and similar portable devices, demographic information can be inferred from the keystroke dynamics resulting from interaction with a keyboard. There are several works that, over the years, have explored this interaction and obtained interesting performances for different soft attributes. Tsimperidis and Arampatzis [154] demonstrate how this biometric attribute can be used to determine the user's gender, age, handedness, and level of confidence with the hardware. With the rapid development of technologies, and in particular touchscreen mobile devices such as tablets or smartphones, research activities have also focused on so-called touch biometrics. Numerous studies have used this new biometric trait to recognize a subject who uses a touch device, either by relying solely on the information obtained from this type of interaction or by combining it with other biometric features. However, there are currently no exhaustive works in literature that address the situation mentioned above. Additionally, almost all of the research utilizes verification protocols that, by not ensuring that training and testing phases are carried out on unseen users, introduce a biometric bias into the experimentation. The state-of-the-art in soft biometrics using touch screen behavior data reveals great accuracy and performance levels, which leads one to believe that behavioral traits can be consistently processed for soft biometrics analysis. When experimental methods are not strictly one-left-out, the models are more likely to be trained to recognize the subject's identity than the desired soft features. One of the first works to use touch data to acquire information about the gender and subjects' level of experience with hardware is presented in [7]. Due to the extremely uneven nature of the dataset (56 men and 15 women), the authors chose to focus on a subset of only 9 subjects for each class. Performance is 88% with a single stroke (defined as a vector of 5 components, including touch position coordinates, timestamp, pressure, area covered, and the number of points belonging to the hit) and 99% with a sequence of ten strokes. However, they claim to utilize a 3-fold cross-validation, thus not taking into account the fact that there may be samples from the same individual in both the training and testing phases, hence risking affecting the same prediction. The same problem applies to classifying a user's tactile experience. The authors divided users into four categories: inexperienced, moderately experienced, experienced, and very experienced. Experiments were conducted on 24 participants, 6 from each class. The accuracy reported ranges from 81% for a single stroke to 100% for twenty strokes. In each experiment, three well-known ML classifiers were evaluated. Buriro et al. [29] have also conducted research on age, gender, and operating-handedness classification. Using the publicly available dataset TDAS, they test different classifiers, of which Random Forest is the best-performing. The accuracy obtained is 82.8% for gender recognition, 95.5% for operating handedness, and 87.9% for age prediction, through a random selection of 80% of the data samples as training. Even in this instance, merely dividing the samples without taking into account the possibility that training and test samples could contain the same subject could have a significant impact on overall performance. The work in [80], on the other hand, focuses solely on gender recognition by utilizing touchscreen movements and other behavioral data derived from accelerometer, gyroscope, and orientation sensor readings. The authors evaluated their method by employing two distinct devices and, this time, placing samples from the same subject either only in the training set or only in the test set. Considering only a selection of four gestures and applying fusion techniques across the several classifiers employed in the two experiments yields an average accuracy of approximately 90%. There is also a comparison of the individual performances of both the sensors and the touch features, and it is evident that for each gesture, the contribution of the sensors is very strong.

All of the above observations served as inspiration for our work [38], which aims to demonstrate how, through a rigorous protocol adapted to a real case study in which a first-time user uses a touch device, performance decreases dramatically. In any case, it is clear that this biometric feature has the potential to discriminate for this type of classification. To the best of our knowledge, this study is the first to employ soft biometric analysis using the datasets described below. The Touch Dynamics based multi-factor Authentication Solution (TDAS) dataset is the sole exception.

DATASETS Although there are many benchmarks in literature reporting touch dynamics data, there are few publicly accessible datasets that include subjects' soft biometric features. RHU KeyStroke Benchmark [56], Keystroke Dynamics Android platform (KDAp) database [8], and Touch Dynamics based multi-factor Authentication Solution (TDAS) dataset [152] were the three databases we employed in this work. These datasets are frequently discussed in literature for the purposes of identification and authentication.

The RHU Keystroke benchmark has four "key event" features: Press-to-Press (PP), Press-to-Release (PR), Release-to-Press (RP), and Release-to-Release (RR), which store, respectively, the time of the event between two key pressures, one key pressure and one key release, one key release and a key pressure, and finally two key releases. All participants entered the password "rhu.university" 15 times over three different time periods during the acquisition process. The four characteristics mentioned above also have sub-features that indicate the different time values for typing the password: RR, RP, and PP each have 13 data, whereas PR has 14 values. As a result, there are 53 features in total for each subject. In the Keystroke Dynamics Android platform (KDAp) database, the typing patterns of 42 individuals were recorded in a controlled environment in two sessions over the course of two weeks. The users input the same password (.tie5Roanl) 30 times through an Android application created for data collection. The system records the finger area, pressure, and timestamp. The features extracted and examined for a total of 71 user-specific touch dynamics characteristics include the key hold time, the down-down time, the up-down time, the key hold pressure, the finger area, the average hold time, the average finger area, and the average pressure. The Touch Dynamics based multi-factor Authentication Solution (TDAS) dataset consists of 150 individuals with 10 samples per subject. The keystroke timings, touch pressure, and touch size represent the three main characteristics extracted. The two numeric inputs, i.e., 4 digits (5560) and 16 digits (13796666624680852), were typed to acquire these properties. Details of the individuals' soft biometric information for each database are shown in Table 4.1.

Table 4.1: Soft biometrics information of users.

Dataset	**Users**	**Password**	**Controlled acquisition**	**Age range, n° per class**	**Gender**	**User touchscreen experience, n° per class**
RHU	51	rhu.university	Yes	7-17, 11 18-29, 30 30-65, 10	26 male 25 female	-
KDAp	42	.tie5Roanl	Yes	20-46 , -	24 male 18 female	0-7, 24 8-9, 18
TDAS	150	5560 (Short) 1379666624680852 (Long)	No	<20, 69 20-40, 46 >40, 35	45 male 105 female	-

EXPERIMENTAL PROTOCOL Several statistical indices, including maximum, minimum, mean, standard deviation, quantiles, and median absolute deviation, are extracted from each dataset prior to model building. Statistical-based preprocessing techniques are useful to prepare data as input for modeling and/or analysis using ML algorithms. In the experimental phase, we used six well-known ML algorithms: AdaBoost, Decision Tree, Random Forest, Support Vector Machine, Nearest Neighbors, and Gaussian Naive Bayes for binary (i.e., gender and user tactile experience) and multi-class (age) classification. The best hyperparameters of the model were found using the Grid-search tuning optimization technique. For each dataset, we used two different strategies:

- an *independent* dataset subdivision by subject designed to prevent biometric factors from biasing training and testing;
- a *random* subdivision of the dataset where 70% is used for training and the remaining 30% is used to generate the test set.

The use of two distinct sample splitting procedures was guided by the observation that soft biometric analysis is often influenced by subject identification. The way samples are distributed throughout training and testing sets can have a significant impact on results when the goal of the experiment is to classify demographic characteristics rather than authenticate the subject. On the other hand, similar searches in literature often use a random selection of samples for training and testing. This allows for a fair comparison of works, but as the results will show, identity-bias has a significant impact on the level of performance achieved. A further experimental evaluation was performed on the characteristics of each dataset. The goal is to draw attention to which biometric touch feature has the greatest impact on performance as a whole. It is also interesting to see which of the three macro categories (time, pressure, and area covered) found in the three data sets performs better than the others or, conversely, is not suitable for the classification task.

RESULTS In line with the first strategy, the maximum accuracy for estimating gender and age on RHU dataset using Decision Tree and Random Forest is 61% and 68%, respectively. The temporal differences between two key pressures and two key releases constitute the most discriminating factors in both gender and age estimation tasks. The highest accuracies for estimating gender and user tactile experience (obtained with Support Vector Machine and GaussianNB) for the KDap database are 86% and 79%, respectively. It is important to note that the age estimation task could not be applied as the dataset is heavily biased in favor of the age group 21–23. In addition, a binary classification was used to evaluate the user's tactile experience by dividing it into two classes, one for levels 0 to 7 and the other for levels 8 to 9. In this way, it was possible to balance the number of subjects in each of the two groups. The down-down time and the finger area are the most crucial characteristics for the aforementioned estimation tasks. The TDAS database is very unbalanced. The distribution of the samples consists of 105 females and 45 males, as previously stated in Table 4.1. For this reason, we randomly divided the female subjects into two different subsets, performing the experiments on both. By calculating the average of the performances achieved in each of the two subgroups, the results for the gender estimate were obtained. Support Vector Machine and Nearest Neighbors achieved the highest accuracy of 62% and 30%, respectively. Note that regardless of the classifier used, performance is particularly poor for the age estimation task. Finally, on the TDAS dataset, the most distinctive feature for gender estimation is the finger touch size, while the time release and time pressure characteristics are decisive in determining age.

The second strategy adopted in the experiments involves randomly dividing each dataset into 70% for training and the remaining 30% for the test set. The performances achieved are very promising, as can be seen from Table 4.2, which summarizes all of the experimental results. More specifically, we can see that the TDAS dataset reached 92% for gender estimation using Nearest Neighbors (as opposed to 62% in the first strategy). Compared to the results of Buriro et al. [29], our performance is significantly better. The accuracy rate reported by the authors is 82.8%. This is achieved despite the work adopting the synthetic minority oversampling technique (SMOTE) and a 50:50 random train:test division strategy. Regarding age-related results, the TDAS dataset surprisingly reaches 84% with Nearest Neighbors (versus 30% in the first strategy). This time, compared to the work of Buriro et al., our results have a slight decline (84% against 87.9%).

Table 4.2: Results for gender, age and user-touch experience estimation.

Datasets	Gender	Age	User-touch experience
RHU	Subject independent: 61% Random: 76.30%	Subject independent: 68% Random: 74.71%	-
KDAp	Subject independent: 86% Random: 88%	-	Subject independent: 79% Random: 81%
TDAS	Subject independent: 62% Random: 92%	Subject independent: 30% Random: 84%	-

4.4 CONCLUSIONS

TD refers to an individual's regular patterns or rhythms while typing on a touchscreen device. These behavioral patterns provide enough information to serve as a powerful biometric identifier. Compared to other biometric modalities, touch biometrics is a cost-effective, user-friendly, and continuous user authentication mechanism. TD as a soft biometric for demographic classification represents a current research topic that has not yet been fully explored. As a result, we investigated how ML classification algorithms can handle touch-interaction behavior data to classify users based on age, gender, and experience level of smartphone usage. The results obtained aim to show how lightweight continuous verification can be achieved by the analysis of soft biometrics and improve the identification mechanism through additional features such as soft biometric traits. Our study also highlights a typical bias affecting the experimentation of approaches for biometric behavioral analysis.

Nowadays, touch biometrics has unrivaled usability and huge potential for cybersecurity applications. New feature extraction and classification algorithms continue to be in high demand. On the assumption that each mobile device is used by a single user, behavioral profiling approaches will be explored. In particular, a user's identity can be verified through their applications' usage in a continuous and transparent manner by monitoring a subject's calling or location activities and using historical calling information. Undoubtedly, mobile authentication solutions based on behavioral traits do not attain the same performance as their counterparts based on physiological features, such as face or fingerprint. The restricted amount of data collected during individual acquisition sessions per user is also a crucial factor to consider. It is essential, in the context of mobile behavioral biometrics, to separate the concepts of user and device. The amount of biometric data included throughout the device's entire operating time is likely to be less than during capture sessions designed to extract the more distinctive features of mobile HCI. So, it would be interesting to find out how much of the success of identification can be attributed to the models' ability to extract and recognize features of the device instead of the user. Based on the above assumptions, to compensate for the inability to obtain a large database, future work will involve the generation of synthetic data, a technique that has proven successful in similar fields. We will further analyze the fusion of typing behavior patterns with other biometric modalities to provide a comprehensive and secure authentication solution.

5 CONCLUSIONS AND FUTURE WORKS

In this Thesis, beginning with the wide concept of behavioral biometrics and with a particular focus on the recent AI technologies in biometric pattern recognition, we focused on their potential by assessing the contexts and environments in which their usage becomes essential. We analyzed both novel techniques to successfully recognize and estimate a behavioral trait (as in the case of head pose) as well as data that presented both new opportunities and challenges (as in the case of facial expressions in the visual domain). Finally, we demonstrated that through user typing behavior on mobile devices, subjects' demographic classification can be processed feasibly and reliably. Based on our experience gained in this context, we can assert that behavioral biometric features are not intended as an alternative to classical biometrics but rather as an additional source of information. The distinction between behavioral biometrics and physical features is quite labile. The same behavioral characteristic can be used in a variety of ways to obtain different types of information. Behavioral biometrics have an exceptionally broad variety of applications, which is another significant finding of our analysis. In contrast to the recognition task, which can be applied to multiple situations without departing from the main objective, that is, the identification of the subject, behavioral biometrics is more flexible and can be applied in contexts completely unrelated to user identification. It is possible to analyze these considerations from a purely technical point of view in relation to the specific methodologies that we have presented.

In the field of human activity detection, automated pose estimation is becoming an intriguing topic of research. Head pose represents an important visual cue in numerous areas, such as human intent, motivation, and attention, among others. We proposed various implementations of the fractal encoding approach that investigate training-free strategies prior to combining the fractal parameters obtained with well-known regression models.

Partitioned Iterated Function Systems are used to represent the self-similarity properties of two images exhibiting similar head rotation, which is a substantial shift from prior research that employs CNN approaches. The training time, which depends on the size of the training set and the number of epochs, is an interesting parameter that should not be underestimated, even though CNN-based models are currently the best-performing. The time achieved by the proposed approaches highlights their suitability for real-time operation, even if compared with competing methods that use significantly more performant architectures. Nowadays, particularly in the automotive context, HPE is one of the most important factors for monitoring attention and analyzing driver behavior. A precise estimation of the driver's head pose is crucial for analyzing driver attention and behavior monitoring during driving. To achieve this objective, the placement and selection of the most suitable sensing device are critical. Specifically, the final system should be able to function in various illumination conditions, which can substantially alter the image quality and visual performance. For this reason, it will be of great interest to apply this method with different sensors, like thermal, infrared, or, even more interestingly, depth images that actually surpass traditional RGB sensors.

Facial expression analysis is a valuable source for assessing human attitude and behavior. In recent years, building a system capable of automatically identifying facial expressions from images and videos has been the subject of intensive research. Early studies of emotional expression were primarily concerned with determining whether or not perceivers could infer emotions from static depictions of prototypical facial muscle configurations thought to communicate anger, contempt, fear, sadness, and surprise. Currently, DL algorithms, particularly CNN architectures, are achieving promising results. Therefore, we developed a CNN-based approach to categorize the principal facial expressions (according to the categorical model). To enhance the performance of the proposed system, some novel data augmentation techniques have been applied to enrich the learning parameters of the proposed CNN model. Furthermore, for fine-tuning the trained CNN model, a trade-off between data augmentation and deep learning features was performed. The results achieved are highly competitive with state-of-the-art methods. We conducted a further study to analyze the correlations between facial expressions, head pose errors, and facial keypoint distances, obtaining a set of configurations that can help HPE method authors better predict and handle HPE errors related to facial expressions in an uncontrolled environment. Even though significant progress has been made as a result of the widespread use of DL-based algorithms, the majority of existing techniques that employ 2D features are incapable of resolving the challenging issues of illumination and pose variations, which could be naturally overcome by 3D techniques. Further, in the past two decades, the research community has demonstrated that facial expressions are "*multimodal, dynamic patterns of behavior*" involving facial action, vocalization, body movement, gaze, gesture, head motions, touch, etc. In light of this, in addition to exploring multiple data sources, future research should focus its attention on developing novel multimodal biometric applications and showing which fusion approaches are more suitable for emotion recognition.

Finally, with the rapid and widespread adoption of new technologies and the incorporation of more advanced sensors into mobile devices, biometric recognition can be performed in real-world applications by leveraging each user's behavioral typing patterns, based on such interactions as typing rhythm, finger-swiping speed, and device-holding posture. The advantages of using TD as a biometric trait are numerous, one of which is certainly its non-intrusiveness. Over recent years, the research efforts on the identification of so-called soft biometrics, such as age, gender, ethnicity, degree of confidence with a certain hardware, and so on, have paid off with interesting results and the definition of potential application fields. TD as a soft biometric for demographic classification represents an emerging research area that has not yet been fully explored. Despite this, our preliminary study shows how lightweight continuous verification can be achieved by the analysis of soft traits, improving the identification mechanism. We also highlight a typical bias affecting the experimentation of approaches for biometric behavioral analysis, showing how, through the two different strategies of splitting data into training and test sets, it is possible to confirm the expected result. Several aspects should unquestionably be investigated further, perhaps by utilizing new data acquired from the use of numerous built-in sensors on mobile devices or by fusing the typing behavioral patterns with other biometric modalities to provide a comprehensive and secure authentication solution. Therefore, future research will incorporate a variety of typing behavior data and the exploration of fresh data fusion techniques.

PUBLICATIONS

[1] Andrea F Abate, Paola Barra, Chiara Pero, and Maurizio Tucci. “Head pose estimation by regression algorithm.” In: *Pattern Recognition Letters* 140 (2020), pp. 179–185.

[2] Andrea F Abate, Paola Barra, Chiara Pero, and Maurizio Tucci. “Partitioned iterated function systems by regression models for head pose estimation.” In: *Machine Vision and Applications* 32.5 (2021), pp. 1–8.

[3] Andrea F Abate, Lucia Cimmino, Fabio Narducci, and Chiara Pero. “Biometric Face Recognition Based on Landmark Dynamics.” In: *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE. 2020, pp. 601–605.

[4] Paola Barra, Riccardo Distasi, Chiara Pero, Stefano Ricciardi, and Maurizio Tucci. “Gradient boosting regression for faster Partitioned Iterated Function Systems-based head pose estimation.” In: *IET Biometrics* 11.4 (2022), pp. 279–288.

[5] Silvio Barra, Sanoar Hossain, Chiara Pero, and Saiyed Umer. “A Facial Expression Recognition Approach for Social IoT Frameworks.” In: *Big Data Research* (2022), p. 100353.

[6] Carmen Bisogni, Lucia Cascone, Jean-Luc Dugelay, and Chiara Pero. “Adversarial attacks through architectures and spectra in face recognition.” In: *Pattern Recognition Letters* 147 (2021), pp. 55–62.

[7] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. “FASHE: A FrActal Based Strategy for Head Pose Estimation.” In: *IEEE Transactions on Image Processing* 30 (2021), pp. 3192–3203.

[8] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. "Hp2ifs: head pose estimation exploiting partitioned iterated function systems." In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 1725–1730.

[9] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. "PIFS Scheme for HEad Pose Estimation Aimed at Faster Face Recognition." In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.2 (2022), pp. 173–184.

[10] Carmen Bisogni and Chiara Pero. "Ifepe: On the impact of facial expression in head pose estimation." In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 486–500.

[11] Guido Bozzelli, Maurizio De Nino, Chiara Pero, and Stefano Ricciardi. "AR Based User Adaptive Compensation of Metamorphopsia." In: *Proceedings of the International Conference on Advanced Visual Interfaces*. 2020, pp. 1–5.

[12] Andrea Casanova, Lucia Cascone, Aniello Castiglione, Weizhi Meng, and Chiara Pero. "User recognition based on periocular biometrics and touch dynamics." In: *Pattern Recognition Letters* 148 (2021), pp. 114–120.

[13] Andrea Casanova, Lucia Cascone, Aniello Castiglione, Michele Nappi, and Chiara Pero. "Eye-Movement and Touch Dynamics: A Proposed Approach for Activity Recognition of a Web User." In: *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 2019, pp. 719–724.

[14] Lucia Cascone, Michele Nappi, Fabio Narducci, and Chiara Pero. "Touch keystroke dynamics for demographic classification." In: *Pattern Recognition Letters* 158 (2022), pp. 63–70.

[15] Marco Castaldo, Aniello Castiglione, Barbara Masucci, Michele Nappi, and Chiara Pero. "Energy Awareness and Secure Communication Protocols: The Era of Green Cybersecurity." In: *International Symposium on Security in*

Computing and Communication. Springer. 2019, pp. 159–173.

[16] Aniello Castiglione, Michele Nappi, Fabio Narducci, and Chiara Pero. "Fostering secure cross-layer collaborative communications by means of covert channels in MEC environments." In: *Computer Communications* 169 (2021), pp. 211–219. ISSN: 0140-3664.

[17] Aniello Castiglione, Michele Nappi, and Chiara Pero. "Towards the Design of a Covert Channel by Using Web Tracking Technologies." In: *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Springer. 2019, pp. 231–246.

[18] Debbrota Paul Chowdhury, Sambit Bakshi, Chiara Pero, Gustavo Olague, and Pankaj Kumar Sa. "Privacy Preserving Ear Recognition System Using Transfer Learning in Industry 4.0." In: *IEEE Transactions on Industrial Informatics* (2022).

[19] Lucia Cimmino, Michele Nappi, Fabio Narducci, and Chiara Pero. "M2FRED: Mobile Masked Face REcognition Through Periocular Dynamics Analysis." In: *IEEE Access* 10 (2022), pp. 94388–94402.

[20] Lucia Cimmino, Chiara Pero, Stefano Ricciardi, and Shaohua Wan. "A method for user-customized compensation of metamorphopsia through video see-through enabled head mounted display." In: *Pattern Recognition Letters* 151 (2021), pp. 252–258.

[21] Alamgir Sardar, Saiyed Umer, Chiara Pero, and Michele Nappi. "A novel cancelable FaceHashing technique based on non-invertible transformation with encryption and decryption template." In: *IEEE Access* 8 (2020), pp. 105263–105277.

[22] Saiyed Umer, Ranjeet Kumar Rout, Chiara Pero, and Michele Nappi. "Facial expression recognition with tradeoffs between data augmentation and deep learning features." In: *Journal of Ambient Intelligence and Humanized Computing* 13.2 (2022), pp. 721–735.

BIBLIOGRAPHY

[1] Andrea F Abate, Paola Barra, Carmen Bisogni, Michele Nappi, and Stefano Ricciardi. "Near real-time three axis head pose estimation without training." In: *IEEE Access* 7 (2019), pp. 64256–64265.

[2] Andrea F Abate, Paola Barra, Chiara Pero, and Maurizio Tucci. "Head pose estimation by regression algorithm." In: *Pattern Recognition Letters* 140 (2020), pp. 179–185.

[3] Andrea F Abate, Paola Barra, Chiara Pero, and Maurizio Tucci. "Partitioned iterated function systems by regression models for head pose estimation." In: *Machine Vision and Applications* 32.5 (2021), pp. 1–8.

[4] Andrea F Abate, Carmen Bisogni, Aniello Castiglione, and Michele Nappi. "Head pose estimation: An extensive survey on recent techniques and applications." In: *Pattern Recognition* 127 (2022), p. 108591.

[5] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, and Ruben Tolosana. "Multilock: Mobile active authentication based on multiple biometric and behavioral patterns." In: *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications*. 2019, pp. 53–59.

[6] Nawal Alioua, Aouatif Amine, Alexandrina Rogozan, Abdelaziz Bensrhair, and Mohammed Rziza. "Driver head pose estimation using efficient descriptor fusion." In: *EURASIP Journal on Image and Video Processing* 2016.1 (2016), pp. 1–14.

[7] Margit Antal, Zsolt Bokor, and László Zsolt Szabó. "Information revealed from scrolling interactions on mobile devices." In: *Pattern Recognition Letters* 56 (2015), pp. 7–13.

[8] Margit Antal and László Zsolt Szabó. "An evaluation of one-class and two-class classification algorithms for keystroke dynamics authentication on mobile devices."

In: *2015 20th International Conference on Control Systems and Computer Science*. IEEE. 2015, pp. 343–350.

[9] Mikel Ariz, José J Bengoechea, Arantxa Villanueva, and Rafael Cabeza. "A novel 2D/3D database with automatic face annotation for head tracking and pose estimation." In: *Computer Vision and Image Understanding* 148 (2016), pp. 201–210.

[10] Stylianos Asteriadis, Dimitris Soufleros, Kostas Karpouzis, and Stefanos Kollias. "A natural head pose and eye gaze dataset." In: *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. 2009, pp. 1–4.

[11] T. Baltrusaitis, P. Robinson, and L. Morency. "3D Constrained Local Model for rigid and non-rigid facial tracking." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2610–2617.

[12] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. "OpenFace: An open source facial behavior analysis toolkit." In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp. 1–10.

[13] Paola Barra, Silvio Barra, Carmen Bisogni, Maria De Marsico, and Michele Nappi. "Web-shaped model for head pose estimation: An approach for best exemplar selection." In: *IEEE Transactions on Image Processing* 29 (2020), pp. 5457–5468.

[14] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obregón, and Modesto Castrillón-Santana. "Gait analysis for gender classification in forensics." In: *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Springer. 2019, pp. 180–190.

[15] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obregón, and Modesto Castrillon-Santana. "Gender classification on 2D human skeleton." In: *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE. 2019, pp. 1–4.

[16] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obregón, and Modesto Castrillón-Santana. "Gotcha-i: A multiview human videos dataset." In: *International Symposium on Security in Computing and Communication*. Springer. 2019, pp. 213–224.

[17] Paola Barra, Carmen Bisogni, Michele Nappi, and Stefano Ricciardi. "Fast quadtree-based pose estimation for security applications using face biometrics." In: *International Conference on Network and System Security*. Springer. 2018, pp. 160–173.

[18] Paola Barra, Riccardo Distasi, Chiara Pero, Stefano Ricciardi, and Maurizio Tucci. "Gradient boosting regression for faster Partitioned Iterated Function Systems-based head pose estimation." In: *IET Biometrics* 11.4 (2022), pp. 279–288.

[19] Silvio Barra, Sanoar Hossain, Chiara Pero, and Saiyed Umer. "A Facial Expression Recognition Approach for Social IoT Frameworks." In: *Big Data Research* (2022), p. 100353.

[20] Nayan Kumar Subhashis Behera, Tanmay Kumar Behera, Michele Nappi, Sambit Bakshi, and Pankaj Kumar Sa. "Futuristic person re-identification over internet of biometrics things (IoBT): Technical potential versus practical reality." In: *Pattern Recognition Letters* 151 (2021), pp. 163–171.

[21] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. "Emotionet challenge: Recognition of facial expressions of emotion in the wild." In: *arXiv preprint arXiv:1703.01210* (2017).

[22] Alphonse Bertillon and Robert Wilson McClaughry. *Signaletic instructions including the theory and practice of anthropometrical identification*. Werner Company, 1896.

[23] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. "FASHE: a fractal based strategy for head pose estimation." In: *IEEE Transactions on Image Processing* 30 (2021), pp. 3192–3203.

[24] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. "Hp2ifs: head pose estimation exploiting partitioned iterated function systems." In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 1725–1730.

[25] Carmen Bisogni, Michele Nappi, Chiara Pero, and Stefano Ricciardi. "PIFS Scheme for HEad Pose Estimation Aimed at Faster Face Recognition." In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.2 (2022), pp. 173–184.

[26] Carmen Bisogni and Chiara Pero. "Ifepe: On the impact of facial expression in head pose estimation." In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 486–500.

[27] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. "Poseidon: Face-from-depth for driver pose estimation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4661–4670.

[28] Michael D Breitenstein, Daniel Kuettel, Thibaut Weise, Luc Van Gool, and Hanspeter Pfister. "Real-time face pose estimation from single range images." In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[29] Attaullah Buriro, Zahid Akhtar, Bruno Crispo, and Filippo Del Frari. "Age, Gender and Operating-Hand Estimation on Smart Mobile Devices." In: *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2016, pp. 1–5.

[30] Kenneth L Campbell. "The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety." In: *Tr News* 282 (2012).

[31] Virginio Cantoni, Lucia Cascone, Michele Nappi, and Marco Porta. "Demographic classification through pupil analysis." In: *Image and Vision Computing* 102 (2020), pp. 10–3980.

[32] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. "VGGFace2: A Dataset for Recognising Faces across Pose and Age." In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), pp. 67–74.

[33] Yuanzhouhan Cao, Olivier Canévet, and Jean-Marc Odobez. "Leveraging convolutional pose machines for fast and accurate head pose estimation." In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1089–1094.

[34] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. "A vector-based representation to enhance head pose estimation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1188–1197.

[35] Andrea Casanova, Lucia Cascone, Aniello Castiglione, Weizhi Meng, and Chiara Pero. "User recognition based on periocular biometrics and touch dynamics." In: *Pattern Recognition Letters* 148 (2021), pp. 114–120.

[36] Andrea Casanova, Lucia Cascone, Aniello Castiglione, Michele Nappi, and Chiara Pero. "Eye-Movement and Touch Dynamics: A Proposed Approach for Activity Recognition of a Web User." In: *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 2019, pp. 719–724.

[37] Lucia Cascone, Carlo Medaglia, Michele Nappi, and Fabio Narducci. "Pupil size as a soft biometrics for age and gender classification." In: *Pattern Recognition Letters* 140 (2020), pp. 238–244.

[38] Lucia Cascone, Michele Nappi, Fabio Narducci, and Chiara Pero. "Touch keystroke dynamics for demographic classification." In: *Pattern Recognition Letters* 158 (2022), pp. 63–70.

[39] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. "Deep, landmark-free fame: Face alignment, modeling, and expression esti-

mation." In: *International Journal of Computer Vision* 127.6 (2019), pp. 930–956.

[40] Jiawei Chen, Jonathan Wu, Kristi Richter, Janusz Konrad, and Prakash Ishwar. "Estimating head pose orientation using extremely low resolution images." In: *2016 IEEE Southwest symposium on image analysis and interpretation (SSIAI)*. IEEE. 2016, pp. 65–68.

[41] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. "Facial motion prior networks for facial expression recognition." In: *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE. 2019, pp. 1–4.

[42] Savina Colaco and Dong Seog Han. "Facial keypoint detection with convolutional neural networks." In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE. 2020, pp. 671–674.

[43] Meltem Demirkus, James J Clark, and Tal Arbel. "Robust semi-automatic head pose labeling for real-world face video sequences." In: *Multimedia Tools and Applications* 70.1 (2014), pp. 495–523.

[44] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. "Retinaface: Single-shot multi-level face localisation in the wild." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5203–5212.

[45] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4690–4699.

[46] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark." In: *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE. 2011, pp. 2106–2112.

[47] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. “Collecting large, richly annotated facial-expression databases from movies.” In: *IEEE multimedia* 19.03 (2012), pp. 34–41.

[48] Katerine Diaz-Chito, Jesus Martinez Del Rincon, Aura Hernández-Sabaté, and Debora Gil. “Continuous head pose estimation using manifold subspace embedding and multivariate regression.” In: *IEEE Access* 6 (2018), pp. 18325–18334.

[49] Vincent Drouard, Sileye Ba, Georgios Evangelidis, Antoine Deleforge, and Radu Horaud. “Head pose estimation via probabilistic high-dimensional regression.” In: *2015 IEEE international conference on image processing (ICIP)*. IEEE. 2015, pp. 4624–4628.

[50] Vincent Drouard, Radu Horaud, Antoine Deleforge, Sileye Ba, and Georgios Evangelidis. “Robust head-pose estimation based on partially-latent mixture of linear regressions.” In: *IEEE Transactions on Image Processing* 26.3 (2017), pp. 1428–1440.

[51] Shichuan Du, Yong Tao, and Aleix M Martinez. “Compound facial expressions of emotion.” In: *Proceedings of the national academy of sciences* 111.15 (2014), E1454–E1462.

[52] Hossein Ebrahimpour-Komleh, Vinod Chandran, and Sridha Sridharan. “Face recognition using fractal codes.” In: *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*. Vol. 3. IEEE. 2001, pp. 58–61.

[53] Paul Ekman and Wallace V Friesen. “Constants across cultures in the face and emotion.” In: *Journal of personality and social psychology* 17.2 (1971), p. 124.

[54] Paul Ekman and Wallace V Friesen. “Facial action coding system.” In: *Environmental Psychology & Nonverbal Behavior* (1978).

[55] Olufisayo Ekundayo and Serestina Viriri. “Facial expression recognition: a review of methods, performances and limitations.” In: *2019 Conference on Information Communications Technology and Society (ICTAS)*. IEEE. 2019, pp. 1–6.

[56] Mohamad El-Abed, Mostafa Dafer, and Ramzi El Khayat. "RHU Keystroke: A mobile-based benchmark for keystroke dynamics systems." In: *2014 International Carnahan Conference on Security Technology (ICCST)*. 2014, pp. 1–4.

[57] Elakkiya Ellavarason, Richard Guest, Farzin Deravi, Raul Sanchez-Riello, and Barbara Corsetti. "Touch-dynamics based behavioural biometrics on mobile devices–a review from a usability and performance perspective." In: *ACM Computing Surveys (CSUR)* 53.6 (2020), pp. 1–36.

[58] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5562–5570.

[59] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. "Random forests for real time 3d face analysis." In: *International journal of computer vision* 101.3 (2013), pp. 437–458.

[60] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. "Joint 3d face reconstruction and dense alignment with position map regression network." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 534–551.

[61] Wallace V Friesen, Paul Ekman, et al. "EMFACS-7: Emotional facial action coding system." In: *Unpublished manuscript, University of California at San Francisco* 2.36 (1983), p. 1.

[62] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras." In: *Proceedings of the symposium on eye tracking research and applications*. 2014, pp. 255–258.

[63] Francis Galton. *Finger prints*. 57490-57492. Macmillan and Company, 1892.

[64] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. "The CAS-PEAL large-scale Chinese face database and baseline evaluations." In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38.1 (2007), pp. 149–161.

[65] Darshan Gera and S Balasubramanian. "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition." In: *Pattern Recognition Letters* 145 (2021), pp. 58–66.

[66] Jeffrey M Girard, Jeffrey F Cohn, László A Jeni, Simon Lucey, and Fernando De la Torre. "How much training data for facial action unit detection?" In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, pp. 1–8.

[67] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. "Challenges in representation learning: A report on three machine learning contests." In: *International conference on neural information processing*. Springer. 2013, pp. 117–124.

[68] Nicolas Gourier, Daniela Hall, and James L Crowley. "Estimating face orientation from robust detection of salient facial structures." In: *FG Net workshop on visual observation of deictic gestures*. Vol. 6. Citeseer. 2004, p. 7.

[69] R Gross and I Matthews. "T. and Baker S. Cohn, JF and Kanade. Multi-pie." In: *Eighth IEEE International Conference on Automatic Face and Gesture Recognition*. 2008.

[70] Hatice Gunes and Björn Schuller. "Categorical and dimensional affect analysis in continuous input: Current trends and future directions." In: *Image and Vision Computing* 31.2 (2013), pp. 120–136.

[71] Aryaman Gupta, Kalpit Thakkar, Vineet Gandhi, and PJ Narayanan. "Nose, eyes and ears: Head pose estimation by locating facial keypoints." In: *ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2019, pp. 1977–1981.

[72] Erwin Haasnoot, JS Barnhoorrr, Luuk J Spreeuwers, Raymond NJ Veldhuis, and Willem B Verwey. “Towards understanding the effects of practice on behavioural biometric recognition performance.” In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 558–562.

[73] Luis Hernández-Álvarez, Lorena González-Manzano, José María de Fuentes, and Luis Hernández Encinas. “Biometrics and Artificial Intelligence: Attacks and Challenges.” In: *Breakthroughs in Digital Biometrics and Forensics*. Springer, 2022, pp. 213–240.

[74] Geoffrey E Hinton. “Deep belief networks.” In: *Scholarpedia* 4.5 (2009), p. 5947.

[75] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. “Quatnet: Quaternion-based head pose estimation with multiregression loss.” In: *IEEE Transactions on Multimedia* 21.4 (2018), pp. 1035–1046.

[76] Yunxin Huang, Fei Chen, Shaohe Lv, and Xiaodong Wang. “Facial expression recognition: A survey.” In: *Symmetry* 11.10 (2019), p. 1189.

[77] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[78] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. “Facial expressions of emotion are not culturally universal.” In: *Proceedings of the National Academy of Sciences* 109.19 (2012), pp. 7241–7244.

[79] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. “Soft biometric traits for personal recognition systems.” In: *International conference on biometric authentication*. Springer. 2004, pp. 731–738.

[80] Ankita Jain and Vivek Kanhangad. “Gender recognition in smartphones using touchscreen gestures.” In: *Pattern Recognition Letters* 125 (2019), pp. 604–611.

[81] Govind Jeevan, Geevar C Zacharias, Madhu S Nair, and Jeny Rajan. "An empirical study of the impact of masks on face recognition." In: *Pattern Recognition* 122 (2022), p. 108308.

[82] Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. "Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network." In: *Neurocomputing* 333 (2019), pp. 231–239.

[83] Mohamed Kas, Y Ruichek, R Messoussi, et al. "New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach." In: *Information Sciences* 549 (2021), pp. 200–220.

[84] Vahid Kazemi and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1867–1874.

[85] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. "A survey of the recent architectures of deep convolutional neural networks." In: *Artificial intelligence review* 53.8 (2020), pp. 5455–5516.

[86] Khalil Khan, Nasir Ahmad, Farooq Khan, and Ikram Syed. "A framework for head pose estimation and face segmentation through conditional random fields." In: *Signal, Image and Video Processing* 14.1 (2020), pp. 159–166.

[87] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization." In: *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE. 2011, pp. 2144–2151.

[88] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models." In: *IEEE Transactions on pattern analysis and machine intelligence* 22.4 (2000), pp. 322–336.

[89] Stéphane Lathuilière, Rémi Juge, Pablo Mesejo, Rafael Munoz-Salinas, and Radu Horaud. "Deep mixture of linear inverse regressions applied to head-pose estimation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4817–4825.

[90] Huadong Li and Hua Xu. "Deep reinforcement learning for robust emotional classification in facial expression recognition." In: *Knowledge-Based Systems* 204 (2020), p. 106172.

[91] Jing Li, Jiang Wang, and Farhan Ullah. "An end-to-end task-simplified and anchor-guided deep learning framework for image-based head pose estimation." In: *IEEE Access* 8 (2020), pp. 42458–42468.

[92] Shan Li and Weihong Deng. "Deep facial expression recognition: A survey." In: *IEEE transactions on affective computing* (2020).

[93] Shan Li, Weihong Deng, and JunPing Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[94] Shancang Li, Li Da Xu, and Shanshan Zhao. "The internet of things: a survey." In: *Information systems frontiers* 17.2 (2015), pp. 243–259.

[95] Yunji Liang, Sagar Samtani, Bin Guo, and Zhiwen Yu. "Behavioral biometrics for continuous authentication in the internet-of-things era: An artificial intelligence perspective." In: *IEEE Internet of Things Journal* 7.9 (2020), pp. 9128–9143.

[96] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei. "3D head pose estimation with convolutional neural network trained on synthetic images." In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 1289–1293.

[97] Xiaofeng Liu, BVK Vijaya Kumar, Ping Jia, and Jane You. "Hard negative generation for identity-disentangled facial expression recognition." In: *Pattern Recognition* 88 (2019), pp. 1–12.

[98] Yuanyuan Liu, Wei Dai, Fang Fang, Yongquan Chen, Rui Huang, Run Wang, and Bo Wan. "Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition." In: *Information Sciences* 578 (2021), pp. 195–213.

[99] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. "Clip-aware expressive feature learning for video-based facial expression recognition." In: *Information Sciences* 598 (2022), pp. 1–82–195.

[100] Yuanyuan Liu, Zhong Xie, Xi Gong, and Fang Fang. "Deep Transfer Feature Based Convolutional Neural Forests for Head Pose Estimation." In: *Pacific-Rim Symposium on Image and Video Technology*. Springer. 2017, pp. 5–16.

[101] Yuanyuan Liu, Zhong Xie, Xiaohui Yuan, Jingying Chen, and Wu Song. "Multi-level structured hybrid forest for joint head detection and pose estimation." In: *Neurocomputing* 266 (2017), pp. 206–215.

[102] Zhaoxiang Liu, Zezhou Chen, Jinqiang Bai, Shaohua Li, and Shiguo Lian. "Facial pose estimation by deep learning from label distributions." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

[103] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE. 2010, pp. 94–101.

[104] Daniel Lundqvist, Anders Flykt, and Arne Öhman. "Karolinska directed emotional faces." In: *Cognition and Emotion* (1998).

[105] Changwei Luo, Juyong Zhang, Jun Yu, Chang Wen Chen, and Shengjin Wang. "Real-time head pose estimation and face modeling from a depth image." In: *IEEE Transactions on Multimedia* 21.10 (2019), pp. 2473–2481.

[106] Iiris Lusi, Sergio Escalera, and Gholamreza Anbarjafari. "SASE: RGB-Depth Database for Human Head Pose Estimation." In: *Computer Vision – ECCV 2016 Workshops* (Nov. 2016), pp. 325–336.

[107] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. "Coding facial expressions with gabor wavelets." In: *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE. 1998, pp. 200–205.

[108] Motion Imagery Standards Board (MISB). "MISB Standard 0601." In: *UAS Datalink Local Metadata* (2014).

[109] Ahmed Mahfouz, Tarek M Mahmoud, and Ahmed Sharaf Eldin. "A survey on behavioral biometric authentication on smartphones." In: *Journal of information security and applications* 37 (2017), pp. 28–37.

[110] Emanuele Maiorana, Himanka Kalita, and Patrizio Campisi. "Mobile keystroke dynamics for biometric recognition: An overview." In: *IET Biometrics* 10.1 (2021), pp. 1–23.

[111] Farkhod Makhmudkhujaev, Mohammad Abdullah-Al-Wadud, Md Tauhid Bin Iqbal, Byungyong Ryu, and Oksam Chae. "Facial expression recognition with local prominent directional pattern." In: *Signal Processing: Image Communication* 74 (2019), pp. 1–12.

[112] Biometric Technology Market. *[Online]*. URL: https://www.precedenceresearch.com/biometric-technology-market (visited on 07/10/2022).

[113] David Matsumoto. "More evidence for the universality of a contempt expression." In: *Motivation and Emotion* 16.4 (1992), pp. 363–368.

[114] John McCarthy. "What is artificial intelligence." In: *URL: http://www-formal. stanford. edu/jmc/whatisai. html* (2004).

[115] Hayet Mekami, Abdennacer Bounoua, and Sidahmed Benabderrahmane. "Leveraging deep learning with symbolic sequences for robust head poses estimation." In: *Pattern Analysis and Applications* 23.3 (2020), pp. 1391–1406.

[116] Benouis Mohamed. "A novel technique for human face recognition using fractal code and bi-dimensional subspace." In: *International Conference on Hybrid Artificial Intelligence Systems*. Springer. 2015, pp. 87–98.

[117] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild." In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[118] Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. "Training on the job: Behavioral analysis of job interviews in hospitality." In: *Proceedings of the 18th acm international conference on multimodal interaction*. 2016, pp. 84–91.

[119] Euclides N Arcoverde Neto, Rafael M Barreto, Rafael M Duarte, Joao Paulo Magalhaes, Carlos ACM Bastos, Tsang Ing Ren, and George DC Cavalcanti. "Real-time head pose estimation for mobile devices." In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2012, pp. 467–474.

[120] Pariwat Ongsulee. "Artificial intelligence, machine learning and deep learning." In: *2017 15th international conference on ICT and knowledge engineering (ICT&KE)*. IEEE. 2017, pp. 1–6.

[121] Mehmet Akif Ozdemir, Berkay Elagoz, Aysegul Alaybeyoglu, Reza Sadighzadeh, and Aydin Akan. "Real time emotion recognition from facial expressions using CNN architecture." In: *2019 medical technologies congress (tiptekno)*. IEEE. 2019, pp. 1–4.

[122] Yanwei Pang, Yuan Yuan, Xuelong Li, and Jing Pan. "Efficient HOG human detection." In: *Signal processing* 91.4 (2011), pp. 773–781.

[123] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition." In: *British Machine Vision Conference*. 2015.

[124] Xi Peng, Junzhou Huang, Qiong Hu, Shaoting Zhang, and Dimitris N Metaxas. "Three-dimensional head pose estimation in-the-wild." In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, pp. 1–6.

[125] Hugo Proenca, Joao C Neves, Silvio Barra, Tiago Marques, and Juan C Moreno. "Joint head pose/soft label estimation for human recognition in-the-wild." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.12 (2016), pp. 2444–2456.

[126] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition." In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 121–135.

[127] Ines Rieger, Thomas Hauenstein, Sebastian Hettenkofer, and Jens-Uwe Garbas. "Towards real-time head pose estimation: Exploring parameter-reduced residual networks on in-the-wild datasets." In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer. 2019, pp. 123–134.

[128] Nataniel Ruiz, Eunji Chong, and James M Rehg. "Fine-grained head pose estimation without keypoints." In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 2074–2083.

[129] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[130] Hamid Sadeghi and Abolghasem-A Raie. "Histnet: Histogram-based convolutional neural network with chi-squared deep metric learning for facial expression recognition." In: *Information Sciences* 608 (2022), pp. 472–488.

[131] Anwar Saeed and Ayoub Al-Hamadi. "Boosted human head pose estimation using kinect camera." In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, pp. 1752–1756.

[132] Khalid Saeed. *New directions in behavioral biometrics*. CRC Press, 2016.

[133] Arthur L Samuel. "Some studies in machine learning using the game of checkers. II—Recent progress." In: *IBM Journal of research and development* 11.6 (1967), pp. 601–617.

[134] Gao-li Sang, Hu Chen, Ge Huang, and Qi-jun Zhao. "Unseen head pose prediction using dense multivariate label distribution." In: *Frontiers of Information Technology & Electronic Engineering* 17.6 (2016), pp. 516–526.

[135] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[136] Mingzhen Shao, Zhun Sun, Mete Ozay, and Takayuki Okatani. "Improving head pose estimation with a combined loss and bounding box margin adjustment." In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE. 2019, pp. 1–5.

[137] Chao Shen, Yong Zhang, Xiaohong Guan, and Roy A Maxion. "Performance analysis of touch-interaction behavior for active smartphone authentication." In: *IEEE Transactions on Information Forensics and Security* 11.3 (2015), pp. 498–513.

[138] Lu Sheng, Jianfei Cai, Tat-Jen Cham, Vladimir Pavlovic, and King Ngi Ngan. "A generative model for depth-based robust 3D facial pose tracking." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4488–4497.

[139] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning." In: *Journal of big data* 6.1 (2019), pp. 1–48.

[140] T Sim, S Baker, and M Bsat. "The cmu pose illumination and expression database teee trans. pattern analysis and machine intelligence." In: (2003).

[141] Giuseppe Stragapede, Ruben Vera-Rodriguez, Ruben Tolosana, Aythami Morales, Alejandro Acien, and Gaël Le Lan. "Mobile behavioral biometrics for passive authentication." In: *Pattern Recognition Letters* 157 (2022), pp. 35–41.

[142] Ioannis Stylios, Spyros Kokolakis, Olga Thanou, and Sotirios Chatzis. "Behavioral biometrics & continuous user authentication on mobile devices: A survey." In: *Information Fusion* 66 (2021), pp. 76–99.

[143] Jie Sun and Shengli Lu. "An Improved Single Shot Multibox for Video-Rate Head Pose Prediction." In: *IEEE Sensors Journal* 20.20 (2020), pp. 12326–12333. DOI: 10.1109/JSEN.2020.2999625.

[144] Wenyun Sun, Haitao Zhao, and Zhong Jin. "A visual attention based ROI detection method for facial expression recognition." In: *Neurocomputing* 296 (2018), pp. 12–22.

[145] Xiao Sun, Pingping Xia, Luming Zhang, and Ling Shao. "A ROI-guided deep architecture for robust facial expressions recognition." In: *Information Sciences* 522 (2020), pp. 35–48.

[146] Zhe Sun, Raymond Chiong, and Zheng-ping Hu. "Self-adaptive feature learning based on a priori knowledge for facial expression recognition." In: *Knowledge-Based Systems* 204 (2020), p. 106124.

[147] Zhe Sun, Zheng-Ping Hu, Meng Wang, and Shu-Huan Zhao. "Discriminative feature learning-based pixel difference representation for facial expression recognition." In: *IET Computer Vision* 11.8 (2017), pp. 675–682.

[148] Kalaivani Sundararajan and Damon L Woodard. "Deep learning for biometrics: A survey." In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–34.

[149] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1701–1708.

[150] Zhijie Tang, Xiaocheng Wu, Bin Fu, Weiwei Chen, and Hao Feng. "Fast face recognition based on fractal theory." In: *Applied Mathematics and Computation* 321 (2018), pp. 721–730.

[151] Pin Shen Teh, Ning Zhang, Andrew Beng Jin Teoh, and Ke Chen. "A survey on touch dynamics authentication in mobile devices." In: *Computers & Security* 59 (2016), pp. 210–235.

[152] Pin Shen Teh, Ning Zhang, Andrew Beng Jin Teoh, and Ke Chen. "TDAS: a touch dynamics based multi-factor authentication solution for mobile devices." In: *International Journal of Pervasive Computing and Communications* (2016).

[153] Issa Traore, Mohammed Alshahrani, and Mohammad S Obaidat. "State of the art and perspectives on traditional and emerging biometrics: A survey." In: *Security and Privacy* 1.6 (2018), e44.

[154] Ioannis Tsimperidis and Avi Arampatzis. "The Keyboard Knows About You: Revealing User Characteristics via Keystroke Dynamics." In: *International Journal of Technoethics (IJT)* 11.2 (2020), pp. 34–51.

[155] Cigdem Turan and Kin-Man Lam. "Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study." In: *Journal of visual communication and image representation* 55 (2018), pp. 331–341.

[156] Saiyed Umer, Ranjeet Kumar Rout, Chiara Pero, and Michele Nappi. "Facial expression recognition with tradeoffs between data augmentation and deep learning features." In: *Journal of Ambient Intelligence and Humanized Computing* 13.2 (2022), pp. 721–735.

[157] JA Unar, Woo Chaw Seng, and Almas Abbasi. "A review of biometric technology along with trends and prospects." In: *Pattern recognition* 47.8 (2014), pp. 2673–2688.

[158] Roberto Valle, José M Buenaposada, and Luis Baumela. "Multi-task head pose estimation in-the-wild." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.8 (2020), pp. 2874–2881.

[159] Michel Valstar, Maja Pantic, et al. "Induced disgust, happiness and surprise: an addition to the mmi facial expression database." In: *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Paris, France. 2010, p. 65.

[160] Paul Viola and Michael J Jones. "Robust real-time face detection." In: *International journal of computer vision* 57.2 (2004), pp. 137–154.

[161] Minh Thanh Vo, Trang Nguyen, and Tuong Le. "Robust head pose estimation using extreme gradient boosting machine on stacked autoencoders neural network." In: *IEEE Access* 8 (2019), pp. 3687–3694.

[162] Weiwei Wang, Xiaoyan Chen, Shuangwu Zheng, and Haiqing Li. "Fast head pose estimation via rotation-adaptive facial landmark detection for video edge computation." In: *IEEE Access* 8 (2020), pp. 45023–45032.

[163] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. "Toward practical smile detection." In: *IEEE transactions on pattern analysis and machine intelligence* 31.11 (2009), pp. 2106–2111.

[164] Lior Wolf, Tal Hassner, and Itay Maoz. "Face recognition in unconstrained videos with matched background similarity." In: *CVPR 2011*. IEEE. 2011, pp. 529–534.

[165] Shen Yuong Wong, Keem Siah Yap, Qingwei Zhai, and Xiaochao Li. "Realization of a hybrid locally connected extreme learning machine with DeepID for face verification." In: *IEEE Access* 7 (2019), pp. 70447–70460.

[166] Yue Wu and Qiang Ji. "Facial landmark detection: A literature survey." In: *International Journal of Computer Vision* 127.2 (2019), pp. 115–142.

[167] Jiahao Xia, Libo Cao, Guanjun Zhang, and Jiacai Liao. "Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks." In: *Ieee Access* 7 (2019), pp. 48470–48483.

[168] Weicheng Xie, Xi Jia, Linlin Shen, and Meng Yang. "Sparse deep feature learning for facial expression recognition." In: *Pattern Recognition* 96 (2019), p. 106966.

[169] Luhui Xu, Jingying Chen, and Yanling Gan. "Head pose estimation using improved label distribution learning with fewer annotations." In: *Multimedia Tools and Applications* 78.14 (2019), pp. 19141–19162.

[170] Mingliang Xue, Xiaodong Duan, Wanquan Liu, and Yan Ren. "A semantic facial expression intensity descriptor based on information granules." In: *Information Sciences* 528 (2020), pp. 113–132.

[171] Jucheng Yang, Yarui Chen, Chuanlei Zhang, Dong Sun Park, and Sook Yoon. "Introductory Chapter: Machine Learning and Biometrics." In: *Machine Learning and Biometrics*. IntechOpen, 2018.

[172] Lie Yang, Yong Tian, Yonghao Song, Nachuan Yang, Ke Ma, and Longhan Xie. "A novel feature separation model exchange-GAN for facial expression recognition." In: *Knowledge-Based Systems* 204 (2020), p. 106217.

[173] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1087–1096.

[174] Ahmad Zairi Zaidi, Chun Yong Chong, Zhe Jin, Rajendran Parthiban, and Ali Safaa Sadiq. "Touch-based continuous mobile device authentication: State-of-the-art, challenges and opportunities." In: *Journal of Network and Computer Applications* 191 (2021), p. 103162.

[175] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. "FDN: feature decoupling network for head pose estimation." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 12789–12796.

[176] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. "From facial expression recognition to interpersonal relation prediction." In: *International Journal of Computer Vision* 126.5 (2018), pp. 550–569.

[177] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. "From facial expression recognition to interpersonal relation prediction." In: *International Journal of Computer Vision* 126.5 (2018), pp. 550–569.

[178] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. "Facial expression recognition from near-infrared videos." In: *Image and vision computing* 29.9 (2011), pp. 607–619.

[179] Rui Zhao, Tianshan Liu, Jun Xiao, Daniel P.K. Lun, and Kin-Man Lam. "Deep Multi-task Learning for Facial Expression Recognition and Synthesis Based on Selective Feature Sharing." In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 4412–4419.

[180] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. "A comprehensive survey on automatic facial action unit analysis." In: *The Visual Computer* 36.5 (2020), pp. 1067–1093.

[181] Yu Zhong and Yunbin Deng. "A survey on keystroke dynamics biometrics: approaches, advances, and evaluations." In: *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics* 1 (2015), pp. 1–22.

[182] Xiangxin Zhu and Deva Ramanan. "Face detection, pose estimation, and landmark localization in the wild." In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2879–2886.

[183] Xiangxin Zhu and Deva Ramanan. "Face detection, pose estimation, and landmark localization in the wild." In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2879–2886.

[184] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. "Face alignment across large poses: A 3d solution." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 146–155.

[185] Visa survey: consumers ready to switch from passwords to biometrics. *Goodbye, passwords. Hello, biometrics. [Online]*. URL: https://usa.visa.com/visa-everywhere/security/how-fingerprint-authentication-works.html (visited on 07/10/2022).