Facial Biometrics on Mobile Devices: Interaction and Quality Assessment

Chiara Lunerti

A Thesis submitted to the University of Kent for the Degree of Doctor of Philosophy in the subject of Electronic Engineering

May 2019

Abstract

Biometric face recognition is a quick and convenient security method that allows unlocking a smartphone device without the need to remember a PIN code or a password. However, the unconstrained mobile environment brings considerable challenges in facial verification performance. Not only the verification but also the enrolment on the mobile device takes place in unpredictable surroundings. In particular, facial verification involves the enrolment of unsupervised users across a range of environmental conditions, light exposure, and additional variations in terms of user's poses and image background.

Is there a way to estimate the variations that a mobile scenario introduces over the facial verification performance?

A quality assessment can help in enhancing the biometric performance, but in the context of mobile devices, most of the standardised requirements and methodology presented are based on passport scenarios. A comprehensive analysis should be performed to assess the biometric performance in terms of image quality and user interaction in the particular context of mobile devices.

This work aimed to contribute to improving the performance and the adaptability of facial verification systems implemented on smartphones. Fifty-three participants were asked to provide facial images suitable for face verification across several locations and scenarios. A minimum of 150 images per user was collected with a smartphone camera within three different sessions. Sensing data was recorded to assess user interaction during the biometric presentation. Images were also recorded using a Single Lens Reflex camera to enable a comparison with conditions similar to a passport scenario.

Results showed the relationship within five selected quality metrics commonly used for quality assessment and the variables introduced by the environment, the user and the camera. Innovative methodologies were also proposed to assess the user interaction using sensors implemented in the smartphone. The analysis underlined important issues and formulated useful observations to enhance facial verification performance on smartphone devices.

Acknowledgments

I wish to express my sincere gratitude to those who have supported me throughout the whole duration of my PhD studies.

First of all, I would like to express my sincere gratitude to my supervisor, Dr Richard Guest, for his guidance, motivation and academic support. I am forever thankful for his professionalism and for all the opportunities that I had the chance to take during these years. I would like to thank my colleagues at the University of Kent for useful and insightful comments and encouragements that help me overcome the difficulties encountered.

A heartfelt thanks also to Raul, Ramon, Judith and the whole GUTI team that hosted me during my Erasmus at the University Carlos III of Madrid. They welcomed me like a family member, and I treasure every moment of that experience. I would also like to thank Patrizio Campisi to make me feel at home and welcome me in his research group during the writing of the thesis.

A special thanks to all my friends and colleagues that found the time to help me with my data collection and offered their valuable biometric data.

Finally, I want to thank my family, mum, dad and my brother, my boyfriend, and all my friends all over the World for always support me and believe in me, for reminding me that I'm never alone, despite the distance.

Table of Contents

List of Tables

List of Figures

Introduction

1.1 Biometrics on mobile devices

Mobile devices have brought a considerable change in everyday life. Smartphones, tablets, and laptops can be used to access sensitive data such as contacts, emails, and calendars at any time. They are ubiquitous both for business and personal tasks, from saving images to a photo gallery to interacting with financial information. As such, sensitive data has the risk of being accessed by unauthorised users. What makes these devices so essential is their mobility, but it also makes them easy to get lost or stolen. It is therefore of critical importance to prevent and improve the security of mobile devices through appropriate and effective authorisation processes.

Recently, biometrics have been increasingly used ahead of PIN and password for protecting access to mobile devices. Biometric systems prevent users from having to remember passwords and also provide safety against attacks such as shoulder-surfing [1].

The adoption of biometrics on mobile devices is promoted through several aspects [2]:

- Firstly, the consideration that a capture device/sensor for several biometric modalities is already included on the mobile device – e.g. every device includes a microphone, that can be used for voice recognition. Likewise, it is the ubiquity of a camera and a touchscreen that can be used for face and signature/writing authentication. Recently, mobile devices incorporated a specific fingerprint sensor allowing the use of fingerprint verification.
- Users already in possession of the devices would only need to acquire an application program implying a reduction in the cost of deployment.
- Furthermore, the adoption of biometrics on mobile devices might be of help for the acceptability issue that had always been present when developing a biometric system, as people are familiar with their device and are more likely to adopt the use of biometrics to unlock it.

There are, however, several challenges that the implementation of biometrics on mobile devices must address. For instance, the available sensors can vary in number and location depending on the device model. Also, different operating systems (OS) and devices embed different biometric methods which may influence the opinion that the population have towards different security methods. Knowing the end-users' opinion is crucial to understand their choices in the adoption of a specific security modality on their mobile devices.

1.2 Research motivations

In implementing practical, usable and appropriate security systems, it is crucial to understand the users' insights on security technologies to provide the right level of protection [3]. Which is why, as the first step in our research, we carried out an investigation on the users' opinion related to different techniques when applied in specific real-life scenarios and seeks to assess whether the awareness of storing sensitive information influences this decision.

We collected and analysed public perceptions of authentication methods on mobile devices across over 400 participants who took part in an online survey, providing their opinions on the systems they experience and use on their mobile devices, and their willingness to use different modalities in real-life scenarios. The results indicated a range of considerations including that biometrics is gaining more acceptance as a solution for security on mobile devices and that the awareness of storing sensitive data on a device influences the approach to security method adoption.

Among the survey outcomes, it outstood that facial recognition is still a modality that needs to be worked on: despite being widely known and already used in several applications, users are still reluctant of using it, and the performance of this specific modality is influenced by several aspects when used in everyday life. For these reasons, we decided to focus on facial verification and in particular when implemented on a smartphone device. The adoption of face recognition on mobile devices has many advantages. As well as ease of use, it can be easily implemented on smartphones as it only requires the use of the already embedded frontal camera.

However, there are also many challenges that need to be taken into consideration when implementing face recognition on smartphones. For instance, the frontal camera usually has less resolution compared to the rear-facing one, and this can limit the quality of the facial images.

The smartphone's mobility implicates that the authentication can happen under a considerable variability of conditions. There is no control in the way the user will interact with the device, neither where the interaction will happen, making the surrounding environment an ulterior variable to take into consideration. The environment where the authentication takes place is impossible to predict, as light exposure depends on the user's position and the time of the day. Also, the facial image's background will not be uniform, as there can be many elements of "noise" behind the users, including other people's faces.

Another aspect that influences the quality and performance of mobile authentication is the user's acceptability and their interaction with the technology. Since biometric authentication requires the presentation of a person's characteristic, the user itself is an active part of the authentication process. To ensure good quality samples for facial recognition, users should feel comfortable during the biometric presentation, and it

should be easy for them to understand how to present the biometrics to the sensor. Therefore, implementing a biometric system on a mobile device implies testing not only the performance of the system but also the interaction that the user has with the sensor.

It is difficult to analyse these aspects in a lab-based experiment because it is hard or impossible to recreate realistic variability of real-life scenarios. With this research, we aim to assess the influence that the environment and the user's interaction have on the face recognition's performance when used on smartphones.

Our study aims to analyse to what extent the variability of light exposure and background in facial images influence the quality metrics and the biometric matching scores to assess the performance of the system in two different conditions that include indoor and outdoor locations. Furthermore, we analysed the level of ease of use and comfort that the user felt in taking the images under these two conditions.

1.3 Thesis structure

This thesis is composed of eight Chapters in total. The state-of-the-art is presented in Chapter 2, followed by the description and results obtained from the online survey that we conducted to understand the public perceptions of mobile biometrics, that is presented in Chapter 3. Chapter 4 illustrates the procedures followed for the collection of the data necessary for this study. It also provides a description of each data type considered in the analysis and how it has been pre-processed and selected.

The following three chapters present the results and the contributions of this thesis. Chapter 5 describes the face detection assessment and the number of Failures to Detect (FTD) that occurred using three different algorithms in the database. It also indicates the analysis assessment of the different factors that have influenced the detection of the facial areas in the images and to what extent.

The next chapter, Chapter 6, presents the image quality assessment, as well as the statistical analysis of the influencing elements that have been considered when measuring image quality. Finally, Chapter 7 describes the performance of the verification systems across the different scenarios considered to assess user interaction and environmental surroundings.

Chapter 8 concludes this work and summaries the results obtained and the observation made from this study. It also provides some considerations for future work.

State of the art

1.1 Face recognition on mobile devices

Face recognition provides a quick, easy to use and reliable modality to authenticate on mobile devices. However, the use of this biometric modality in the mobile context brings relevant issues that need to be addressed. Current research is working on enhancing the performance and the acceptability on facial verification by considering the critical related challenges that the use of facial biometric systems brings when implemented on smartphone devices.

The main effort in research has been focusing on enhancing the performance of the facial verification system by considering live detection and anti-spoofing techniques [4]. Other mentioned acceptability issues relate to privacy and concerns on whether the biometric data is stored and secured on the owner device and not available to use by third-party. Face recognition also brings usability issues when considering that the system should enable access to the device in any environmental condition, even in darkness.

The "Quality Labelled Faces in the Wild" (QLFW) database [5] was released to investigate the effect of unconstrained environment conditions over facial verification images. The database presents 13,233 images of 5,749 subjects taken in different light exposure and user's pose conditions, including variability in focus, demographics, and camera resolution.

Although research in images taken "in the wild" is advancing [6], obtaining enhanced verification performance in an unconstrained environment, in the context of mobile devices, it is crucial to asses realistic scenarios that involve the collection of images by the users with smartphone cameras. While taking the images, not only the subject moves to present different head pose and posture, but also the camera's movements introduce variations that can affect the image quality. In face verification, the majority of implementation standards and best practices focus on specific scenarios, such as electronic IDs or passports. Best practice needs to be adapted to the unconstrained environmental variables introduced by the mobility of the device.

2.1 Facial Image Quality

Facial Image Quality (FIQ) assessment can be used to estimate and enhance the biometric system performance by identifying and rejecting those images that are not conformed with the requirement before the authentication.

The ISO/IEC 29794-5 Biometric Sample Quality Technical Report (TR) [7] provides methodologies and guidelines to assess the image quality of facial images for biometric

authentication. Several factors can affect the facial image quality, including the subject characteristics and the acquisition process, that includes the environmental conditions in which the presentation is taking place. The TR suggests a distinction between static and dynamic characteristics of both the subject and the acquisition process. In Table 2.1 there is an example of the characteristics considered in FIQ assessment distinguished between the static and dynamic.

Table 2.1: ISO/IEC 29794 TR proposed characterisation of Facial Quality.

	Subject characteristics	Acquisition Device
Static	Morphological characteristics: - Anatomical characteristics (e.g. eyes position, head dimensions) - Ethnicity - Injuries and scars Not permanent characteristics - heavy makeup - glasses - permanent jewellery	- Static proprieties of background - Physical proprieties (resolution and contrast) - Camera characteristic (sensor resolution)
Dynamic	- Head-pose - Opened\closed eyes - Subject Posing	- Dynamic characteristics of backgrounds - Variation in lighting - Position of the subject in the image - Partial occlusion

The ISO/IEC 29794 TR reports a series of metrics and indications to assess the image appearance: quality scores can be calculated to estimate the illumination strength based on the distribution of pixels values over the image histogram. Image quality metrics included in the TR are:

- Image Brightness
- Image Contrast
- Perceive Contrast considering the spatial frequency
- Exposure
- Focus, Blur and Sharpness

The metrics are described and provide different methods to calculate the scores. The TR mainly refers to the static characteristic of the users for 2D portrait images that are also specified in the International Standard ISO/IEC 19794-5:2011 Biometric data interchange formats – Part 5: Face image data [8]. The Standard specifies the face image format for face recognition, including recommendations and best practice for the collection of facial images. Aspects that should be included and considered in facial recognition are the digital image attributes and photography properties, such as image

resolution and camera positioning, but also the scene constraints like lighting or the user pose.

Following the requirements specified in the Standard result in enhancement on the verification accuracy. In particular, many of the recommendations described in the Standard focus on the specific scenario of electronic ID or Passport images. The illumination and the pose variation are two critical aspects that have been assessed for image quality applications in the gate access scenario, as they cause severe lowering in the performance of the recognition system [9].

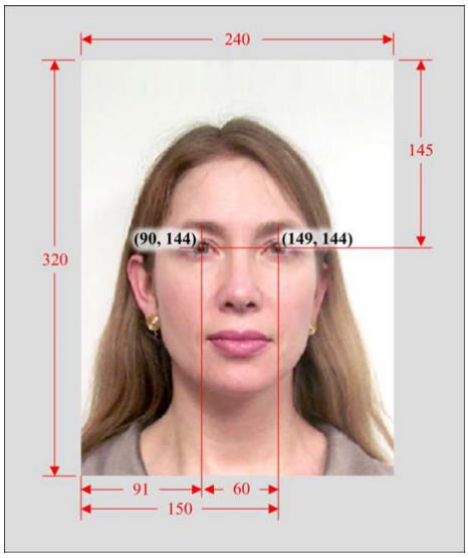


Figure 2.1: Geometric requirements for compliance with the ISO/IEC 19794-5 [8].

The application of the ISO/IEC 19794-5 Standard to improve the accuracy of the verification system is challenging to apply when the images are taken with non-cooperative users in an unconstrained environment. In an uncontrolled environment, the challenge given by the variations introduced by the lighting conditions and the user's movements and pose.

The National Institute of Standards and Technology (NIST) reported that in the past few years there had been an advance in the recognition technology for facial images, in mainly thanks to the use of convolutional neural networks (CNN) [10]. The importance of having facial image quality and the effect on the performance of the system was also considered with the Face Recognition Vendor Test (FRVT) performed by NIST: the problems of assessing quality with a unique way is still ongoing. NIST is running an assessment of the algorithms [11].

There are several studies undertaken concerning the assessment of image quality for face recognition and environmental factors, notably different light exposure and pose of the user, but only a few were focused on mobile devices. One of the main approaches in dealing with "poor" quality images is to enhance the performance of the verification

system by rejecting those images that presented an FIQ score that is lower than a selected threshold that defines a “good” quality image.

The author in [12] presented, in 2007, two algorithms for Quality Assessment (QA) concerning the blurriness of the image, the user’s head pose and facial expressions and the lighting conditions. The first introduced approach defined measurements to assess the level of degradation of the facial images, while the second approach classifies the intensity of facial expressions within “Good quality” or “Poor quality”. The methods were assessed through a polynomial function that predicted the face recognition performance using the Eigenface technique, that involves the extraction of significant features from a facial image that enables the verification comparisons [13]. The algorithms were assessed using the Face Recognition Technology (FERET) database [14]. The blur in the images was artificially added using a Gaussian filter. The QA results showed that the algorithms were able to estimate the image quality and applying an acceptance threshold for each quality metric, it could be possible to classify and select “Good quality” images to obtain higher performance.

In work presented in [15] also in 2007, the authors proposed an approach for standardisation that enables to assess the differences in facial symmetry due to non-frontal lighting and user’s pose during the facial image presentation. The method proposed by the authors was tested using a dataset of 10 subjects that took facial images under 65 different light conditions and nine different poses. The images were selected from the Yale Face Database B [16] that enables the assessment of facial images under different light and pose condition.

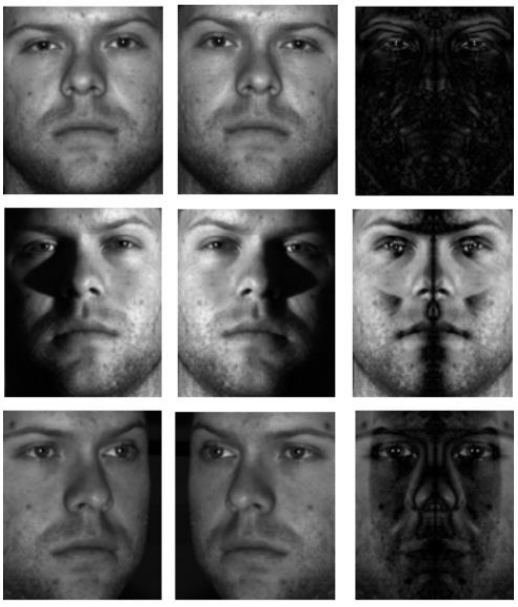


Figure 2.2: Original images (Left 1-3 column), the mirror version of the image (middle 1-3 column) and corresponding symmetrical difference images (Right 1-3 column) [15].

The proposed method divided the facial region symmetrically into the left and right side and assessed the symmetry in light and pose by checking the locally-filtered pixel

values and their mirrored corresponding location. If the image is left-right symmetric, the difference between the selected image feature and the filtered pixel locations would be zero; else the values will represent the asymmetric difference. An example is shown in Figure 2.2. The presented method for QA resulted effective to assess the lighting and user pose symmetry and was included in the ISO/IEC 29794-5 Technical Report.

A different approach to image quality is to identify fiducial face points in the facial image that are resilient to the different light and user's pose conditions, as the method proposed by the authors in [17]. The methodology uses Toeplitz matrices to identify 25 landmark points and test them over a database of 30 users verifying images taken in an unconstrained environment. The algorithm achieved 90% success rate showing resilience to the variations added in light and user's pose, although these results worsen when increasing the size of the database, indicating the need of future work to be able to use it over a larger scale of subjects.

In 2014, the work in [18], presented by Abaza et al., presents an evaluation of common metrics used for QA and introduced an alternative FIQ measure to predict the matching performance by requesting another sample in the case where a donated image did not conform to quality requirements. The method was assessed using open source experimental databases that involved images collected under different light and pose conditions. The authors artificially added the variation in quality to analyse the different variation in intensity for the common quality metrics considered, that included: Contrast, Brightness, Focus and Sharpness and Illumination. Results presented an enhancement in the system performance when rejecting the images that were classified as "low-quality", obtaining an improvement from 60.67% to 69.00% of correct biometric verification when using a distribution-based algorithm (Local Binary Patterns) and from 92.33% to 94.67% when using commercial software (PittPatt).

The author in [19] presented an evaluation of FIQ metrics considering facial images taken with a smartphone device. Facial angles, illumination conditions and distance from the camera device were assessed over a database of 101 subjects that collected 22 facial images using two different mobile devices: a Samsung Galaxy S7 and an iPhone 6 Plus. The images were collected by 48 subjects on a second experimental session. The study over the light and pose variations was performed by asking the participants to take images within fixed, established positions. Two images were taken with different yaw angles (head turned to the right or the left), and six more by varying the user pose by the roll (head tilt to the right or the left) and pitch (head leaning to the front or the back) angles. Authors evaluate the quality metrics specified in the ISO/IEC 29794-5 TR considering the traditionally employed framework presented in Figure 2.3. FIQ metrics considered were Lighting and Pose Symmetry, Brightness, Contrast, Global Contrast Factor (GCF), Exposure, Blur and Sharpness. Furthermore, they proposed a new quality metric as an overall score for the input image to specifically address smartphone images for facial verification. Results demonstrate that the metrics resulted in nearly equal or better performance to the other quality assessment methodologies in the collected database.

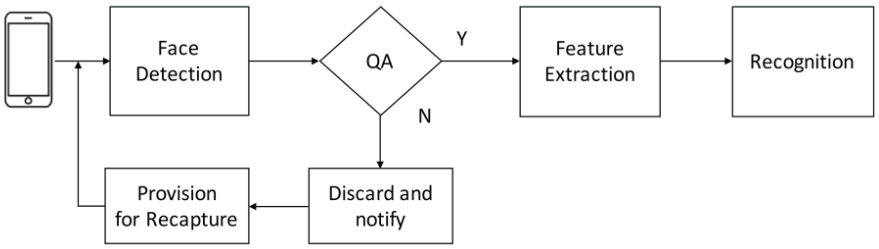


Figure 2.3: The quality assessment framework for facial image recognition [19].

To enhance the biometric performance of facial verification implemented on mobile devices, the authors in [20] proposed a generic FIQ metric that considered the differences between the enrolment and the sample images. The FIQ metrics considered for assessing quality were Brightness, Contrast, Focus and Illumination. By assessing these quality metrics, the quality measure proposed considers the facial image to have a "good" quality when the condition between the verification image and the enrolment are similar according to the FIQ values. The analysis was performed considering 1,050 images collected by ten subjects with a smartphone camera. The values calculated from each FIQ metric considered were combined using three methods: mean, geometric mean and weighted mean. The results showed that the proposed generic quality metric reported higher correlation coefficient values with the biometric performance.

The related work presented over FIQ assessment indicates that there is not a unique method to investigate the quality of an image. Only a few studies focused on smartphone authentication images, where there are factors that can influence the biometric performance, such as the resolution of the front-mounted camera used for the image capture. There is a lack of study that assesses how the FIQ metrics varies within facial images collected on a mobile device. Also, there is a need to assess image quality over realistic mobile scenarios. Most of the database used for the QA are artificially modified to estimate the intensity of the noise in the image. Moreover, the variation in user's pose and lighting is controlled presenting images taken with head position or light exposure.

2.2 User's interaction on mobile biometrics

Face recognition on mobile device involves a self-assessed unsupervised verification of a biometric characteristic. The user is a critical integral part of the biometric system, and as such, the interaction with the smartphone during the mobile authentication is one of the main aspects that should be considered to ensure high verification performance. However, there are only a few studies that assess user interaction on face recognition. Even less when considering mobile devices. Automated face recognition presented several problems in being accepted, mostly due to usability issues.

The NIST Visualization and Usability Group [21] has been working on usability in biometrics since 2005. Most of the NIST publications are based on the definition of usability from the ISO 9241-11:2018 Ergonomic requirements for office work with visual

display terminals (VDTs) – Part 11: Guidance on usability [22], where efficiency, effectiveness and satisfaction are the primary metrics. Publications by NIST are comprehensive and cover topics such as ergonomics, user acceptance or accessibility.

One of the main NIST's contributions is a handbook regarding usability and biometrics released in 2008: "*Usability and Biometrics: Ensuring Successful Biometric Systems*"[3]. This handbook helps to determinate the impact that the user's interaction has on the system performance and introduces the user-centred design in biometrics. It also defines some guidelines that can help designers and developers of biometric devices. The design places the user in the centre, as all the qualities and demographic characteristics that users bring should be considered to enhance the performance of biometric systems together with the instruction and feedback that they receive. The NIST user-centred design process described in the handbook consists of first, an analysis of the context of use, secondly a definition of the user and organisational requirements, then the designed solution that meets the requirements, and finally an evaluation of the design (Figure 2.4).

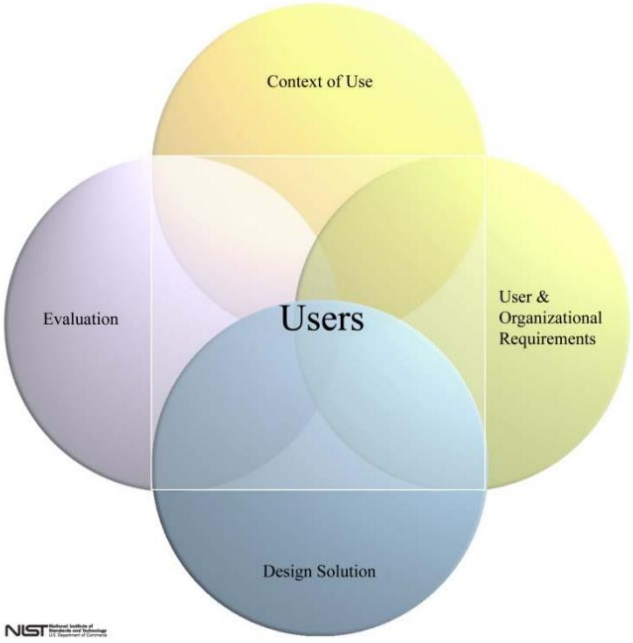


Figure 2.4: NIST Usability Model [3].

As the first step to designing a usable system, the NIST model indicates the analysis of the context of use to define the users' needs and expectations and how their demographics and abilities can affect the biometric system. In this phase, it is also necessary to understand the environment in which the users will use the system and the goals they try to achieve. Once the context of use has been determined, the following step will be defining which requirements to consider because this decision will have a significant impact on the user experience of the system. After that, it will be possible to develop the design solution, and the last essential step will be the evaluation of the system that helps to identify issues that need to be resolved. According to NIST, the best

approach to evaluate the system is to combine both qualitative and quantitative evaluations. The evaluation of the system is a critical component of any system design process. To help designers in this critical evaluation, the last chapter of the handbook gives precise guidelines and information on different usability methods and techniques.

Only a few recent studies on user interaction have moved to mobile biometric scenarios. In works made by UC3M [23] [24] users were required to interact with biometrics embedded in mobile devices within the most common scenarios (an example is in Figure 2.5).

Figure 2.5. Users were interacting with biometrics in different scenarios during one of the UC3M ergonomics experiments [24].

Conti et al. [25] (2014) analysed the usability of a fingerprint reader linked to an Android-based mobile device. Although they acknowledged the importance of video-recording sessions (and even using eye-tracking systems or think loud approach), the authors argued that those methods could impact the user's experience. Instead, in order to collect feedback from users, it was prepared a grid analysis of critical situations accompanied by a short final questionnaire. Furthermore, the Android application included ways to track user's interaction such as time spent on performing the task. Authors conclude that some real-time operations concerning human-machine interactions can slow down due to the biometric authentication process and this could make the users feel annoyed.

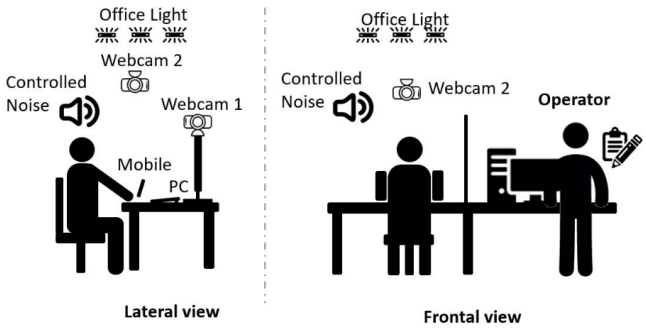


Figure 2.6: Camera setting scenarios utilised in [26].

Video-recording the users while interacting with the mobile device is the most used approach when assessing usability. A study presented in [26] investigate the usability of face and voice modality on a smartphone device. The proposed settings involve the disposition of webcams to enable the recording of user interaction (Figure 2.6).

The challenge is to collect the data for usability analysis without making the subjects feeling uncomfortable or closely under observation. It would be useful to use the

embedded sensors already implemented on the device to assess the usability. A possible solution could be using sensing data collected from smartphone sensors like the gyroscope or the accelerometer since this type of data is already used for detecting user activity, especially for continuous authentication [27].

2.3 Objective and research questions

Some open issues and challenges need to be addressed to enhance the performance and the acceptability of the facial recognition system on mobile devices. Despite the improvements in designing efficient algorithms that perform with images taken “in the wild”, there are still many issues that need to be addressed.

The assessment on image quality can provide an estimation over the variation given by the environmental conditions and the user interaction. In the particular context of mobile authentication, the users will require to access the device at any time, so there is a considerable variation in light exposure, image background and the surrounding environment. Furthermore, the users can move freely during image acquisition, and the capture device is moving with them since the camera is implemented on the smartphone. Subjects can present a large variability in head pose and facial expression.

Research has focused on images taken in an unconstrained environment, but the attention was more addressed on video surveillance applications. There is not a unique methodology or metric to measure image quality, although algorithms are being tested and proposed. One of the main objective in our research is to identify FIQ metrics that are commonly used in state of the art to assess image quality. The study investigates how FIQ values variate in the unconstrained context of mobile biometrics and to obtain adapted context requirements when using facial verification on mobile devices.

Moreover, there is a lack of studies that involve real-life scenarios. Often the variation in terms of user pose and light condition have a fixed position, or the noise elements are artificially added to the images. For this reason, we decided to collect a database that could simulate locations and environmental conditions that can occur in real life.

Few studies consider the usability of biometrics, and they usually involve video recording the users during the interaction. Videos are difficult to examine and usually require an extended amount of time since they often involve a visual examination. It would be useful to assess user interaction using only the data obtained from the device. Moved by these reasons, we consider the inclusion in our experimental design of sensing data provided by the accelerometer and the gyroscope to assess the user interaction.

Public Perceptions of Biometric User Authentication on Mobile Devices

3.1 Introduction

During the past few years, the numerous security systems adopted to protect smartphone access had been changing, updating, and enhancing to respond to the different users' needs and preferences. Personal Identification Numbers (PINs) and passwords are two modalities that have been traditionally used to protect access across a range of device manufacturers and Operating Systems (OSs). In 2008 the Android OS also introduced a personalised graphical pattern that allows the unlocking of the device by connecting at least four dots on a 3x3 grid (Figure 3.1). However, all these security methods are vulnerable to attacks such as shoulder-surfing or are easy to replicate or guess [28], [29]. Shoulder-surfing is the terminology used when impostors secretly observe users typing their password to replicate it when accessing the device.

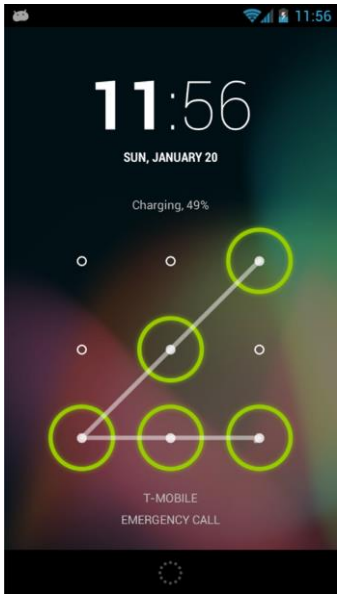


Figure 3.1: An example of the Unlock Pattern from Android [30].

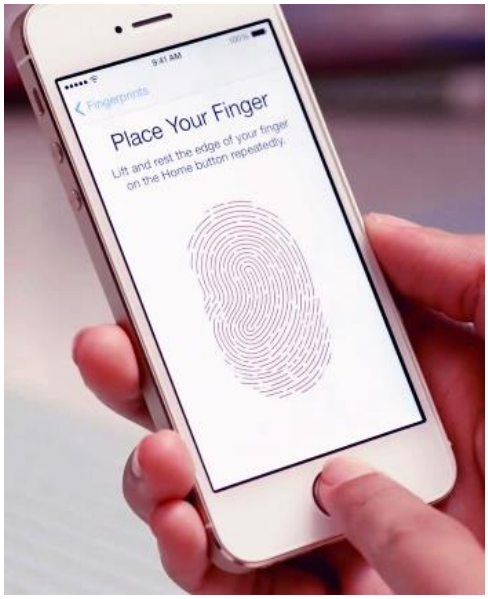


Figure 3.2: Touch ID on iPhone [31].

In 2011, Google introduced into Android 4.0 "Ice Cream Sandwich" a face recognition system called Face Unlock. This security method allows unlocking the device using the front-mounted device camera. In recent years the system has been updated and improved until it was replaced in 2014 with Smart Lock [32] in the Android 5.0 "Lollipop".

Smart Lock is a group of security options that includes, along with the more traditional PIN and unlock pattern, a face verification system, called Trusted Face, that unlocks the screen of the device when the owner's face is detected by the front camera, and a voice recognition system, called Trusted Voice [33].

In 2015, Android also introduced fingerprint recognition from the OS Android 6.0 "Marshmallow" used both for unlocking the screen of the device and for allowing users to authorise online payments and get access to specific apps. In additions to these biometric modalities, in April 2017, Samsung released its first Android device with iris recognition [34]. More recently, in 2019, the new smartphone released by Samsung, the Galaxy S9 [35], implemented an Intelligent Scan face recognition, that combines face and iris recognition for a more reliable and secure solution to protect access to the device.

On Apple devices, the Touch ID fingerprint recognition was released in 2013 [36], available on the iPhone 5S and later, iPad Air 2 and later, iPad Pro, and the iPad Mini 3 and later (Figure 3.2). This system can be used to unlock the device, to make purchases in the various Apple digital media stores (iTunes Store, the App Store, iBooks Store) and to authenticate Apple Pay in stores and within apps (using an iPhone 6 or later). At the end of 2017, Apple Inc. released the iPhone X with a series of sensors (Figure 3.3) that enabled a 3D scan of the user's face [37]. The smartphone is implemented with an infrared flood light that can allow the detection of the face regardless of the illuminance conditions.

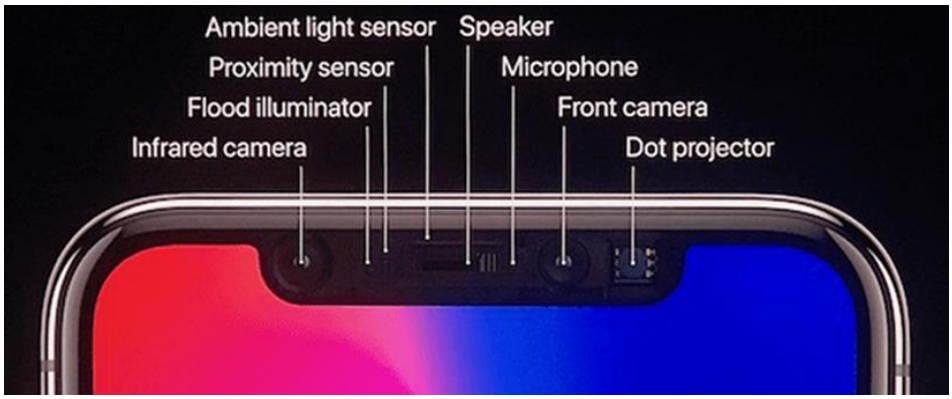


Figure 3.3: The 3D scanning technology implemented by Apple Inc. in iPhone X [38].

With the OS Windows 10, Microsoft introduced Windows Hello that supports face, fingerprint and iris recognition. From the end of 2016, it has been possible to use fingerprint and iris recognition not only on laptop devices but also on Windows phones. The capture of a good iris image requires some additional technology implemented: the device uses infrared light to illuminate the eyes of the user and a specific camera that works in different light conditions that take a picture of the iris [39].

Despite the numerous options that smartphones companies provide, there are also users that decide not to protect the access to their own devices, due to the frequency of unlocking the smartphone's screen and the time required for the authentication [40]. The

compromise between efficiency, satisfaction and the right level of security has been the object of many recent studies in the past few years.

In evaluating authentication methods on mobile devices, it is vital to assess the opinions of end-users concerning their adoption and use of the plethora of methodologies available. Users are implicitly and actively engaged with an authentication process since, to be authenticated, they need to interact directly with a sensor. For this reason, acceptability and user satisfaction are fundamental aspects to be considered in an evaluation analysis because they can significantly influence the outcome of an authentication system [41].

This work presents an online survey that was undertaken to address and investigate these issues. The aim was to understand users' perceptions of current and future security techniques on mobile devices. The outcome of this survey allows a contemporary assessment of these issues in an ever-changing technological landscape.

The term “mobile device” had been used in our online survey to indicate a portable computing device such as a smartphone or tablet computer. Questions were related to the level of familiarity and trust that users have about traditional and innovative security methods. In particular, the survey focusses on understanding how users perceive personal data stored on a mobile device and the importance they place on being able to protect it. Furthermore, the outcomes reported information on the awareness of the security modalities available on smartphones, including biometrics, and how people would trust them depending on different real-life scenarios. The questionnaire was designed using the website SurveyMonkey [42]. To ensure that all the participants were informed of the content of the questionnaire, definitions of specific terminologies such as “biometrics” and “sensitive data” were presented to participants before related questions.

3.2 Background

At present, passwords and PINs are still the most common security techniques used for the protection of systems and sensitive data, as they are adopted not only as a security method, but also as a second mean of authentication in case the system in place fails the verification. Like all authentication methods, a breach of the security template/information can compromise the integrity of the system. The security attached to conventional “knowledge-based” tokens is therefore essential.

In this context, a survey conducted in 2000 [43] reported on public attitudes towards passwords across 175 subjects (80% male, 74% aged below 35). In this study, the majority of the participants (59%) had a professional computing/engineering background. Of the participants, 91% used a password to protect their personal computers. However, an alarming 34% declared that they had never changed their passwords, and often used the same password for accessing different devices. Respondents also self-declared that they were compromising the protection of their passwords by writing them down (15%) or sharing them with other people (29%).

Ideally, for increased password security it is necessary to increase the entropy, adding uppercase, lowercase and special characters. As a consequence, in the context of a mobile device, the input of a complex password often requires more time and could require a switch to a second or a third page of the virtual keyboard for the entry of special characters.

The authors in [44] reported a study that assessed usability and shoulder surfing susceptibility when inputting a password on eight different virtual keyboards, with the participation of 80 people. Each participant had to enter five passwords; each password was individually generated following fixed patterns of increasing complexity. Entry time and mean error rate (counting of mistyped characters) were the metrics used to assess usability. In the second phase, an experimenter assumed the role of the victim, and the participant acted as a shoulder surfer, noting the password. The results showed that there is a significant difference regarding entry time and mean error rate between different virtual keyboards. The virtual keyboard that presented the lowest performance in the usability analysis was also the most resistant against shoulder surfing attacks. Their findings showed that it is essential to understand and find a compromise between system security requirements and the usability of a PIN or a password, especially concerning timing and entry accuracy.

3.2.1 The general perception of biometric systems

A number of previous studies have assessed public perception on the use of biometric technologies. As an emerging technology, many studies have explored the general principles of biometric modality implementations. However, these assessments have been mainly focused on “fixed” systems such as border controls or desktop computer access.

Moody, in 2004 [45], described a survey conducted to understand the acceptability of a range of common types of biometric systems and usage scenarios. The responses were collected from 300 participants (64% male) of whom only 6% had ever used a biometric system before. The results showed that, at that time, the use of biometrics was deemed acceptable for highly-personal data such as medical records, but not for ATM transactions and online payments. Interestingly, 43% of participants agreed or strongly agreed that biometrics are an invasion of privacy. Participants were subsequently presented with a series of scenarios where they had to express a selection preference for a particular biometric modality from fingerprint, iris, retina, voice recognition, and handwriting recognition. Fingerprint was the preferred modality for logging into personal computers (53%) and for physical access to buildings (58%). Iris and retina scans were the two modalities that people tended to trust more but were often confused for each other. However, iris/retina were also deemed the most intrusive for presentation (41% for iris and 47% for retina).

Moving forward, in 2007 a survey conducted by L. A. Jones et al. [46] reported the outcomes of a survey completed by 115 participants concerning several authentication

technologies, including biometrics. From the survey, participants declared to have familiarity with fingerprint (51.3%), signature (47%) and voice recognition (43.5%). Face recognition, together with hand geometry, is a biometric modality that is still not really known by the population, as around half of the participants responded to be unfamiliar with them. In terms of acceptability, password is the modality most accepted (70.4%) in the financial domain, followed by fingerprint (67%) and signature (63.5%). An interesting outcome from the survey highlight that fingerprint is the modality most accepted in health care (by 58.3%, more than password that was considered acceptable by 50.4%) while signature is the modality preferred in the retail domain (48.7% vs the 44.4% that preferred passwords). Face and voice are the biometric modalities that have expressed more privacy concerns among the ones presented in the survey.

In 2010, the authors in [47] investigated the acceptability of three different biometric systems. The study is based on two experiments; the first involved 100 participants that required the use of a PIN and two biometric systems randomly selected between signature, hand geometry and face. The second experiment also included 100 participants, but it involved the use of only the contact-less hand geometry system. 30% of the participants took part in both experiments. According to the surveys conducted at the conclusion of the experiment, the hand-based system resulted in being the most accepted biometric modality among the participants, and it obtained the most favourable response when considering privacy. Signature resulted in being the modality that participant found more comfortable to use, probably thanks to the habituation factor that influences the acceptability of the systems.

In a further survey conducted in 2010 [48], participants were asked to give their general perception of biometric technologies, and their opinions on keystroke dynamics and face verification systems. A survey was conducted for two months with the participation of 70 volunteers encapsulating students and employees. Less than half (43.5%) expressed a good knowledge of biometric technology. During the study, participants were asked to complete a questionnaire after testing both biometric systems. The keystroke system was preferred for managing access to computer systems by 56.52% of the participants, while 26.1% preferred face recognition systems. On the contrary, when considering physical access, a face system is preferred by 36.23% of participants, and keystroke system by 14.5%. High concern about data privacy issues was reported in the case of face recognition system (46.6%).

This survey was extended in 2012 [49] with 100 volunteers. A decade later than Moody's study, a significantly larger percentage (90.9%) now agreed that the use of biometrics is much more appropriate than secret-based solutions against fraud. There was, however, still a high percentage (47.5%) of participants that had concerns for their privacy when using face recognition. Participants showed acceptance for both the biometric systems, but they were significantly more satisfied (according to a Kruskal-Wallis analysis) with the use of the keystroke system (88.9%) over the face modality (75.8%).

Blanco-Gonzalo et al. [50] conducted a study in 2015 to assess the acceptance of fingerprint recognition. 600 participants were asked to use three planar semiconductor fingerprint sensors, and to answer questions related to their experience and opinions using biometric systems (specifically on fingerprint) and PIN, both before and after the experiment. Around 70% of the users regarded fingerprint to be faster than PIN, and around 80% considered this modality to be more comfortable and secure.

In summary, participants responded quite positively to biometrics, declaring that they are familiar with the majority of the modalities, and the number of participants that would adopt biometric in future had increased throughout the years. Generally, fingerprint is the biometric modality more accepted, especially in scenario where physical or virtual access is needed, while signature is the modality that is more accepted and comfortable to be used when associated to retail and financial use. Face and voice are less known and accepted, in particular participant often associate face recognition with privacy concerns. Overall, across the literature, the number of studies examining user's satisfaction and opinions on authentication mechanisms is low (and often only as part of the evaluation of the performance of a specific system, thus not applicable to all cases). Furthermore, the majority of studies focus on desk-based systems.

3.2.2 Mobile devices

When looking more specifically into the context of mobile devices, only a few studies investigate the users' opinion on mobile authentication systems and focus on real-life scenarios using biometric modalities. In 2005, Clarke and Furnell [51] presented a survey assessing users' attitudes towards security technologies on mobile devices. 297 responses from mobile users were collected over two years through an online survey. Even in 2005, an encouraging 83% of respondents were in favour of using biometrics for the protection of their device. Fingerprint (81%) and voice (79%) authentication were the two modalities that users were most aware of, and would use on mobile devices. Participants were more familiar with iris authentication (76%) than with face (67%). Surprisingly, only 39% would use face as security modality, less than the 43% that would use hand geometry. Hand geometry was known by 49% of participants and keystroke by 44%.

In 2010, a survey that included 548 participants was conducted by [52]. Questions related to the usage and security level of their mobile devices and their opinions of using biometrics as an alternative modality from the one they have adopted. More than the half of the participants (54.4%) responded positively to a possible use of biometrics. When participants willing to use biometrics were asked which modality they would use, fingerprint was the most popular one (87%), but also speaker (20%), face (19%) and gait (9%) recongition were mentioned.

Authors in [53] assessed the usability of the Android Face Unlock system and the Apple Inc. Touch ID. In September of 2014, the authors extended their study with an online survey where 109 and 89 participants were asked questions related to the perception and influence of adoption of the two security systems respectively. 16% of the participants

are Face Unlock users and 36% had previously used the face recognition system, but then decided to stop using it by the time of the survey. Among the reasons of this decision, the majority declaired that they tried the Face Unlock out of curiosity but the technology did not appeal them enough to actually use it. Among the main reasons decleared by the remaining 48% that had never used Face Unlock, not knowing about the technology and the security concerns are the ones that most stood out. In contrast, 69% of participants are current Touch ID users and only 18% decided not to continue using it. As well, the main reason for this decision was trying the technology but not feeling the necessity to use it afterward.

In parallel with our study, in 2016, a survey conducted by Harbach, M. et Al. [54] recorded the opinions of 8286 participants across 8 countries (Australia, Canada, Germany, Italy, Japan, Netherlands, the United Kingdom, and the United States). In line with our survey, the study aimed to understand the perceptions that users have on security systems when implemented on mobile devices, but while our study focused on biometrics and specific scenarios, the one presented by Harbach, M. et Al. assessed the general opinion of sensitive data and security across different nationalities. Significant results from the statistical analysis revealed that demographics and nationality are important variables that influence security adoption. Older participants were less likely to secure their smartphones, mainly considering their protection to be not necessary, while countries like Germany and Japan, that showed a higher level of awareness of sensitive data, were also the countries more likely to consider important the protection of their smartphones.

Table 3.1: Summary of recent surveys of user's perception and adoption of mobile security.

Authors	Year	Participants	Security systems	Main outcomes
Clarke and Furnell, [51]	2005	297	Fingerprint, voice, iris, face, hand geometry and keystroke	83% knows about biometrics. Fingerprint and voice are widely known, while face is still not well accepted.
Breitinger and Nickel, [52]	2010	548	Fingerprint, voice, face and gait	54.4% responded positively to biometrics. Fingerprint is the modality that participants are more likely to use (87%).
Bhagavatula et al. [53]	2015	198	Face and fingerprint	There is a high percentage of participants that decided not to continue using face recognition, compared to the people that abandoned the use of fingerprint
Harbach et al. [54]	2016	8286	Biometrics in general	The perception of sensitive data and demographics have an effect in the security adoption in mobile devices.

Table 3.1 summarise the main outcomes of previous and related studies. Overall, from the studies conducted this past decade, we can notice a positive response to biometrics when adopted on mobile devices, but there is a huge gap between modality like fingerprint, overall accepted and used, with other modalities like face and voice recognition. There is also a lack of a study that consider the user's perspective about security modalities when used in specific scenarios.

We designed an online survey to understand the awareness and perception that participants have with each individual security modality. In particular, compared to previous studies, this online survey investigates the users' opinion on different techniques when applied in specific real-life scenarios and seeks to assess whether the awareness of storing sensitive information influences this decision. The survey also inspects the user's attitudes towards more innovative biometrics such as continuous authentication.

3.3 The online survey

For the current study, a total of 402 participants took part in the survey. Recruitment was online and lasted a month (April 2016). It should be noted that at the time of the survey, not all the security methods described in the introduction of this Chapter were available in the market. For example, the Face ID technology was not available when the online survey was distributed.

Responses were collected mainly from the UK and a minority group from Spain and Italy. There were no sufficient participants from each of the three considered countries to allow a proper comparison between the groups, discarding geographic location as a variable for assessment. The requirements for participation were: (a) being a current user of a mobile device, (b) being aged 18 or over, and (c) have the ability to communicate in English. The term "mobile device" had been restricted to indicate a portable computing device such as smartphones or tablet computers.

The questionnaire consisted of a total of 19 questions, structured into three thematic sections:

- Section A: use of sensitive data;
- Section B: current security modalities;
- Section C: future and emerging modalities.

Questions were related to the level of familiarity and trust that participants have in traditional and innovative security methods. A series of questions use a five-point Likert Scale to understand how much the users agree or disagree with a particular statement [55]. The use of a Likert Scale allows to express the intensity with which the participant agrees or disagrees with each statement.

3.3.1 Demographics

The first four questions collected information on demographics and the type of device participants use. Since the questionnaire was compiled online, unconstrained answers and, indeed, completion, was allowed meaning that some participants did not complete all the questions.

The responses came from two forms of recruitment:

- 249 (123 male, 126 female) responses were from the audience provided by the SurveyMonkey website service and included a range of UK participants that own a mobile device. Out of the total number of responses from this audience, 170 (90 male, 80 female) completed all the sections of the survey.
- 153 (82 male, 71 female) responses were from participants that were contacted by email or social networks like Twitter and Facebook, and the area of origin is not restricted to the UK. 108 (54 male and 54 female) of the total number of responses completed the survey. The age range of participants is more focussed on the range 21-29, accounting for more than half of the total number of participants.

In the following results description, statistics are presented *only for fully completed surveys*. Gender is balanced within all three Sections (51% male). Age ranges are shown in the pie chart in Figure 3.4.

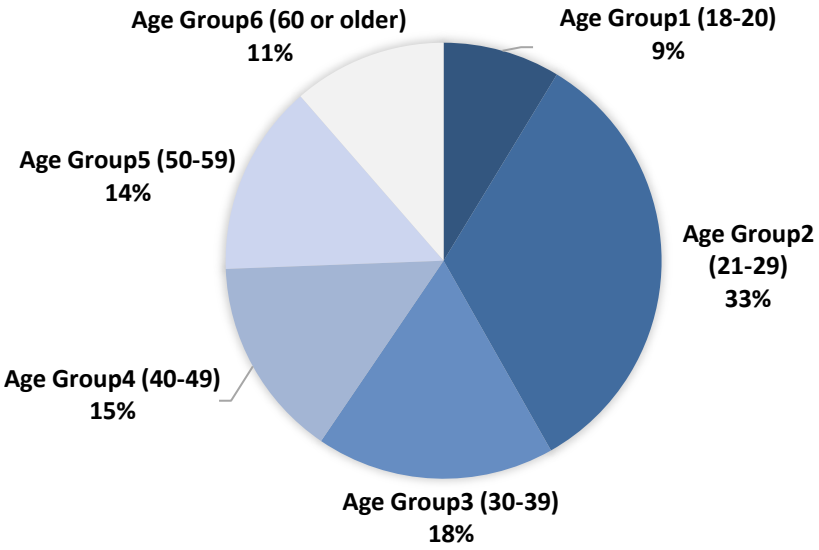


Figure 3.4: Percentages for each Age Group.

During the analysis of the data, specific information on the type of device, model, and OS used by each participant was acquired, thereby enabling an analysis of if these elements influence opinions that participants might have on particular security methods. Participants can be divided into three different groups depending on which OS they are currently using: Android, Windows or iOS. Since participants may own more than one device, they were invited to indicate multiple OS. Android and iOS are the most popular: 50% and 47% of the total number of participants use them respectively. Female

participants tend to own devices running iOS (53% of the total number of iOS users), whereas 57% of Android users are male participants. Participants that use Windows (13%) are of balanced within gender.

Groups of participants can also be considered according to a combination of different OSs used: 19 participants use both Android and iOS, four use iOS and Windows, and 11 participants use Android and Windows. Five of the total number of people use all the three OSs suggested, and only two declared that they use a different OS, one SailfishOS [56] and the other one Symbian (an open-source OS [57]).

Table 3.2: OSs used across age groups and gender.

Age groups per age and gender:		Group A: Android	Group B: iOS	Group C: Windows
Age Group1 (18-20)	**Male**	6	6	2
	Female	9	12	1
Age Group2 (21-29)	**Male**	43	29	8
	Female	28	33	4
Age Group3 (30-39)	**Male**	22	16	3
	Female	13	19	7
Age Group4 (40-49)	**Male**	15	13	4
	Female	15	15	6
Age Group5 (50-59)	**Male**	12	12	6
	Female	15	13	5
Age Group6 (60+)	**Male**	18	14	6
	Female	9	8	5

Whilst there is no significant difference between age groups of participants using different OSs, it can be seen in Table 3.2 that participants between the ages of 18 and 20 seem to prefer iOS devices, whereas Android is preferred for participants between 21 and 29 years old. Windows has approximately the same number of users in all the age groups.

3.4 Sensitive Data on Mobile Devices

Participants were given the following definition of sensitive data taken from the UK Data Protection Act [58]:

" 'Sensitive personal data' is defined in Section 2 of the UK Data Protection Act as personal data consisting of information relating to the data subject with regard to racial or ethnic origin; political opinions; religious beliefs or other beliefs of a similar nature; trade union membership; physical or mental health or condition; sexual life; the

commission or alleged commission by the data subject of any offence; or any proceedings for any offence committed or alleged to have been committed by the data subject, the disposal of such proceedings or the sentence of any court in such proceedings ".

It should be noted that this was the definition of sensitive data active in the year of the survey data collection (2016). This has subsequently been superseded by a new Data Protection Act (2018) accounting for the General Data Protection Regulation (GDPR) that is valid in Europe [59]. It would be interesting for future work to have a further comparison in the user's perception of data protection and observe the difference between the two time periods (pre/post-GDPR).

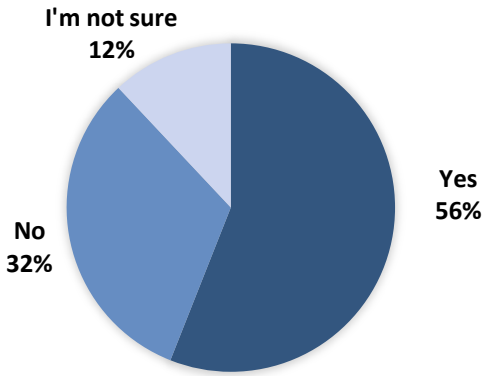


Figure 3.5: The percentage of participants that believe in storing sensitive data on their devices.

From this online survey, more than the half of the total participants (56%) believe that they currently store sensitive data on their mobile device, 32% do not consider having any sensitive data and 12% are not sure (Figure 3.5). If people are aware that they store sensitive data, they might be more cautious in protecting them than people that do not believe that their device holds any such data.

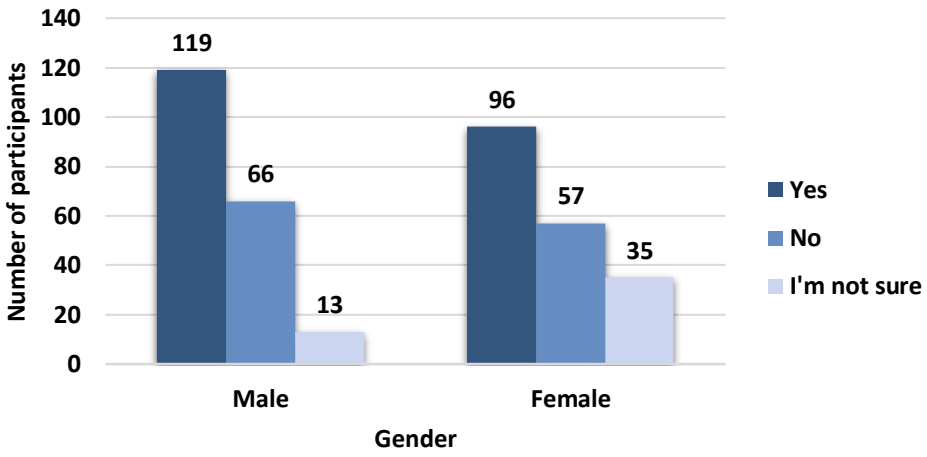


Figure 3.6: Participants that consider that they store sensitive data divided by gender.

In Figure 3.6, it is possible to observe in greater detail the differences between genders concerning sensitive data, where there is a significant difference for χ^2 (2) = 12.95, p = 0.002. The 12% of participants that were not sure of storing sensitive data were mainly female (73%). Approximately 10% more of male subjects considered that they stored sensitive data, while the difference between genders that do not believe that they store any sensitive data is around 8%.

Significant differences can also be noticed looking at the age groups for χ^2 (10) = 24.1, p = 0.007. Figure 3.7 shows the number of participants answering this question divided into age groupings. Participants that believe their device holds sensitive data are more concentrated in the age range between 20 and 39.

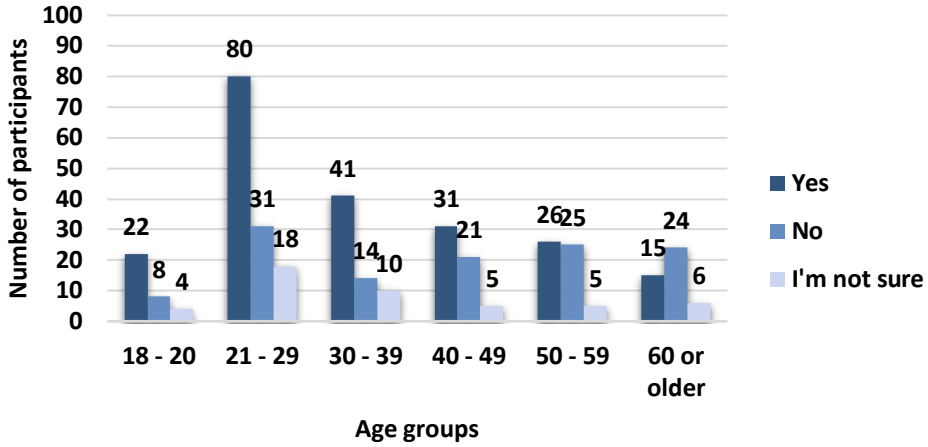


Figure 3.7: Participants that consider that they store sensitive data divided by age groups.

In proportion, in older age groups (between 40 to 60 or older) the number of participants that do not believe their device holds sensitive data are more or closer in number to the ones that answered "Yes". A possible explanation for this difference could be a misinformation that older generation have on security and type of data stored on a mobile device, and it underlines the importance of providing the right information to all mobile users about storing personal information and the risks involved in terms of cyber security. There was not significant difference to notice in the answer to this question between groups using different operating systems.

Participants were also asked what information they consider essential to protect on their mobile device (Figure 3.8) and they were invited to indicate a scale of importance from 1 (not at all important) to 5 (extremely important). Generally, specific apps, such as those interacting with sensitive information (e.g. medical records, health care) or potential financial use are ranked highly (4.46), followed by emails, messages, and other "note" content (4.15). Photographs (3.93) and contacts (3.80) are also considered relatively important, whereas less importance was assigned to the protection of accessing memberships, travel cards (3.54), social networks (3.48) and browsers (3.38). Protecting information from calendars is rated neutral (3.05).

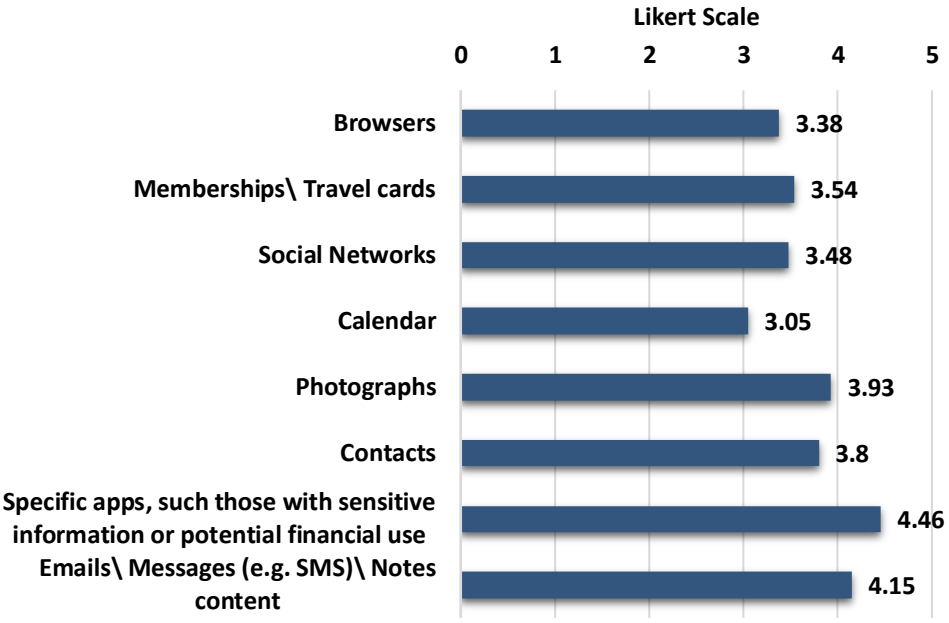


Figure 3.8: Average values on a scale from 1 to 5 related to the importance that participants associated with each element.

Participants were also asked to provide suggestions on items that were not stated in the list but considered necessary to protect. Some interesting applications were protecting call history, location information, stored passwords and passwords hints, and access to the documents stored in the cloud.

Participants in the youngest group (age range 18-20) rate contacts 3.24 in the Likert Scale (Table 3.3) and the importance for the protection of this item increases with age (χ^2 (20) = 43.34, p = 0.002).

Table 3.3: Perceived relative importance of data security across mobile application.

Application	Age Group 1	Age Group 2	Age Group 3	Age Group 4	Age Group 5	Age Group 6	All
Emails\Messages\Notes	4.21	4.06	4.26	4.28	4.11	4.05	4.15
Specific apps	4.21	4.47	4.48	4.49	4.37	4.64	4.46
Contacts	3.24	3.61	3.92	3.86	4.07	4.15	3.80
Photographs	4.06	3.94	4.17	3.79	3.71	3.89	3.93
Calendar	2.79	2.78	3.28	3.36	3.09	3.31	3.05
Social Networks	3.65	3.60	3.78	3.53	2.98	3.18	3.48
Memberships\Travel cards	3.26	3.48	3.55	3.59	3.46	3.91	3.54
Browsers	3.18	3.38	3.49	3.36	3.13	3.73	3.38

Participants aged between 18 and 39 considered the protection of social networks more important than participants aged 40 or older (χ^2 (20) = 36.42, p = 0.014). Possible explanations could either be that a younger population is more active and use more Social Networks than older users, or that younger participants are more aware of the privacy

risk and concern that these types of internet services can bring. There is also a significant difference between genders for the photograph category (χ^2 (4) = 17.22, p = 0.002). It is apparent that female participants consider the protection of photos more important than male subjects.

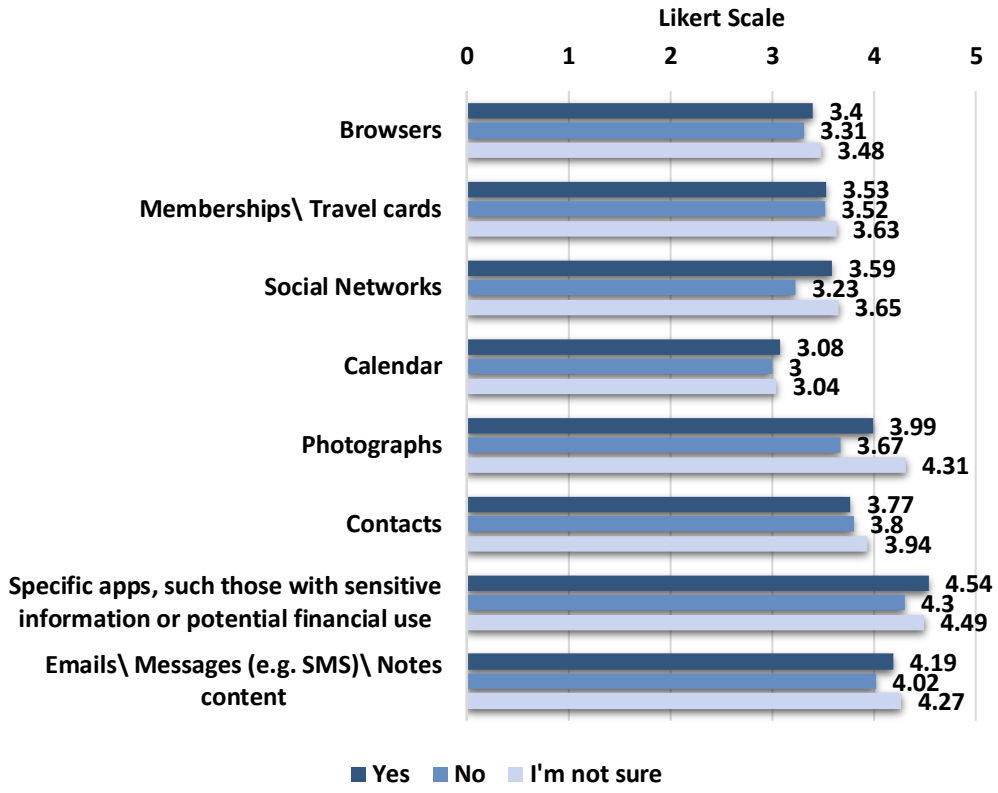


Figure 3.9: The importance level from 1 (not at all important) to 5 (extremely important) assigned by the participants for each app element.

Participants that are unsure of having any sensitive data also assigned a higher level of importance to the protection of the majority of the category when compared to the other two groups. In particular, they consider the protection of photographs (4.31) of slightly more importance than the protection of emails, messages and notes content (4.27) ($p < 0.001$ with χ^2 (8) = 29.84). Participants that believed that are storing sensitive data considered it extremely important to protect specific apps such as those with confidential information or potential financial use (4.54) (χ^2 (8) = 16.25, p = 0.039). The priority of protecting each element divided according to their response of whether they believe they store sensitive information is shown in Figure 3.9.

These results highlight the importance of being informed in what type of data there is on a mobile device, as this information resulted in influencing the perception of security that different stored elements need. Overall, more than the half of the total participants believe to store sensitive data, but there is an evidence of misinformation among demographic groups regarding the sensitivity of data stored.

3.5 Common security modalities

After providing participants with a definition of biometric face, voice and fingerprint verification, they were asked if they have had experience of using each authentication system. Out of 291 completed responses, more than the half (52%) have experienced fingerprint verification, 23% face and 17% voice recognition. The number of people that had experienced fingerprint systems is more concentrated in the two age groups of 21-29 and 30-39, with a significant difference of χ^2 (5) = 27.62, p = < 0.001.

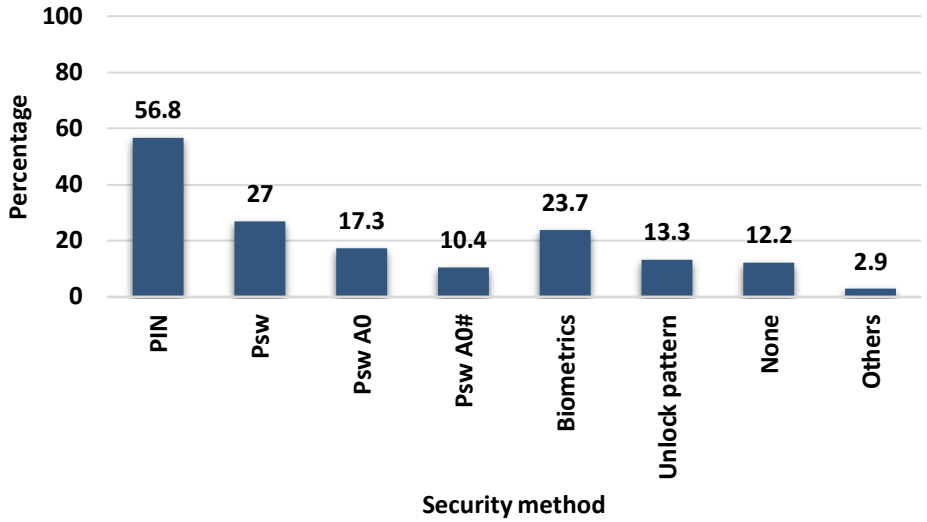


Figure 3.10: Current mobile security methods used by participants. Psw stands for password, Psw A0 stands for alphanumerical password and Psw A0# stands for alphanumerical password with special characters.

Participants were also asked to indicate which security method they are currently using to protect their mobile device. Considering passwords, they were asked to distinguish whether they use a password that contains alphanumerical and/or special characters. Since some participants use more than one device, they were free to indicate more than one modality. Figure 3.10 indicates the percentage use of each modality across the participants. The majority (56.8%) are using a PIN. There is a significant difference between OSs: the majority of the participants that use a PIN (60.3%) are iOS users (p = < 0.001 with χ^2 (2) = 50.92).

Passwords are used by 27%, even if only a few participants (10.4% of the total) indicated the use of complicated passwords with alphanumerical characters. Significant correlation results were also obtained between OSs groups: 53.1% of participants that use passwords are iOS users, 36.5% are Android users, and only 10% use Windows platforms (χ^2 (2) = 8.09, p = 0.018). These differences should be seen considering that different OSs not always provide the same type of authentication modalities and that the user could be limited on using only a few of the security methods presented in the list on their mobile device. For example, Unlock Pattern is not available for iOS users.

Only around 23.7% of the total number of participants use biometrics to protect their devices. Participants that currently use biometrics mainly own an iOS device and use Touch ID - they have more experience with fingerprint verification technologies and trust them more (χ^2 (2) = 33.46, p < 0.001). Biometrics is also preferred by younger participants between 18 and 29 (χ^2 (5) = 18.23, p = 0.003) and people that believe to have sensitive data (χ^2 (2) = 7.26, p = 0.026). 12.2% of users do not protect the access to their device at all and are mainly Android users (χ^2 (2) = 24.9, p < 0.001). 13.3% of participants use the Unlock Pattern.

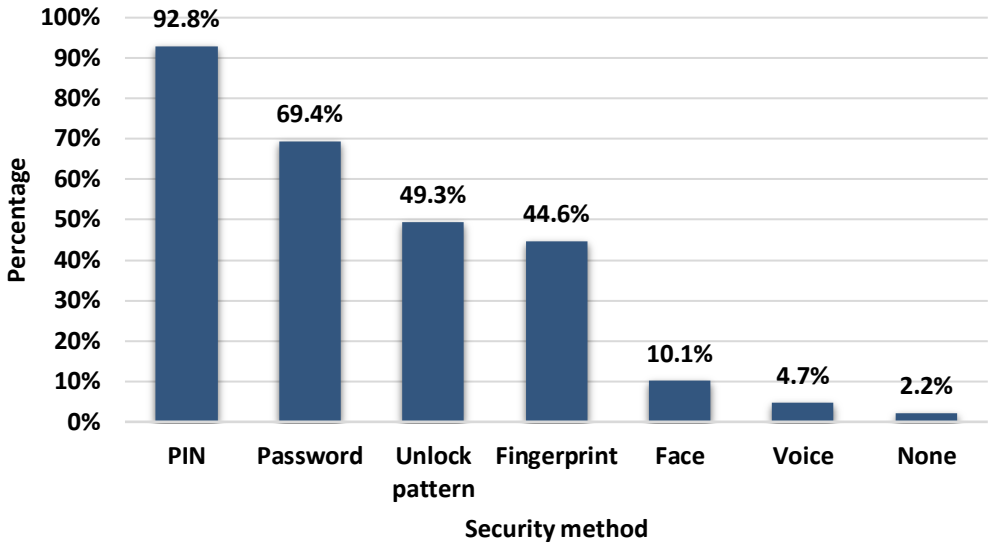


Figure 3.11: Percentages of participants that have experienced various security methods on a mobile device.

Furthermore, participants were asked to indicate the security methods that they have experienced on mobile devices. These results are shown in Figure 3.11 where participants could indicate more than one option. PIN was experienced by almost all the participants (92.8%), and password by almost 70%. Just less than half of the participants had experienced pattern (49.3%) and fingerprint verification systems (44.6%). A few participants had experienced face (10.1%) and voice (4.7%) verification systems. 2.2% have experienced none of the security methods proposed.

Furthermore participants were asked to assign a level of trust indicated from 1 (I would not trust this method at all) to 5 (I would trust this method for sure) for each security modality. These results are illustrated in Figure 3.12. The method that participants trust the most is fingerprint verification (4.12), followed by the traditional password (3.89) and PIN (3.76). Surprisingly, face (3.64) and voice (3.39) are more trusted than the unlock pattern as security methods. These are the average calculated across all the participants, independently on their experience with each method and the availability of these methods on their mobile devices.

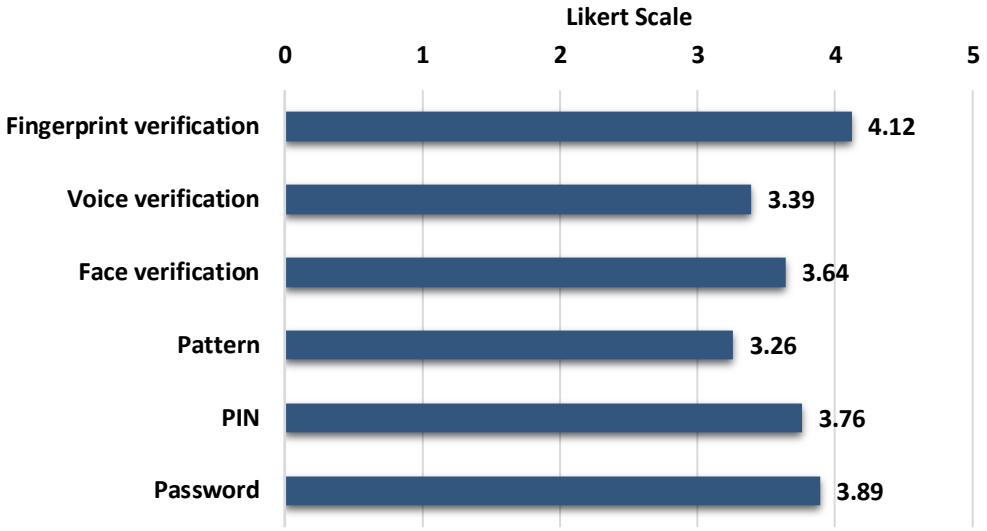


Figure 3.12: Level of trust that participants indicated for each security modality.

These encouraging results shown that despite PIN and passwords are still the most common security modalities adopted, biometrics is gaining trust among the users, especially fingerprint recognition. They also highlight that there is the need to distribute more information regarding voice and face recognition. These two modalities are already implemented in many devices, but the percentages of users that have experienced these two systems are very low.

3.6 Future and new modalities

Participants were asked if they have ever used any of the more innovative biometric verification systems such as iris, gait, and vein verification systems. With the advance of technology, high-quality cameras are now implemented on mobile devices and may be used for iris and vein verification. Despite the fact that iris verification has been implemented and used in smartphone during the past 3 years, it was relatively new at the time of this survey, thus the inclusion of this methodology in this Section. Before each question, a definition of the biometric modality was given to them. As mobile devices have different sensors, additional data can be collected from components such as the internal gyroscope and the accelerometer, and can be used to recognise someone by the way they walk or hold the device.

Touchscreen, GPS, and the keypad can also provide information on an individual and be used for continuous authentication, which is the process of verifying the identity of a user repeatedly (typically in a background task) during the use of a mobile device. Continuous authentication methods assume that the process of authentication is unobtrusive; this is necessary as it is impractical to require users to authenticate themselves explicitly at recurring intervals.

Only a few participants stated that they had experienced innovative methods, not surprisingly, since these modalities are not common, even though some have been already implemented on a mobile device [60], [61]. One of the participants declared that gait verification was one of the modalities they were currently using.

The list of novel continuous authentication elements shown in Figure 3.13 was given to participants as examples of information that can be used as a means to authenticate the user in a non-intrusive way.

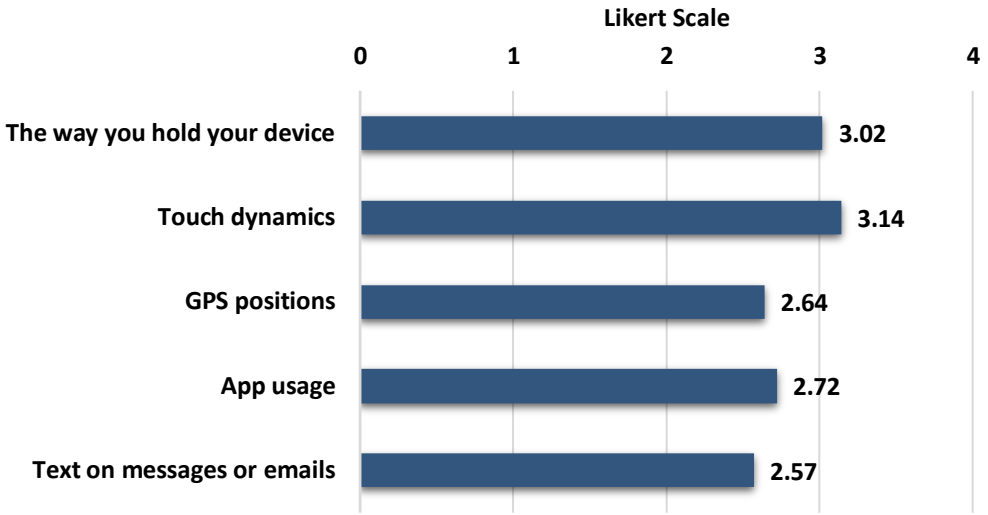


Figure 3.13: The Likert scale of data elements for continuous authentication.

Participants were asked, with regards to obtrusiveness, how happy they would be (from 1, very low, to 5, very high) to use each information for continuous authentication on their mobile device. On average, participants were not happy of using any of the information proposed to continuously authenticate themselves on the device. Results show that the way the user interacts with the device, either through the touchscreen (3.14) or through the way the device is held (3.02), are the elements that people might be happier to use for continuous authentication purposes. Using the device GPS position, or the textual contents of an email or message resulted in an average response of 2.6. Gender has a significant influence in considering the use of textual content of messages or emails for continuous authentication, with χ^2 (4) = 11.36, p = 0.023: females are more indecisive, while males are more polarised in strongly disagreeing or agreeing in using these data.

When asked to the participants the level of trust they would assign to each innovative and new modality presented in this section of the survey, iris verification resulted to be one of the most trusted methods (3.82) (Figure 3.14). Participants replied that they would not trust gait verification (2.60) and that they are not sure if trust vein recognition (3.24) and continuous authentication (3.03). Probably, if participants had the chance to use these modalities, their trust in them can grow.

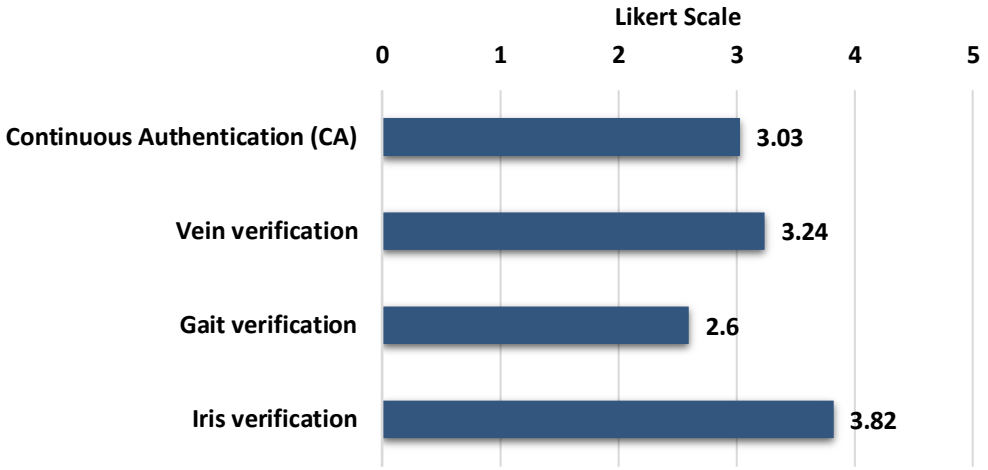


Figure 3.14: Level of trust that participants indicated from 1(ı would not trust this method at all) to 5 (ı would trust this method for sure) for each security modality.

From these outcomes we can underline the importance of providing information to the users, especially which type of data is being collected to perform continuous authentication. Users' negative opinion towards these innovative modalities could change over time, as it was observed with static biometrics in surveys mentioned in Section 3.2.1, especially after experiencing them, but it is fundamental to ensure transparency on what data has been used as it will influence the acceptability and adoption of that particular modality.

3.7 Scenarios

In a final set of questions, participants were asked to evaluate a series of real-life scenarios across which they had to rate on a scale from 1 (very unlikely) to 5 (very likely) in terms of how likely they would use each of the security methods presented in the questionnaire. Since there are fewer responses for the last part of the survey, results are presented only for the 278 participants that completed this part of the questionnaire.

Table 3.4 shows the mean Likert scale value for each modality. The first scenario presented was unlocking the mobile device. Five modalities are rated between 3 and 4 indicating that participants are likely to use them. It is interesting to note that fingerprint and PIN are almost at the same level, indicating that biometrics has become increasingly accepted by the users. Another modality that is often considered trustworthy by the users, even if not experienced before, is iris recognition. It is, in fact, more likely to be used (given the choice) than the unlock pattern implemented on Android devices. Interestingly, voice and face, two modalities already implemented in the context of mobile devices, are not likely to be used.

Table 3.4: Likelihood of modality selection for each scenario

Scenarios	Psw	PIN	Pattern	Face	Voice	Fingerprint	Iris	Gait	Vein	CA
Unlocking the device	3.54	3.9	3.08	2.79	2.53	3.8	3.19	1.92	2.56	2.48
Accessing to a particular App	3.02	3.11	2.47	2.48	2.35	3.02	2.74	1.84	2.29	2.41
Making a purchase online	3.99	3.51	2.43	2.77	2.53	3.36	3.18	1.87	2.54	2.41
Bank transfer	3.91	3.49	2.31	2.81	2.54	3.45	3.28	1.84	2.62	2.35
Accessing Google Services	2.75	2.66	2.28	2.32	2.22	2.64	2.44	1.82	2.19	2.25
Accessing Social Network	3.54	3.14	2.5	2.53	2.39	2.93	2.63	1.8	2.31	2.37

The second scenario considered was protecting the access to a particular app such as for calling or texting, two fundamental functions of a mobile device. Participants indicated that they are unsure of using any specific modality for this purpose or even not likely to use any modality in this scenario. Only fingerprint, password and PIN scored over three on the Likert scale.

The following two scenarios are related to making a purchase online and a bank transfer through the mobile device. For these cases, the likelihood of using a security method is higher, with password the method that has the highest likelihood of use. Fingerprint scores slightly lower than PIN, with iris recognition, also having consideration in these scenarios.

Participants are not likely, or not sure of, protecting the account and the services provided by Google (such as Gmail, Google Maps, etc.) and the access to their social networks. Password is the modality most likely to be used, even if the difference is minimal, as it barely reaches three on the Likert scale.

From these results, we noticed that in general there is a positive acceptance on biometrics. In particular, fingerprint and iris verifications are considered as valid trustworthy alternatives to PINs and passwords. When authenticating on mobile devices the adoption of biometrics is considered more than for the use of financial information. This could not necessary mean that participants do not trust the security of biometrics modalities, but it could also depend on the habituation of using password when it comes to bank transfers and purchases.

3.8 Conclusions and considerations

The research presented in this Chapter aimed to assess users' perspectives of biometric technologies in the context of mobile devices. It is necessary to take into consideration the reliability of the users' responses. Participants were encouraged to respond as honestly as possible to the questions, but obviously, it is not possible to have complete control over the honesty of the answers, especially given the remote nature of the survey collection. They do, however, provide an essential indicator of responses and trends.

Although the majority of the participants claims that they have data that needs to be protected on their mobile device, there is still a high number of people that are not aware or not sure of the presence of sensitive data on their devices. The awareness of having "something to protect" appears to influence the responses that participants gave to the security level they associate with each element. It was identified, for instance, that people who were not sure of storing sensitive data on their devices considered more important the protection of their data compared to people that are aware of having it.

Specific apps, as those interacting with sensitive information, are ranked highly in the scale of elements to protect, followed by the content of emails and messages, photographs and contacts. Less importance was placed on scenarios for accessing memberships or travel cards, social networks and browsers.

There were differences observed between gender and age groups. However, the use of different OSs did not influence the consciousness of storing sensitive data on the mobile device. It may be said that users do not link data protection levels to the choice of OS security. The significant differences observed within demographics highlight that there are still groups of users that are misinformed on the sensitivity of their data and the importance of provide the right type of information to all categories of users.

From the survey's outcome, a shift can be noticed in the biometric systems' acceptance. Compared to previous studies, the population have widely accepted and used biometrics in the context of mobile devices. Fingerprint recognition, in particular, is a modality that participants are most likely to use together alongside the more "traditional" modalities of PIN and passwords. Fingerprint recognition has reached higher acceptance levels and is considered more trustworthy than PIN and passwords in some of the scenarios presented. The reason for this finding could be the successful integration of fingerprint sensors in popular smartphones.

Although face and voice recognition had been implemented on mobile devices for several years, a low percentage of subjects had experienced these two modalities. More consideration should be given to their deployment: although they are widely accepted as biometric technologies, results showed that they are unlikely to be accepted in the context of mobile devices. It should be considered that from the time of the survey, new technologies had been provided from smartphone companies that involve 3D mapping

for face recognition. It would be interesting, for future research, to evaluate the differences after the introduction of this solution, although it might take a few years before this technology can spread over the market, due to the cost of implementation and the availability of this option to the population.

Participants also showed a positive attitude towards the possibility of using iris recognition. This technology has been adopted only recently on mobile devices, but it is surprisingly well accepted. Further research should be conducted to improve usability for iris recognition, as it is difficult to get a good image quality of an iris during the mobile authentication process [62]. Along with the progressive use of these novel modalities, users could become more habituated, inducing higher levels of acceptance.

Even though recently an increasing number of studies have addressed innovative approaches like continuous authentication [63], [64] and gait recognition [65], these technologies, at present, have low acceptance, probably because they are not widely deployed. The most accepted modalities for possible continuous authentication are touch dynamics and the way the user holds the device. Participants considered vein verification, continuous authentication and gait recognition, as modalities that they would trust the least to secure their mobile devices.

When considering different real-life scenarios, it is possible to conclude that PIN and password are still preferred as security methods to protect mobile devices. Participants are more likely to use passwords when it comes to online payments and bank transfers. However, fingerprint verification is considered as a valid alternative, in fact, it is more preferred than passwords for unlocking the screen. In the past few years, near-field communication (NFC) transactions have been adopted to perform contactless payments using a smartphone device. There are more and more apps like Apple Pay, Google Pay and Samsung Pay [66], that allow this kind of transaction authorising the payment with a biometric verification. This could change in future the habituation of having a PIN or a password associated with the financial domain and be encouraging for the adoption of biometrics.

Based on these outcomes, face recognition has been recognised as one of the modalities that should be given more consideration in terms of acceptability and user interaction. Despite being implemented and used in many popular devices, the general opinion is still low. In particular, it would be interesting to investigate the influence on system performance when used in real-life scenarios on user opinion. Likewise, user acceptability and, in particular, interaction can have an effect on the performance on the system itself.

In the following Chapters, this thesis will explore face recognition systems and, more specifically, user interaction with mobile face verification in realistic scenarios. The following Chapter will describe an experimental protocol and data collection and analysis exercise to address the research questions described previously in this work.

Experimental Setup, Preprocessing and Data Extraction

4.1 Introduction

In order to assess the impact that the environment and the user's interaction have on facial images for mobile authentication, we conducted a data collection comprising images collected under varying conditions. We designed a collection process lasting about 30 minutes repeated across three time-separated sessions. During the experiment, participants took facial selfies suitable for verification on a provided mobile device. Participation was voluntary and remuneration was provided following the last session. Full ethical approval was obtained for this experimentation from the Sciences Faculty Ethics Committee prior to the start of the data collection. Facial images and metadata have been collected during the total duration of each session. The experimental setup, data pre-processing and feature extraction are described in detail in this Chapter.

4.2 Experimental configuration

When authenticating using a facial image on a smartphone in a real-life scenario, there are a series of variations introduced by the user and from the surrounding environment that are not predictable when testing such a system in a laboratory. To produce realistic end-use results, the system should be tested in an unconstrained environment, under the same, or at least similar, conditions as to those with which users will be confronted when they use the system on their own. Since there is no existing database comprising images that represent this variability in terms of user poses and non-laboratory-based environments of images taken with a smartphone camera, this study has defined and collected a dataset to specifically address our research questions. With the collection of this database, we addressed three main goals:

- Having facial images collected using a smartphone camera that have lower resolution and less freedom in adjusting the camera settings compared to a fixed system such as that used for passport images.
- Collecting a database for facial verification that can represent a range of realistic scenarios when used on a mobile device.
- For each facial image, having linked metadata from smartphone sensors to be used to assess the user interaction during the biometric presentation.

As well as these main goals, it was an aim to assess facial images with a representative range of variability to verify users across realistic end-use scenarios. Furthermore, to understand in further detail users' perception on mobile facial verification systems, a questionnaire was completed by all participants. The experimental design comprised

three sessions of about 30 minutes each. Each session was separated by a minimum of one day. This ensured that there was the potential for variability in terms of clothing, weather conditions, time of day, etc. It also enabled the collection of a wider range of images for each participant. Details of the experimental setup are explained in the following paragraphs.

4.2.1 Image capturing devices

One of the main objectives of this study is to collect a database of facial images taken with a smartphone camera, where not only the user, but also the acquisition system is moving. The interaction between user and smartphone is unconstrained so it is not possible to predict the exact location nor the external factors that might influence the image taken and hence the verification outcome. Even the distance from user's face to the camera can vary and neither the user's pose nor the camera placement is fixed. One of the first considerations for the data collection was the location as to where users typically access their mobile devices. The data collection was planned to include scenarios where participants are free to take images with no constraints as they would do for daily-life tasks.

Another consideration was to compare a passport-style facial image to a constrained facial image taken with a smartphone camera (Google Nexus 5). This would allow a comparison between a fixed and a mobile scenario. To obtain images in a passport scenario, we used a Single Lens Reflex (SLR) camera (Canon EOS 30D) and followed the procedure defined for the collection of passport images following the photography recommendations described in Section C.2 of the ISO/IEC 19794-5 Biometric sample quality standard [8]. The aim of this investigation is to establish the differences between the two scenarios across the different camera devices, as well as to verify whether the same procedure adopted for facial passport images can be applied to a mobile scenario.

Furthermore, we wanted to check whether enrolling with an SLR image would result in a higher performance for facial verification than an enrolment with images taken with a smartphone camera. We hypothesised that the images taken with the SLR would have a better quality, and therefore more resilient to subsequent verification image variation that an unconstrained scenario can create. The camera specifications for both types of devices are summarised in Table 4.1.

Table 4.1: Camera specifics for the capturing devices [67][68].

Camera specifics	Canon EOS 30D	Google Nexus 5
Type	Digital AF/AE SLR	Selfie camera
Pixels	8.5MP	1.3MP
Focal length (35mm)	35mm	33mm
Sensor Pixel Size	22.5 x 15.0mm	1.95 µm
Autofocus Features	Autofocus 9 point	Fixed focus

The same conditions were applied to all participants. An image of both capture devices is presented in Figure 4.1. Images from the SLR were collected at the beginning of the first session in an experimental laboratory where the environment replicated the constrained and controlled enrolment scenario for passport images. Users were asked to be seated in a chair in front of a solid white background, with fixed artificial light.

Canon EOS 30D **Nexus 5**

Figure 4.1: The two capturing devices used during the data collection: the Single Lens Reflex (SLR) on the left hand-side [69] and the Nexus 5 smartphone on the right hand-side [70].

A total of 6 images were taken with the SLR camera, with the camera operated by an operator. Participants were seated in a chair that was placed 2 m away from the camera and were asked to assume a neutral facial expression and to look directly at the camera mounted on a tripod. Under the same conditions, each participant took 5 images with the smartphone camera (with the camera operated by the participant). The difference between the two types of images is that despite having the same conditions as for passport image collection, while using the smartphone the acquisition camera can be moved unlike the SLR which was at a fixed distance.

To avoid any variability in terms of resolution of the mobile device camera (and settings), the same model of device was provided to each participant. The participants then used the smartphone for the remainder of the session to take images in an unconstrained environment outside the laboratory.

4.2.2 Location types

In order to have an element of control within the unconstrained scenario, we decided to select an approximate area in which the images needed to be taken. A map was given to the participants containing 10 locations that needed to be visited, with an image to be taken at each location. In each session the participants were given a different map (A, B, or C). The locations varied: indoors and outdoors, crowded and less crowded, and were

representative of locations where smartphones are typically used in everyday life (cafés, streets, corridors of a building, etc.). Figure 4.2 illustrates one of the maps followed by the participants. The map shows a section of the University of Kent campus. The participants were guided to 10 different locations starting from the experimental laboratory where they started the data collection. The route finally returned to the experimental room, completing the session.

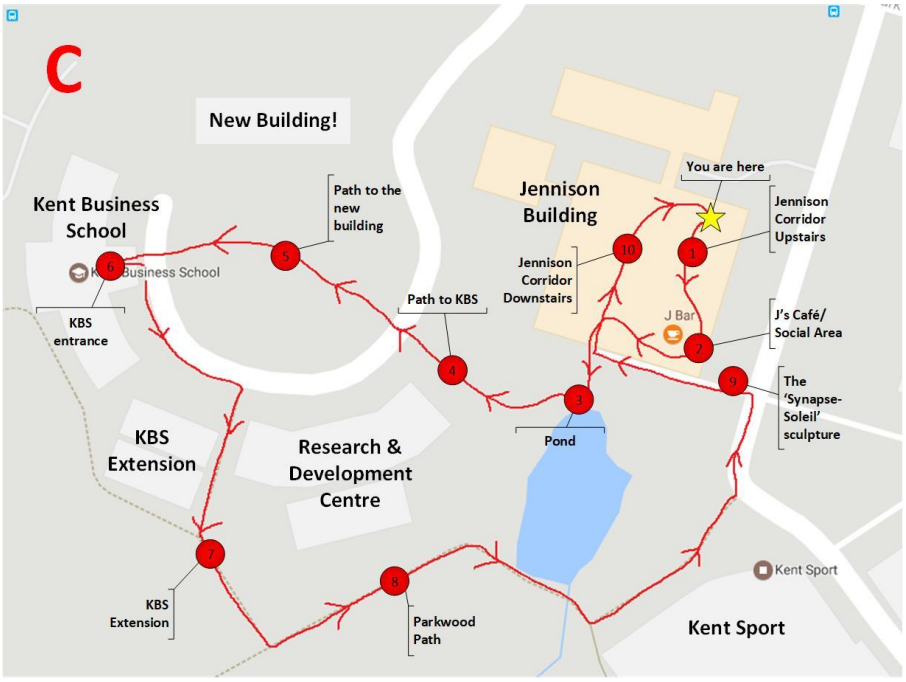


Figure 4.2: An example of one of the three maps used during the data collection.

At each location, participants were instructed to proceed with the acquisition of facial images for perceived biometric verification use. There was a minimum requirement of 5 images for each location, but participants were free to take more images if they wanted. For safety reasons, participants were also warned to not to walk while using the device and the locations were chosen to be both safe for the participants and legitimate areas for smartphone use. Five of the locations were identical across all three maps, while the other five locations differed.

4.2.3 Scenarios

The data collection was structured to assess four different scenarios. A first scenario involves the collection of facial images where the users are sitting on a chair in an experimental room with facial images taken with an SLR camera in a fixed position on a tripod. In further scenarios, the user collects facial images using a smartphone device that has no fixed constraints as with the SLR in terms of position, but the environmental conditions are the same as for the previous scenario - the participant seated on a chair in the same experimental room.

The third and fourth scenarios both involve the collection of facial images outside the laboratory by using the same unconstrained smartphone device. The locations in these two scenarios are considered to be facial images taken indoors and outdoors respectively while both the acquiring device and the user are moving.

A summary description can be seen in Table 4.2.

Table 4.2: Scenarios description.

Scenario	Environment locations	Person fixing	Camera fixing	Type of device used
1	Indoors	Seated	Constrained	SLR
2	Indoors	Seated	Unconstrained	Smartphone camera
3	Indoors	Moving	Unconstrained	Smartphone camera
4	Outdoors	Moving	Unconstrained	Smartphone camera

4.2.4 Application development

To collect the facial images and the background device metadata, we developed an Android app to automate the data collection process. The app was developed and designed using Android Studio [71].

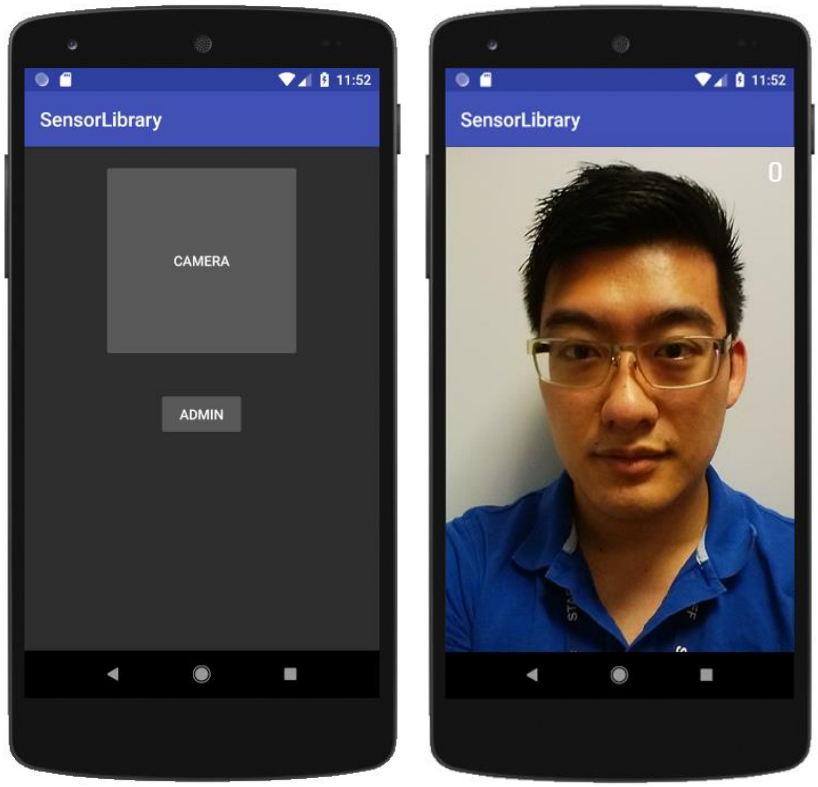


Figure 4.3: Interface of the mobile application used for the data collection.

The only instruction that participants received from the operator was to take the facial selfies for verification: they were advised to ideally present a neutral expression and a

frontal pose to the front-facing device camera, but they were free to move as they deemed necessary, assessing and adjusting for lighting conditions and image background that, in their opinion, was suitable for biometric verification. The user interface of the application is shown in Figure 4.3.

The user had to only interact with a button labelled “Camera” to launch the image capture activity. The application did not have an in-built biometric system, rather the use was exclusively for the collection of images and the metadata. Once the capture screen was launched, a video preview of the participant facial image was presented. The participant could press anywhere on the screen to take the image. A counter was displayed at the top right of the screen informing of the number of images still required to be taken to reach the minimum at a particular location.

Once a session was completed, the operator used the “Admin” button to store the session device metadata on the smartphone. The device metadata had been recorded in the background for the whole session. The facial images were saved at the moment they were taken and stored on the smartphone. At the end of each session, the operator transferred all the files created for the device metadata in a .csv format, and all the facial selfies were saved as jpeg images. After saving all the data from the smartphone, all images and files created for that session were deleted using the “Admin” button to clear the smartphone of any data of the previous participant, so that the next participant could not open or gain access to any data from previous users.

4.2.5 Ethics

As the experiment involved human participation, and the biometric data collected is of a sensitive nature, ethical approval was needed before starting the data collection. An application together with a proposal for the experiment was submitted for the ethical approval. The experimental procedure was reviewed and approved by the Science Research Ethics Advisory Group at the University of Kent [72]. During the first session, participants were informed about the nature of the study, given a Participant Information sheet to read and if they agreed to take part of the data collection, they were asked to sign the Consent Form prior to beginning the study.

4.3 Database description

A total of 9,728 images from 53 participants were collected, from both the smartphone camera and the SLR. We assigned a file name to each image to signify properties of a particular image. The first part of the image filename was P (participant) followed by a number that indicates the participant identifier (from 01 to 53). This was followed by a single letter indicating the sex (M or F) of the participant. The next section of the filename comprised of the letter S (session) followed by the numerical identifier of the session in which the image had been taken (01, 02, or 03). Following this, the map identifier letter of the map used by the participant as locations (A, B, or C) was denoted. This letter was followed by a number signifying the image count within the session. This number has a

minimum limit of 50, but it varied from participant to participant, since some of them decided to take extra images during the data collection. The last part of the file name is the timestamp signifying when the image had been taken; this was separated by an underscore "_" from the first part of the file name. Timestamps were saved in the UNIX epoch time format. For example:

P01MS01A1_1488801830.jpg

indicates an image taken from the participant 01, who is a male, during the first session, where map A was used, and it is the first image of a minimum set of 50, taken at the time 1488801830 (6/03/2017 12:03:50). Of all the participants, only one (P27M) did not complete all three sessions.

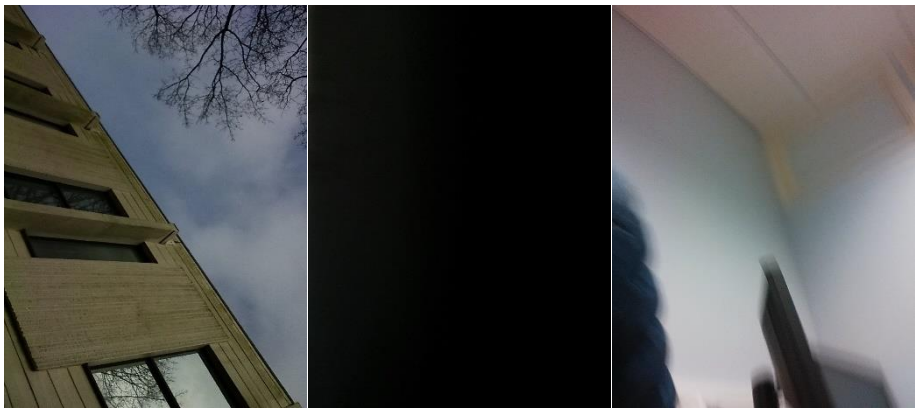


Figure 4.4: Examples of images taken by the participants by mistake.

The database of images was pre-processed to remove captures that the users took by mistake that did not include a facial image (e.g. when walking from one location to another, or when keeping the smartphone in a pocket), as shown in Figure 4.4.

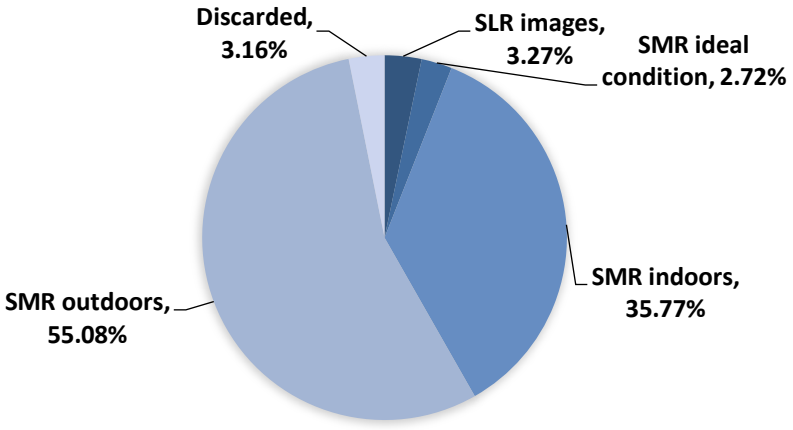


Figure 4.5: Diagram showing the total images collected. SLR is indicating the images collected using a Single Lens Reflex. SMR indicates the images collected with a smartphone camera.

Following this process, a total of 9,421 images were contained in the cleaned database (9,103 from the smartphone camera and 318 from the SLR). Figure 4.5 describes the division of the images. In total, the database contained approximately 180 images per participant.

4.3.1 Demographics

At the beginning of the data collection process, participants were asked to complete a form about their demographics as well as information about their previous experiences with biometrics. Across the dataset, subjects' sex was approximately balanced, with a total of 26 male and 27 female subjects. As most participants were recruited from a university environment, the age groups were skewed as can be seen in Figure 4.6.

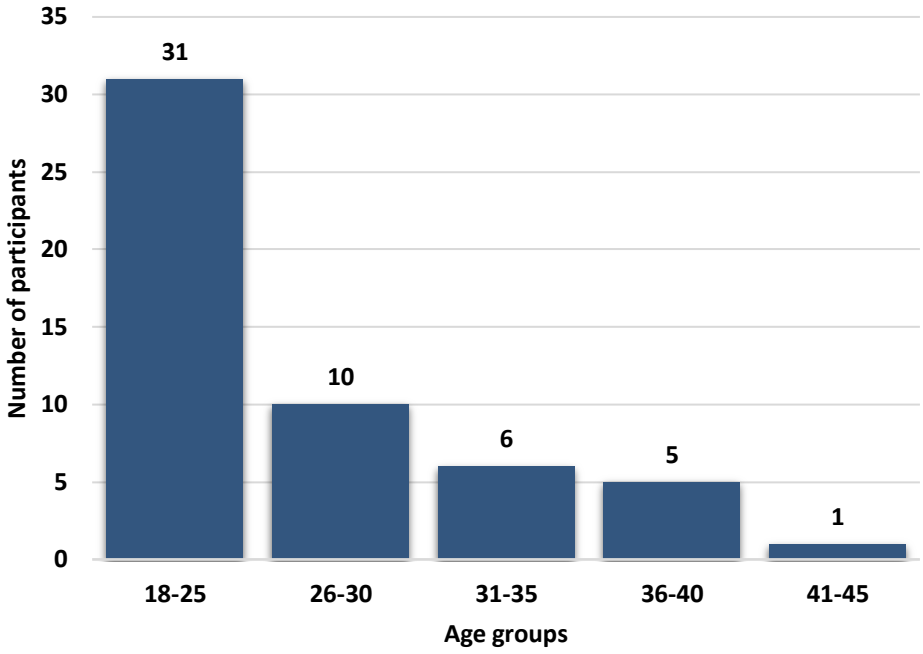


Figure 4.6: Histogram of participants' age.

Information about the participants normal writing hand was requested to establish if a relationship exist between the hand with which the users hold the device while taking the facial images and their handedness. The majority of participants self-declared to be right-handed (47). Only 2 participants declared to be left-handed, with 4 as ambidextrous. Given this distribution, there is not enough information to analysis this relationship. Even a manual visual analysis is not possible to determine the hand used for holding the device from the image, but this could be of interest for future research.

We asked participants to indicate their ethnic origin to enable an analysis as to whether there were differences in detecting and extracting facial features with a different ethnic facial grouping. We divided participant into groups as categorised and described by the NIH [73] to enable uniformity and comparability of data on race and ethnicity. The description and the number of participants for each group are indicated in Table 4.3.

Table 4.3: Description of ethnic groups and number of participants.

Code	Ethnic group	Countries included	Participants
1	**American Indian or Alaska native**	A person having origins in any of the original peoples of North and South America	0
2	**Asian**	Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam	17
3	**Black or African American**	A person having origins in any of the black racial groups of Africa	6
4	**Hispanic or Latino**	A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race	1
5	**Native Hawaiian or other Pacific Islands**	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.	0
6	**Caucasian**	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa	29

Participant were also asked to indicate the completed level of their education. This information was collected to inform us about the participants educational background and check whether they are informed about technology due their academic attainment level. As the study took place in a university, most of the participants had been awarded either a Masters (17) or bachelor's degrees (18), mainly in a scientific discipline. There were also 14 participants at a pre-university qualification level, and 4 participants with a PhD (Figure 4.7).

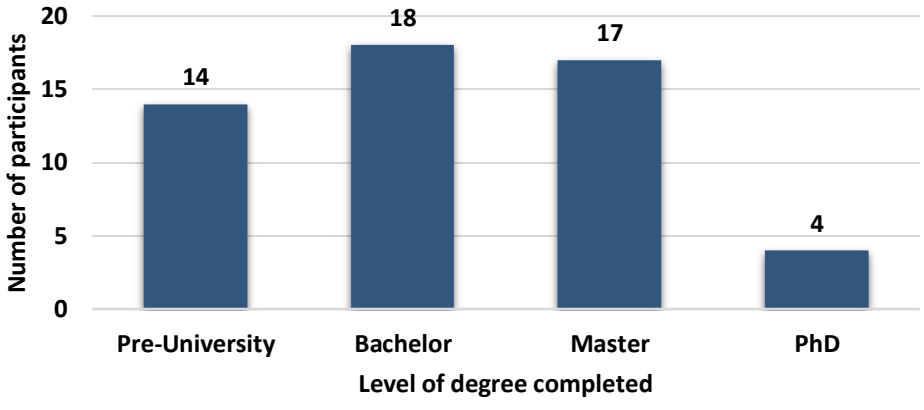


Figure 4.7: Number of participants for each completed levels of degree.

Each participant provided information about which mobile operating system (OS), or systems in case of more than one device, they are currently using, to check if there are any usage differences when comparing their currently used OS to the one proposed for

the data collection. Out of 53 participants, around half of them (26) stated to own only one mobile device (either a smartphone or a tablet), 17 declared to have 2 devices and 9 stated they use 3. Only one person declared the use of 4 mobile devices. The total number of Android users was 38, while there were 25 iOS users. 8 participants declared the use of both Android and iOS, while only 1 Android user has also a Windows mobile device.

The following questions related to biometric experience: 47 participants have heard of biometric systems typically in the media or when they applied for passport. There was a group of 17 participants that have either studied or taken part in a previous study about biometrics in a security context, while a smaller group of 6 people had attended a module at university regarding security and biometrics.

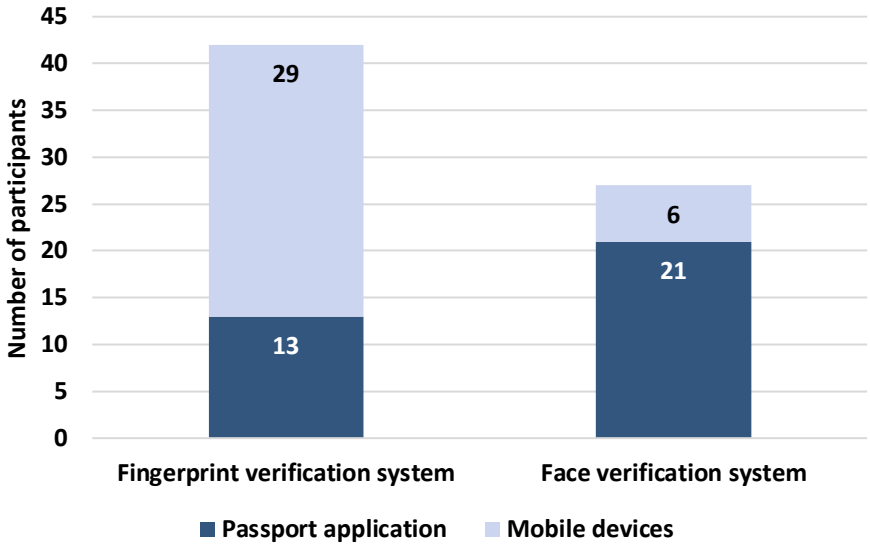


Figure 4.8: Differences in previous experiences that participants had with fingerprint and face verification.

40 participants had experienced a biometric system prior the data collection, but only 20 had experienced them on a mobile device. Face and fingerprint verification systems were the biometric modalities most mentioned by the participants (Figure 4.8). In particular, 42 participants had prior experience with fingerprint recognition, either through airport passport control (13) or mobile devices (29).

21 participants had experience with facial recognition systems, all declaring that they had utilised this technology when crossing the border within automated passport control. Only 6 participants had experienced facial recognition on a mobile device, all in the context of research purposes. This underlines the importance of our study to understand the aspects that influence facial recognition on mobile devices to make it accessible and accepted by the population.

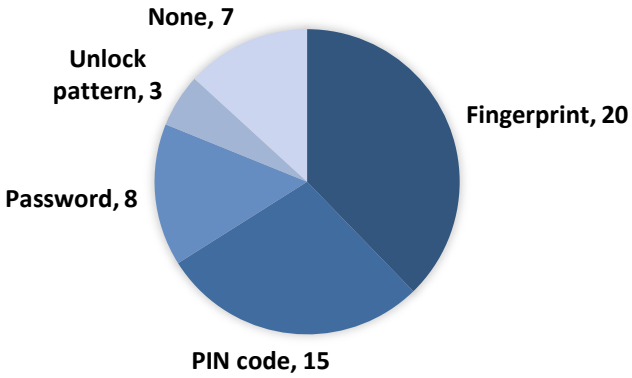


Figure 4.9: Security modalities adopted by the participants on their mobile devices.

Finally, we asked participants if they secure their own devices and what modality they are using Figure 4.9. Out of 53 participants, 46 protect their mobile device with a security system. 20 participants utilise fingerprint recognition on their smartphone, with either a password or a PIN code as a secondary security means. 8 participants use passwords, 15 use a PIN code and 3 an unlock pattern. A total of 7 participants did not protect their device with any security system.

4.3.2 Images

All the images taken with the smartphone camera had been saved on the device in Jpeg format with a resolution of 96dpi and dimensions 960x1280. As previously mentioned in Section 4.2, participants were asked to attend three separate sessions allowing for the collection of a number of facial image variations from the same subject across a range of capture scenarios. As an example, Figure 4.10 shows three images taken in the same location but in different sessions by the same participant.

Session 1 **Session 2** **Session 3**

Figure 4.10: Facial images taken in the same location by the same user in three different sessions.

It can be noticed that there are changes in the surroundings in terms of lighting and background, despite the image being taken in the same location, as the user can move freely and decide where to take the image within the area specified on the map.

Furthermore, the participant presents a different hair style across the three sessions. For example, in Session 3 hair could create occlusion over the face, however this cannot happen in Session 1.

4.3.3 Device metadata

Most Android devices are embedded with sensors that can provide raw data describing the motion, orientation and environmental conditions. The data from these sensors was collected throughout the whole duration of the session to constantly record both the device and the user's movements while collecting the facial images.

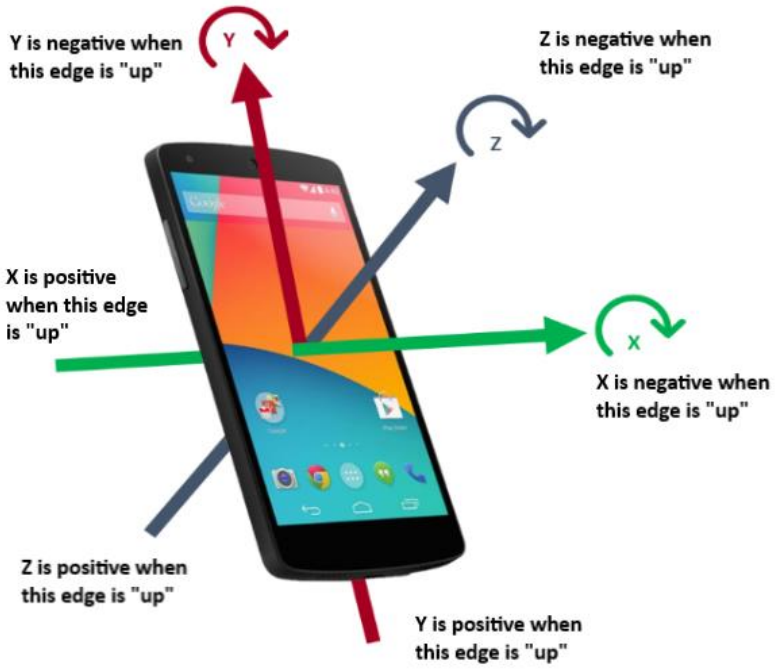


Figure 4.11: Representation of the physical axes of the smartphone.

It was hypothesised that these types of data contained information about the position and movements of both the user and the device at the moment of the biometric sample acquisition. Among the different sensors and types of information available on the chosen mobile device, we selected those that we anticipated were going to provide the information needed to assess the user's interaction and movements with the device. In this Section we explain which sensors were used to collect information during the data collection.

4.3.3.1 Accelerometer

The accelerometer is a hardware-based sensor that allows the detection of device movement. It measures the acceleration force in m/s^2 that is applied to each of the three physical axes (x, y, and z) as shown in Figure 4.11, with returned values including the force of gravity. Each axis returned a value between -20 and 20 at a timestamped frequency of 100 Hz.

4.3.3.2 Gyroscope

The gyroscope is a hardware-based sensor that measures the rate of rotation in rad/s of the device around each of the three physical axes (x, y, and z). The ranges recorded for the x and y axes are between -15 and 15, while the z axis recorded a value between -20 and 20 at a frequency of 100 Hz.

4.3.3.3 Activity

The ActivityRecognition API [74] on Android allows the recording of a DetectedActivity parameter that gives an estimation of the type of activity the device is performing returning a value between 0 to 100 that represents the likelihood that is performing a particular activity. The larger the value, the more likely the event is occurring within the data.

The API returns a value that represents each activity as described in Table 4.4.

Table 4.4: Description of constant values for DetectedActivity [74].

Parameter	Name	Description
0	**IN_VEHICLE**	The device is in a vehicle such as a car
1	**ON_BICYCLE**	The device is on a bicycle
2	**ON_FOOT**	The device is on a user who is walking or running
3	**STILL**	The device is not moving
4	**UNKNOWN**	Unable to detect any activity
5	**TILTING**	The device angle relative to gravity changed significantly (e.g. when the user picks up the smartphone from a desk, or the device is on a user that change from sitting to stand up
7	**WALKING**	A sub-activity of ON_FOOT. The device is on a user who is walking
8	**RUNNING**	A sub-activity of ON_FOOT. The device is on a user who is running

4.3.3.4 Proximity

Proximity is a built-in sensor that collects information about the presence of an object (measured in cm) relative to the screen of the device. The sensor is located on the top front part of the screen as shown in Figure 4.12.

This sensor can be used to detect if the device is located in a person's pocket or is held to a person's ear during a phone call. When an object covers the sensor, the screen turns off when in call mode or when locked, to avoid pocket calls or to accidentally activate other functions. It returns a value of 1 if the object detected is at a distance of less than 5 cm and a value of 0 when the distance is over 5 cm, or no object is detected.

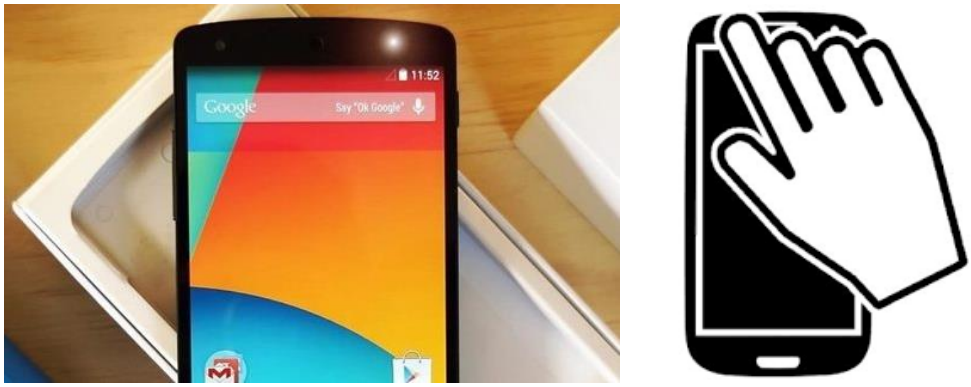


Figure 4.12: Proximity sensor located on the top part of a Nexus 5's screen [75]. On the right hand-side the icon that indicates the proximity sensor when active [76].

4.3.4 Users opinions and perceptions

Participants were required to complete a questionnaire at the end of each session to record their experience during the experiment. There was a total of 15 questions (Table 4.5).

Table 4.5: A summary of the questionnaire and the topics asked at the end of each session.

Number of questions	Description
Questions 1-4	Questions related to the participants' level of comfort of presenting the biometrics in unconstrained environments and in presence or not of other people
Questions 5-7	Questions related to the participants' level of confidence of providing a good biometric presentation in unconstrained indoors or outdoors locations
Questions 8-10	Questions related to the participants' level of confidence of providing a good biometric presentation in presence or not of other people
Questions 11-13	Questions related to the description of how difficult it was for the participants to perform the presentation of facial images on a mobile device
Questions 14-15	Question to check the participants overall experience and the likelihood with which the participants were to use a facial recognition system on their smartphone and whether their opinion changed after each session

The questionnaire was intended to check whether users react differently according to the different scenarios (indoors, outdoors, other's people presence). All answers were measured on a Likert Scale [55] that ranged from 1 (strongly disagree) to 5 (strongly agree). Participant had to indicate for each question to what extent they agreed or disagreed with each statement provided. It was possible to indicate a neutral option by awarding a mark of 3 in the middle of the scale. Table 4.5 provides a description of the questions that were presented in the questionnaire.

4.4 Preprocessing and data extraction

Based on the research questions that we wished to address, we considered our analysis according to the diagram shown in Figure 4.13.

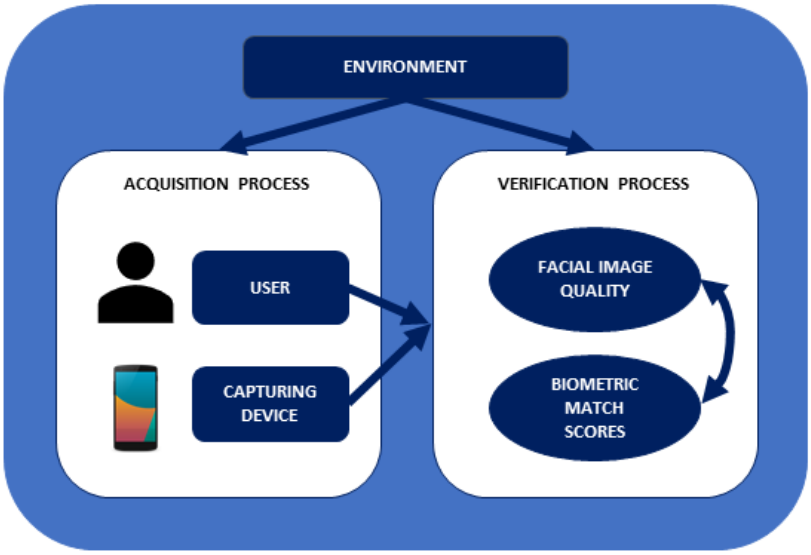


Figure 4.13: Diagram of relationships considered in a mobile face verification system.

The diagram shows the contributory variables that we wanted to investigate, and their relationships indicated by the arrows. These relationships can be explored across different types of environment. The acquisition process in mobile scenarios is not a fixed system. Both the user and the smartphone can move freely. In the verification process, facial image quality and biometric match scores receive influence from the user interaction and the capturing sensor. All variables are under the influence of the different environments.

4.4.1 The environment

In our analysis we have two types of environmental locations: indoors and outdoors. The indoors environment includes the experimental laboratory where the participants took the images while seated using both the SLR and the smartphone camera, and all the images taken in unconstrained scenarios that were acquired inside a building. The outdoors environment corresponds to unconstrained images captured outdoors where both the user and camera can move.

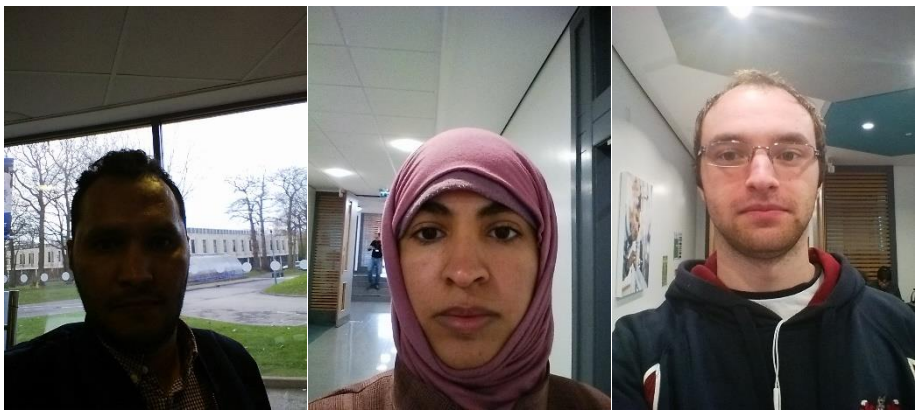
4.4.1.1 Scenarios 1 and 2

The experimental laboratory consisted of a room without any windows, lit by artificial light and white walls forming a background for the image capture. In this location, two different capture systems were considered in order to compare images captured under the same conditions as used for passport images, and a mobile acquisition process using a smartphone camera. Figure 4.14 shows an example of two images from the experimental laboratory; one taken with the SLR and the other when using a smartphone camera.

Figure 4.14: Comparison between an image collected with the SLR (on the left hand-side), and an image collected by the user with the smartphone camera (on the right hand-side).

4.4.1.2 Scenarios 3 and 4

Unconstrained scenarios 3 and 4 were considered under two different location types: images taken with the smartphone camera taken indoors and, separately, outdoors. These two types of location presented different aspects in terms of variation regarding light exposure, background complexity and user pose.



(a) **(b)** **(c)**

Figure 4.15: Examples of images taken in indoors locations.

Indoors locations usually present artificial lights, or a combination of artificial and natural light when images have been taken next to a window for instance in Figure 4.15 (a). The background in this scenario is predominantly white walls (b) and sometimes other people appear or there are paintings or posters with other faces (c).

The fourth scenario considers images that the participants have taken outside. Figure 4.16 presents examples of images taken in location type. Images taken in these scenarios are mainly influenced by weather condition and natural light that changes between the time of the day (a) and the sun position (b). Wind and rain also play a role in these images, for example in (c) the participant's hair moved across their face during the image capture process.

(a) **(b)** **(c)**

Figure 4.16: Examples of images taken in outdoors locations.

4.4.1.3 Background analysis

The image background in each scenario varying depending on the capture location. Outdoors image backgrounds comprised mainly buildings with a monochromatic colour or trees, and sometimes people walking behind the users, while indoors images were characterised by white walls when mainly taken in a corridor but could present wall fixings or other people passing by, especially in crowded locations such as a café.

To assess the complexity of the background, we removed all non-facial areas of the image and subsequently performed a texture analysis. Facial regions were detected using the Viola-Jones method [77]. After isolating the facial area, the background of the image was segmented in four pieces (top, bottom, right and left sides of the facial area). For each segment of the background, we calculated three metrics to quantify the complexity of texture.

Texture analysis is useful to assess the local spatial variability of the pixel intensity values in a region of the image. We considered three statistical metrics, calculated on the images in grayscale:

- **Texture Range**: the local range of the image, calculated as the difference between the maximum value and the minimum value of a 3-by-3 neighbourhood window around the selected pixel. Low values characterise a smooth texture, while higher values are more typical of a rougher texture.
- **Texture Standard Deviation**: the local standard deviation of the image is calculated considering a 3-by-3 window of intensity values around the considered pixel, and considering symmetric padding when calculating the pixel values that are at the border of the image.
- **Texture Entropy**: the local entropy is calculated of a grayscale image considering a 9-by-9 neighbourhood around the selected pixel. The local entropy, calculated according to [78], describes the randomness of an input image.

The resultant texture values were normalised on a scale from 0 to 5, an average has been calculated for each of the background sections and an overall value has been calculated to describe the complexity of the whole background.

Images where a face could not be found were cropped manually using the MATLAB ***imcrop*** command to select the facial region. Cropping the image in this way enabled the analysis of the complexity of the background and an understanding as to why faces were not detected in these images. Face detection has been shown to be affected in particular by the presence of other faces in the background. This can create noise within the background and, depending on the face detection algorithm, can be recognised as the actual user's face and fail the verification. Figure 4.17 presents examples of images where more than one face has been detected and an erroneous attempt has been made to distinguish the user's face.

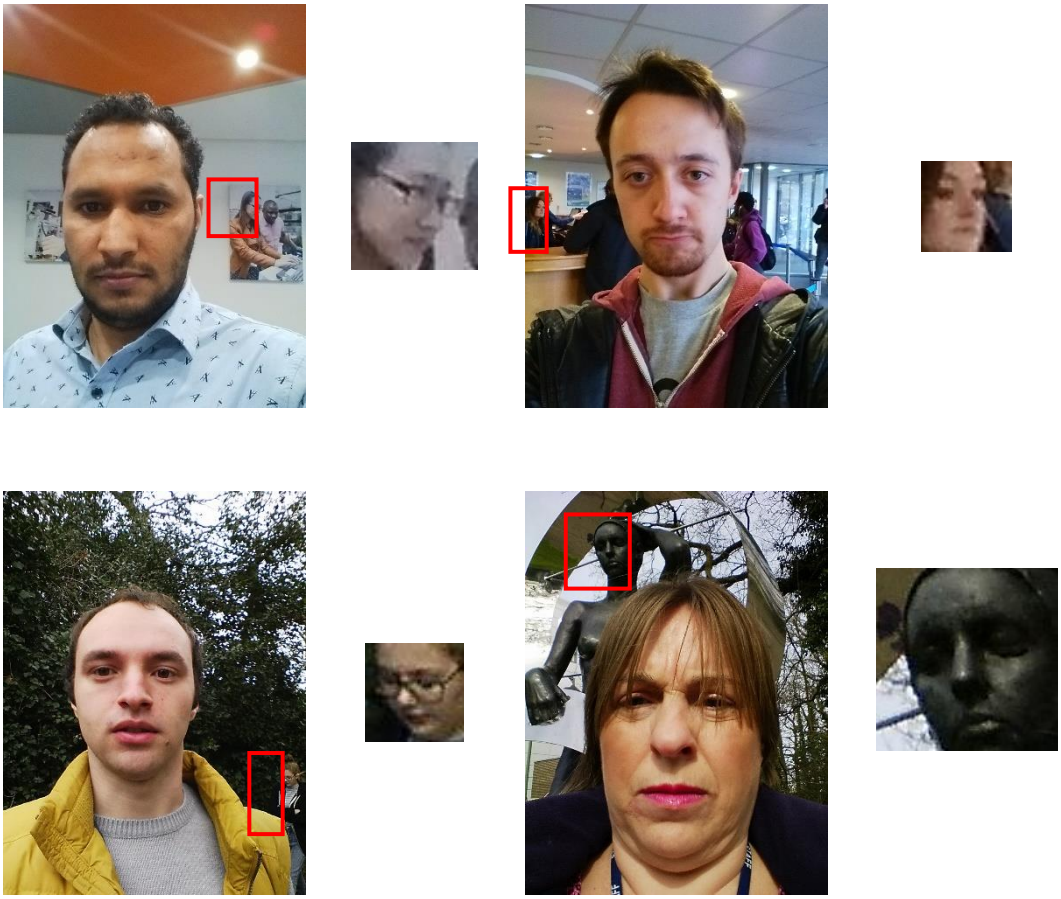


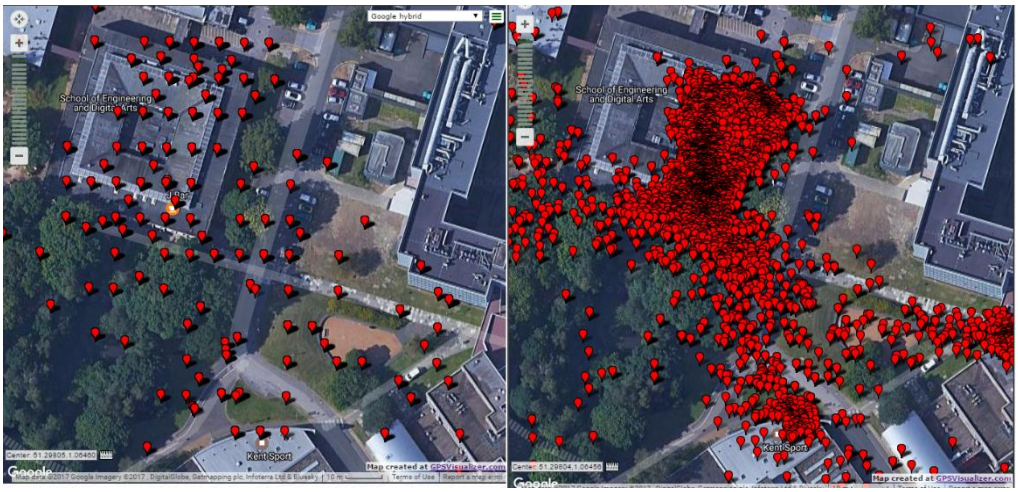
Figure 4.17: Examples of images with more than one face detected.

4.4.1.4 GPS and Wi-Fi location

From each image, we checked the GPS metadata to automatically detect the location as to where the image had been taken and if there was a way to consider GPS as information to distinguish between indoors and outdoors locations.

Unfortunately, a GPS location was not recorded for each image, so despite the fact that we designed the app to record the GPS location, the majority of the images did not have this information recorded. This happened because when the device is inside a building, the location information cannot update, so many images that reported the GPS location were either missing this information or it was not accurate.

Wi-Fi access points can be used to detect the estimated latitude and longitude by sending a Wi-Fi fingerprint to the Google Geolocation API [79]. Using this method, it had been possible to detect the location of a larger number of images. Figure 4.18 shows an example of this comparison within images where the GPS location was available (a) and the location detected using Wi-Fi (b). The area shown is the building where the data collection started. From the images we can see that using this method, more location information has been obtained, but the accuracy is not enough to distinguish between indoors and outdoors, and the way of obtaining this information is limited through knowledge of the Wi-Fi access point information.



(a) **(b)**

Figure 4.18: Data plotted on Google maps for each of the images with GPS (a) and Wi-Fi location (b).

4.4.2 Acquisition Process

The acquisition process describes all the variables that influence the moment in which the user is interacting with the camera and taking pictures. The two main contribution in this process come from both the user itself and the acquisition sensor which are the smartphone and the SLR cameras. As described in Chapter 2, user characteristics are distinguished between static and dynamic in the ISO/IEC 29794-5 Biometric Sample Quality – Part 5: Face image data Technical Report (TR) [7].

Characteristics and facial expressions of the users had been calculated automatically using the Neurotechnology SDK VeriLook 10.0 [80]. The user's pose towards the camera had been calculated following the methodology presented in [81] by Asthana et al., while camera characteristics had been extracted from the image metadata.

4.4.2.1 Static characteristics of users

From the ISO/IEC 29794-5 TR, the static characteristics of the subjects include anatomical characteristics, ethnicity and non-permanent characteristics as heavy make-up or glasses. As static characteristics for the users, we considered demographics, in particular: sex, age, ethnicity, completed education level, operating system used, and previous experience as the background history of the user. In addition, we also considered some non-permanent characteristics.

Figure 4.19: A user that weared glasses in one session and remove them for a subsequent one (left-hand side images) and a participant wearing photochromic lenses during the data collection (right-hand side)

One characteristic is whether the user is wearing glasses. There were 17 participants that wore glasses, but 7 decided to wear them only for a subsection of the session, as shown in Figure 4.19. In particular, there are two participants that wore glasses across all three data collection sessions that comprised photochromic lenses, causing the lenses to become darker when moving from outdoors to indoors. We used Neurotechnology VeriLook 10.0 to automatically detect glasses and dark glasses in the images, and we crossed check with visual examination.

Figure 4.20: A participant not wearing and wearing make up in two different sessions.

Included as non-permanent characteristics was also the presence of heavy make-up. In Figure 4.20, the participant presented make-up in a subsection of the sessions.

Male participants can present the situation where they have removed or grown facial hair between sessions (Figure 4.21). Neurotechnology had been used to automatically detect beard and a visual examination had been carried out to confirm the results.

Figure 4.21: Examples of a single participant with and without facial hair.

4.4.2.2 Dynamic characteristics of the users

One of the main user interaction features is how they interact with the acquiring device. Out of the 53 participants, one subject decided to use the smartphone in a landscape orientation for the second and third session as shown in Figure 4.22.

Figure 4.22: An image from the participant that decided to use the smartphone in the landscape orientation.

The dynamic characteristics described in the ISO/IEC TR 29794-5 relate to the way the user poses in front of the camera. To detect the user pose, we extracted the facial features of the image as described in the study proposed by Asthana et al. [81].

All the participants were given instruction to take the images on the smartphone for the purpose of facial verification but their perception of how to present their face was completely subjective. An example of a range of user's poses is given in Figure 4.23. Some

participants decided, for instance, to take the image from below, because this is how they would normally use their device, as they explained to the operator.

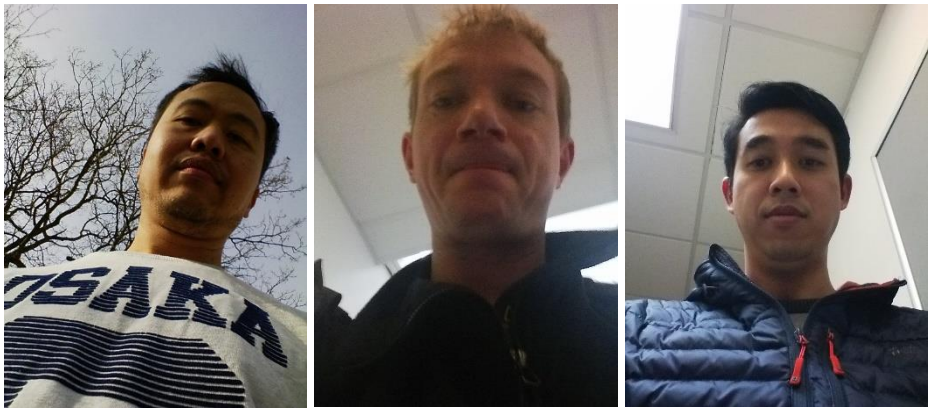


Figure 4.23: Examples of participant that took the images as they would do to unlock their device before using it.

Other subjects presented different facial angles. The facial angles for each image have been calculated in terms of pitch (nod), yaw (bobble) and roll (tilt) rotation of the face. The Discriminative Response Map Fitting MATLAB Code (DRMF 2013) [81] has been used to detect 66 landmark points on the users' faces and the estimations of the head pose.

In Figure 4.24 there is an example of a user that took images from different angles, expressed in degrees, and the respective facial landmarks.

Pitch: -7.27	**Pitch: -5.81**	**Pitch: -12.83**
Yaw: -12.98	**Yaw: 21.06**	**Yaw: 1.66**
Roll: 0.07	**Roll: -5.28**	**Roll: 0.55**

Figure 4.24: Images taken in different angles from the same participant: from the left, right and top.

Some participants were distracted by the surroundings and did not centre their facial images to the camera. This happened for examples in the images shown in Figure 4.25.

Figure 4.25: Examples of images taken from the participants.

A dynamic characteristic is also the reaction that users might have when taking a picture where they may have closed their eyes (Figure 4.26).

Figure 4.26: Users that closed their eyes during the acquisition of the facial image.

The Neurotechnology VeriLook 10.0 allowed the determination as to whether the users had open or close eyes in images with a binary outcome of 0 for closed and 1 for opened. Facial expressions are also the way that a participant interacts with the device during the acquisition process. In Figure 4.27 there is an example of a range of facial expressions made by a single participant during the data collection. The Neurotechnology VeriLook 10.0 algorithm has also been used to estimate the facial expressions and their relative accuracy scores (in the range from 0 to 100) as a percentage of confidence that the user exhibits an expression in a captured image. The facial expression detected were:

- anger
- disgust
- fear
- happiness
- neutral
- sadness
- surprise

Figure 4.27: Different facial expressions present by one of the participants.

4.4.2.3 Users experience and opinions

Another set of variables that was assessed is the user's opinion and experience during the data collection. We divided users into groups following data collection per category of questions according to their responses. In this section we report the results that we analysed from the questionnaires completed by the users.

From the chart shown in Figure 4.28, it is possible to see that overall the level of comfort that the participants felt while presenting their biometrics when no one was present during the data collection ranges between 4.3 to 4.45 on the Likert Scale.

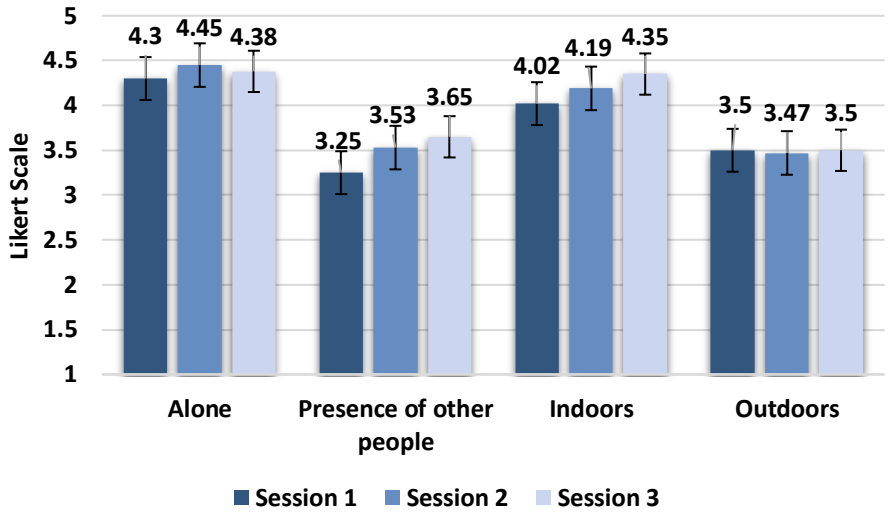


Figure 4.28: Mean Likert values describing the level of comfort that the participants had while taking the images in presence of other people, and in indoors or outdoors locations.

It can also be observed that the comfort expressed by the users increases within the three sessions for both the situations in which the participants had to take a facial image in front of other people, and in a situation where images were taken indoors.

In contrast, when the participants were taking images in outdoors locations, their level of comfort remained stable at around 3.5 across all the sessions. This is probably due to the variability in terms of weather conditions and time of the day that were not present in the other scenarios.

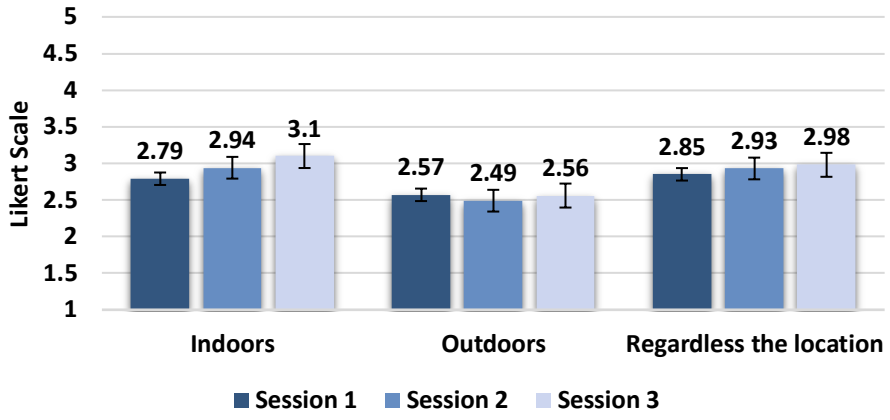


Figure 4.29: Good sample presentation according to the user perceptions depending on the location expressed as the mean of the Likert values.

When asked to compare the location types, the participants selected a neutral response overall when asked if they believed they provided a good sample for verification, regardless of the location in which the facial images were taken (Figure 4.29). Indoors presented an increasing score on the Likert Scale, from 2.79 to 3.1 from the beginning to the end of the data collection. The scores recorded when asking participants' opinions regarding images taken outdoors remain stable across the three sessions, at between 2.49 and 2.56.

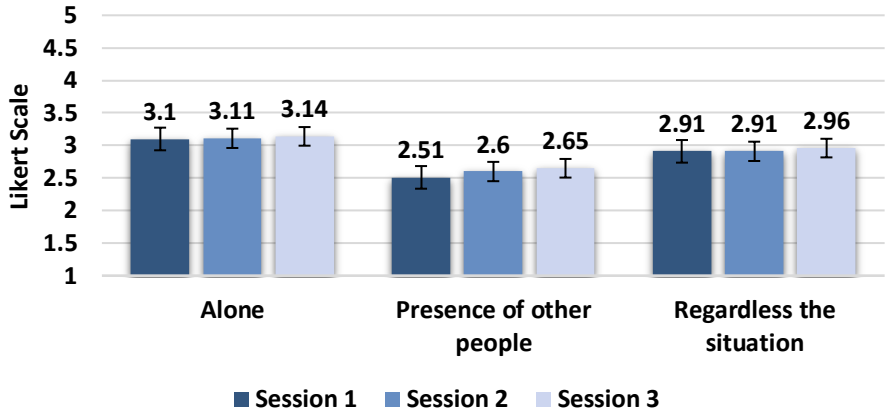


Figure 4.30: Good sample presentation according to the user perceptions depending on the location expressed as the mean of the Likert values.

When asked a similar comparison related to the different situations in which the biometric sample presentation had been made in presence or not of other people, the participants overall selected a neutral score around 2.91 that became 2.96 in the last

session (Figure 4.30). While taking images when alone, the participant reported a neutral score of around 3.1, in presence of other people they reported a score of 2.51 that increased gradually within the sessions until reaching 2.65 in Session 3.

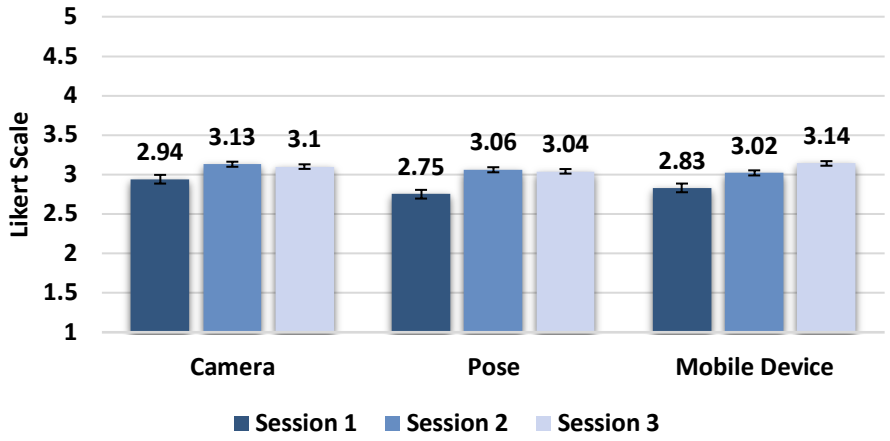


Figure 4.31: Mean Likert values describing the easiness to place the camera for the acquisition, to pose or to use the system on a mobile device.

The following set of questions were related to the collection of facial images in the context of mobile devices. The results as shown in Figure 4.31. Participants indicated a level of easiness of using a mobile device to collect an image with a smartphone that increased from 2.83 to 3.02 reaching 3.14 in the last session. Encouraging results can also be seen as to their opinions on the placing of the device for image capture. This result varied from 2.91 for the first session to approximately 3.1 in the following sessions. Similarly, their opinion on the easiness of presenting a suitable pose towards the camera changed from 2.75 to approximately 3 in the following sessions.

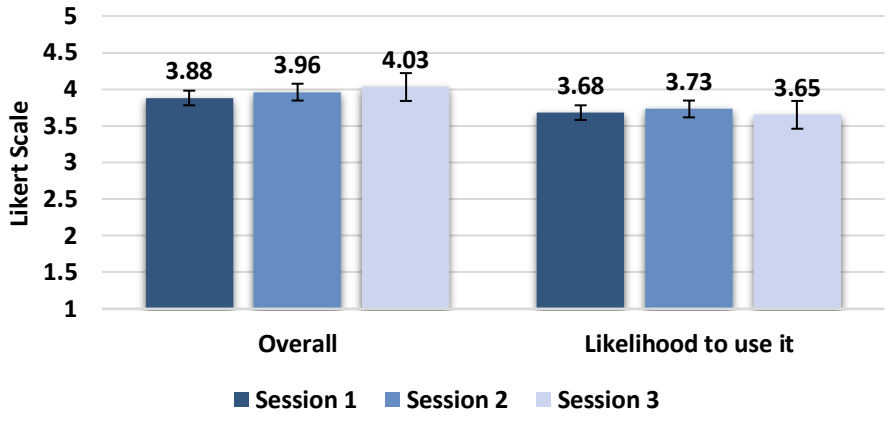


Figure 4.32: Mean values describing the overall experience and likelihood to use it.

Participants rated their overall experience between 3.88 and 4.03 with a slight increase from the first to the last session (Figure 4.32).

4.4.2.4 Static characteristics of the capturing device

According to the ISO/IEC 29794-5 TR, the camera can also be classified as both static and dynamic. For the static characteristics we collected our information from each image. We examined the Exif (Exchangeable image file format) information from each image to establish the variation in capture characteristics. Recent phones allow the owner to access, personalise and modify specific characteristics of the frontal camera. With the Nexus 5 that was not possible as the focus was set to automatic.

The main camera settings that give control over quality are the aperture, camera ISO and shutter speed [82]. Aperture is the size of the hole behind the lens that controls the quantity of light that enters the camera sensor and consequentially regulate the degree of exposure to light. In our experiment, it had a fixed value of 2.9 throughout across all the images taken with both the smartphone camera and the SLR.

Shutter speed is the length of time the camera shutter opens when taking the image. Adjusting the shutter speed allows the control of how moving subjects are recorded. The SLR camera was fixed in position with a tripod, and the shutter speed was set at 1/60 which is suitable for recording images of non-moving subjects. When taking selfies with the smartphone, not only are the subjects moving but also the camera can take a different position depending on how the user is holding the device. It becomes hard to differentiate these types of movements and, for this reason, the settings that we decided to consider in our analysis is the variation in ISO that measures the sensitivity of the camera sensor. The SLR had a fixed value set to 400, while the smartphone camera ISO variates between 100 and 2000.

4.4.2.5 Dynamic characteristics of the capturing device

Dynamic characteristics of the camera were assessed using the accelerometer data. The three-axes acceleration forces were combined to detect whether there was movement during the capture by comparing a non-moving capture. The image shown in Figure 4.34 shows an example of how movement was detected and calculated.

Time-stamp and accelerometer data was recorded with a sampling frequency of 10 Hz for each image collected. We pre-processed and segmented the signal using three window sizes of 1, 3, and 5 seconds before and after each image was taken. We then extracted features that could be used to analyse user and smartphone movements. First, we calculated the magnitude for each image using the below formula, where M is the resultant Magnitude, and A_x, A_y, and A_z are the directional accelerations on each smartphone's axes.

$$M = \sqrt{{A_x}^2 + {A_y}^2 + {A_z}^2}$$

Magnitude can provide information about gait movements [83], whether the users were walking or moving when capturing the images. When the signal does not present any variation, it means that the user stopped walking and is not moving or is performing

minimal movements with the smartphone before the capture process. An example of the two situations is presented in Figure 4.33 and Figure 4.34. From the three different selected time windows, we observed that the overall trend of the magnitude presented peak-to-peak amplitudes within the range of ±3 m/s^2. We empirically selected two thresholds: 1.5 m/s^2, and 2 m/s^2 to differentiate between movement and non-movement. We considered as magnitude features the number of peaks presented in the signals and the amplitude of their variations when they were above the three selected thresholds.

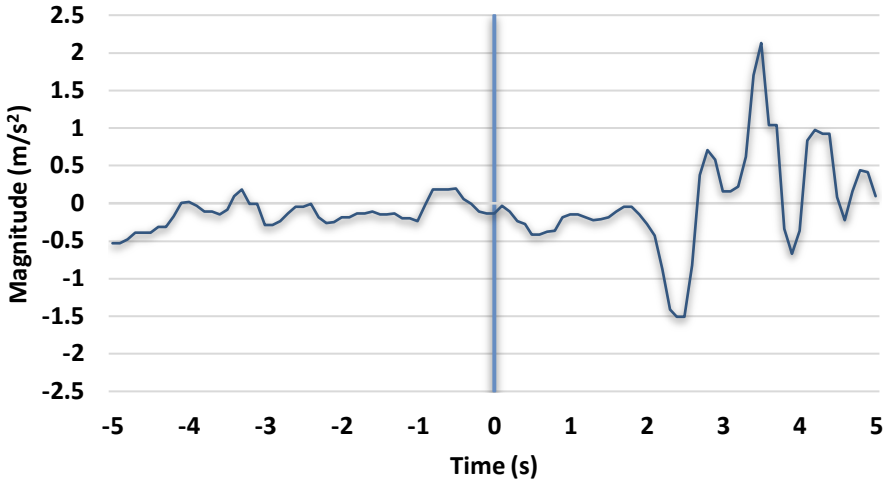


Figure 4.33: Gait movements in a 5-second window before and after an image was taken. The graphs shown a user that was still moving or had not stopped completely before taking an image.

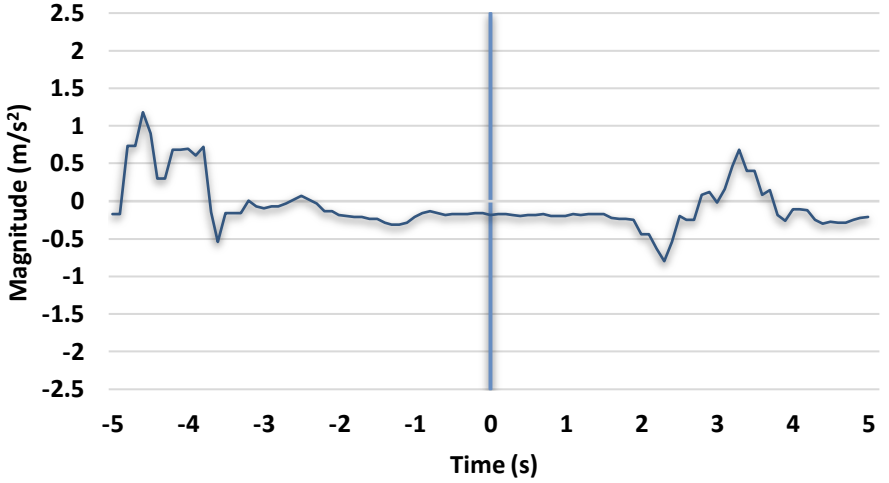


Figure 4.34: Gait signal where a user had stopped or recorded little movement while taking an image

4.4.3 Verification Process

The verification process consists of the quality assessment and verification matching score undertaken by the biometric system to authenticate users on the device. A first step

in the verification process has been made by detecting the facial area of the images, that have been subsequentially assessed to obtain the quality metrics and the verification matching scores.

Facial Image Quality (FIQ) metrics was initially calculated using the Aware PreFace [84], but it was not possible to assess the majority of the facial images in our database since the software could not calculate the metrics for facial images that were not conformed with the passport Standard. Therefore, we calculated FIQ metrics considering the indication in the ISO/IEC 29794-5 TR. Verification decision and matching scores were calculated using a state-of-the-art commercial verification system, Neurotechnology VeriLook 10.0, and an open source verification system, Face Recognition [85], built with deep learning. Only genuine comparisons were considered as for the scenario of biometric verification.

4.4.3.1 Facial Image Quality (FIQ)

To assess the facial quality of the selfies acquired during the data collection, we followed the recommendations of ISO/IEC 29794-5 Technical Report (TR). The TR considers several Facial Image Quality (FIQ) metrics. Out of the several FIQ metrics considered in the TR, we selected five commonly used metrics as illustrated in Chapter 2 to describe quality features.

- Image Brightness refers to the overall lightness or darkness of the image.
- The Image Contrast helps to understand the difference in brightness between the user and the background of the image.
- The Global Contrast Factor (GCF) determines the richness of contrast in details perceived in an image. The higher the GCF, the more detailed the image.
- Image Blur quantifies the sharpness of an image.
- Exposure quantifies the distribution of the light in an image.

Below is a description on how each FIQ metric was calculated.

4.4.3.2 Image Brightness

Image Brightness is a measure of pixels intensities of an image. As defined in the TR, image brightness can be represented by the mean of the intensity values h_i, where $i \in \{0, \ldots, N\}$. The mean of the histogram $\bar{h}$ can be represented by the formula:

$$\bar{h} = \frac{1}{N+1}\sum_{i=0}^{N} h_i$$

where h is the intensity value of each pixel, and N is the maximum possible intensity value. Values have been normalised from 0 to 5, where 0 is the lowest level of brightness and 5 is the highest. An example of both extremes can be observed in Figure 4.35.

(a) (b) (c) (d)

Figure 4.35: Example of brightness. Image (a) has the lowest value recorded for brightness (B = 0), but the algorithm still detected the face from the original image (b). The facial area (c) extracted from the original image (d) has a high brightness value (B = 4.51).

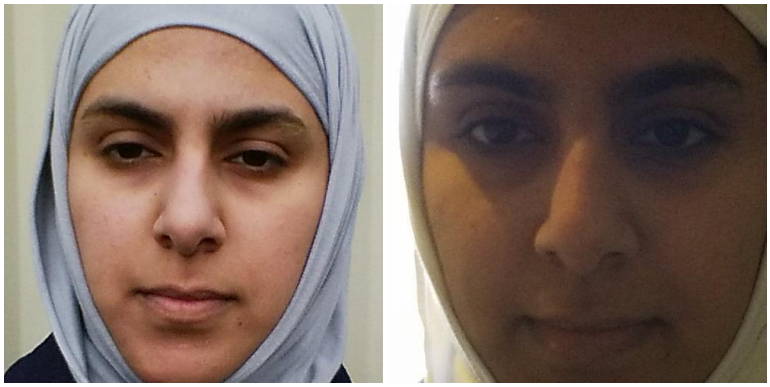
4.4.3.3 Image Contrast

Image Contrast is the difference in luminance of the object in the image. There are different ways to define Image Contrast.

We chose to calculate it from the histogram of the facial region using the following formula:

$$C = \sqrt{\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}(I(x,y)-\mu)^2}{MN}}$$

where $I(x,y)$ is the facial region of size MxN, and μ represents the mean intensity value of the facial region. Values had been normalised to range from 0 (low contrast) to 5 (high contrast). An example can be seen in Figure 4.36.



(a) Contrast = 4.29 **(b) Contrast = 1.96**

Figure 4.36: Examples of two images with high and low contrast level.

4.4.3.4 Global Contrast Factor (GCF)

The Global Contrast Factor (GCF) is described in the TR as the sum of the average local contrasts for different resolutions multiplied by a weighting factor. We calculated the GCF following the methodology presented by Matkovic et al. [86]. The local contrast is calculated at the resolution of the original image as the average difference between the

intensity of neighbouring pixels. Then the local contrast is calculated for decreasing resolutions that are obtained by combining four original pixels into one superpixel, reducing the image width and height to half of the original. This process has been calculated across R iterations. The global contrast is then calculated as a weighted average of local contrasts:

$$GCF = \sum_{k=1}^{R} w_k C_k$$

where C_k is the local contrast for R the number of resolutions considered, and w_k is the weighting factor. The authors defined the optimum approximation for the weighting factor over R resolution levels as:

$$w_k = \left(-0.406385\frac{k}{R} + 0.334573\right)\frac{k}{R} + 0.0877526$$

Where w_k ranges from 1 to the number of resolutions (R) of the image considered.

Contrast values have been normalised to have a scale from 0 to 5 where 0 is the lowest value and 5 the highest, and an example can be seen in Figure 4.37.

(a) GCF = 4.81 **(b) GCF = 0.56**

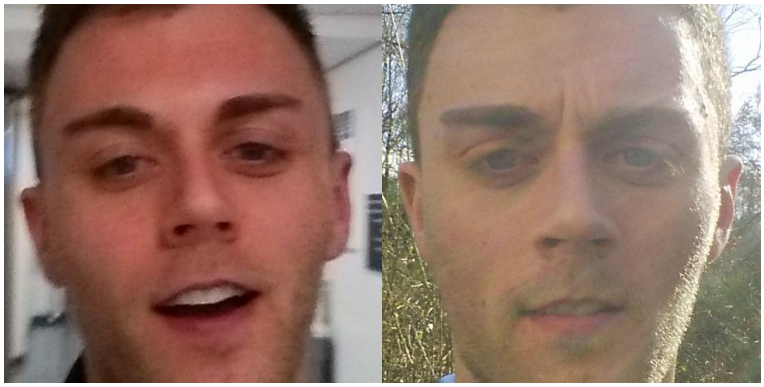
Figure 4.37: Examples of high and low level of Global Contrast Factor.

4.4.3.5 Image Blur

To quantify blur, we applied the work presented by Crete et al. [87]. Their methodology allows the determination of a no-reference perceptual blurriness of an image by selecting the maximum blur among the vertical direction $blur_{ver}$, and the blur across the horizontal axis $blur_{hor}$.

$$Blur = Max(blur_{ver}, blur_{hor})$$

The metric produces a result between 0 and 1, where 0 is the sharp image and 1 is the worst quality. To make the results comparable to previous metrics, we normalised the scale to be in a range from 0 to 5 where 0 is the sharper and 5 is the more blurred. An example can be seen in Figure 4.38.



(b) Blurriness = 4.21 **(a) Blurriness = 0.8**

Figure 4.38: Example of a really blurred image and a sharp one taken from the same participant during the data acquisition.

4.4.3.6 Exposure

Exposure can be characterised by the degree of the distribution of image pixels over the grayscale. As defined in the TR, exposure can be calculated as a statistical measure of the pixel intensity distribution, such as entropy [78].

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where p_i is the histogram of the intensity level for the N possible intensity levels. As with the other FIQ metrics, we normalised the scale to be in a range from 0 to 5 where 0 is the less exposure and 5 is the high exposure. An example can be seen in Figure 4.39.

(a) Entropy = 5 **(b) Entropy = 2.86**

Figure 4.39: Examples of high and low entropy in the images.

4.4.3.7 Biometric matching score: the enrolment

We considered four enrolment scenarios. The first enrolment (E1) included five images taken using the SLR camera under Scenario 1 as previously explained. The second type of enrolment (E2) used images taken with the smartphone camera (Scenario 2). These first two types of enrolment enabled a comparison of different types of cameras under the

same ideal enrolment conditions, with the hypothesis that the SLR would take higher quality images and that it would be resulted in higher verification scores.

The other two types of enrolment replicate real-life situations where the participant is using facial verification for the first time and is required to enrol on the smartphone. We selected the first five images taken indoors from Scenario 3 for the third enrolment (E3) and the first five images taken outdoors (Scenario 4) (E4).

We decided to exclude a combination of images taken indoors and outdoors because we assumed that it would be unlikely that a user will change their location from indoors to outdoors (or vice versa) during enrolment. Table 4.6 summarises the four type of enrolment and their specifics.

Table 4.6: Enrolment scenarios.

Enrolment type	Capturing system	Scenario	Description
E1	SLR	Static, fixed camera	Indoors
E2	Smartphone	Static, moving camera	Indoors
E3	Smartphone	Unconstrained	Indoors
E4	Smartphone	Unconstrained	Outdoors

Once all the images had been selected for the enrolment, we then considered all remaining images from that participant for verification to replicate the typical use of a mobile device.

4.4.3.8 Neurotechnology VeriLook 10.0

Neurotechnology VeriLook 10.0 has its own face detection algorithm. We used the VeriLook SDK to perform biometric verification, recording a Failure to Detect (FTD) when the algorithm could not recognise a face within an image. We calculated a biometric score (BS) as the mean of the one-to-one comparisons of the facial verification image against all five enrolment images and a biometric outcome (BO) as either “Successful” or “Failed” depending on the majority between the five comparisons.

The biometric scores that were output from the system were given on a scale from 0 to 2822, therefore we decided to normalise them to a scale from 0 to 1 to be able to compare the system with other biometric algorithms. For the BO, we kept the default matching threshold that was set to 48 (0.02 on the normalised scale), although we considered it quite permissive.

4.4.3.9 Face_recognition

Face_recognition is an open source algorithm implemented in Python using dlib [88] and OpenFace [89]. Firstly, it uses the Histogram of Oriented Gradients (HOG) method for face detection [90]. Then, the system estimates 68 facial landmark points using the method presented by Kazemi and Sullivan in [91]. With Face_recognition we calculated the encoding for each of the 5 images considered as enrolment and we calculate the matching scores with the one-to-one comparisons with all the image in the verification dataset.

From each comparison, a BS was recorded with a continuous range from 0 to 1 that expressed the level of similarity between the two images. The system presented also a BO that indicated whether the verification had been “succeeded” or “failed”, hence whether the comparison scores were below the default tolerance cut-off of 0.6.

4.5 Summary of the variables considered

Each of the aspects described in this Chapter had been considered for the analysis of user interaction and image quality have on the face verification system when implemented on a smartphone. A summary of all the type of data is presented in Figure 3.6

The analysis has been carried out considering the different passages of the verification process. In the following Chapter we are going to present the aspect that influence face detection when performed in different environmental locations and considering the static and dynamic characteristics of both the camera and the users.

Chapter 6 and Chapter 7 will describe the aspects analysed for respectively image quality and verification matching scores. The observations had been made also with statistical analysis to observe significant variations within the variables. The conclusions observed from this analysis will be used to provide an overall perspective for the issue of mobile face recognition and aimed to create a guideline for developers and future research when designing face verification system to be implemented on mobile devices.

Table 4.7: A summary of all the variables taken into consideration in the analysis.

Type of variables	Name	Description	Scale
Environment	Location types	Experimental lab	0-1
		Indoors	0-1
		Outdoors	0-1
	Background analysis	Background Complexity (Texture metrics)	0-5 (scale)
		Other faces detected in the image	0-1
	Location metadata	GPS or Wi-Fi	Discarded
Acquisition process - User	Static characteristics	Sex	0-1
		Age	18-46 (scale)
		Handedness	Discarded, no enough data for comparison
		Ethnicity	6 groups
		Completed education level	5 groups
		Operating systems used	3 groups
		Previous biometrics experience	0-1
		Glasses	0-1
		Make-up	0-1
		Beard	0-1
	Dynamic characteristics	Landscape mode	Discarded, only 1 user used it
		Facial angles	Facial features locations
		Blink	0-1
		Mouth open	
		Facial expressions	7 types
		User's movement	8 activities
Acquisition process - Camera	Type of camera	SLR or smartphone	0-1
	Static characteristics	ISO	5 groups
		Light Value	
	Dynamic characteristics	Camera movements	Accelerometer features
Verification process	Face detection	FTD	0-1
	Quality assessment	Brightness	0-5
		Contrast	0-5
		GCF	0-5
		Blurriness	0-5
		Entropy	0-5
	Biometric performance	Enrolment 1	Biometric scores
			Biometric outcomes
		Enrolment 2	Biometric scores
			Biometric outcomes
		Enrolment 3	Biometric scores
			Biometric outcomes
		Enrolment 4	Biometric scores
			Biometric outcomes

The Verification Process: Face Detection

5.1 introduction

The first step for the verification process is to locate and segment the facial area in the sample image. Many algorithms in state of the art have been used, enhanced and studied to perform this task. The work presented in this chapter aims to identify the variables that influence mobile face detection and provide a description of their relationship and relevance when considering the mobile context. Two main algorithms were used to perform this task with a further two provided with the face verification systems used to assess biometric performance. The comparison between multiple algorithms was considered to ensure that the observations made have a universal application and that they are not valid for a specific context.

One of the methods used in this analysis was proposed by Viola and Jones [77], a well-known object-detection algorithm principally used for the detection of a face for facial recognition. The method utilises classifiers in a cascade to ensure high performance while reducing the computational time. Despite being released in 2001, it is still widely used in many applications, even in the context of mobile devices [92] [93]. The model presented by Zhu and Ramanan [94] was used as a second facial detection algorithm because its tree-based method is particularly effective in detecting faces within images taken in the "wild". The remaining two algorithms considered are provided as part of the facial recognition systems employed in this study: Neurotechnology VeriLook 10.0 SDK [80] and the dlib Face_recognition system [85]; the latter uses the Histogram of Oriented Gradients (HOG) method whilst the former is a black box proprietary system.

This Chapter presents the analysis of several variables that can influence the outcome of a face detection system when used by unsupervised subjects in a mobile context. The results will be presented for the face detection outcomes obtained under different conditions: across the different scenarios considered for the data collection, the captured environment, the influence that the user's interaction has on the detection system and the quality assessment performed on those images that presented a Failure to Detect (FTD).

5.2 Face detection in each scenario

As described in Section 4.2.3, there are four different scenarios that were considered for data collection. Scenario 1 considers images taken with a Single Lens Reflex (SLR) camera at a fixed distance while the users were seated on a chair in the experimental room. Scenario 2 also took place in the experimental room, but the participants took the facial images using a smartphone camera while they were seated on a chair. Scenarios 3

and 4 were images taken by the participants with a smartphone camera while they were moving between locations comprised of indoors and outdoors environments.

All the face detection algorithms considered were able to detect a facial area for the 318 images presented in Scenario 1. However, the images that were taken with the smartphone camera (Scenarios 2 to 4) reported different results depending on the algorithm used. The frequency and percentages of FTDs of each method are indicated in Table 5.1.

Table 5.1: Frequency and percentage of FTD occurred for each method.

FTD	Viola-Jones	Tree-based method	VeriLook 10.0	Face_recognition
Frequency	289	1	601	395
Percentage	3.2%	0.01%	6.6%	4.3%

As seen from the Table, the tree-based method performed almost perfectly, with only one image out of the 9,103 taken with the smartphone that could not be detected. From the results, it can be seen that each algorithm performed adequately, presenting percentages of FTDs that could be considered acceptable within a repeated sample scenario. Interestingly, not all the FTDs occurred for the same images; for instance, some images could present an FTD when Face_recognition is applied but not when using the Viola-Jones method. An example is given in Figure 5.1: the image shown is the only image that the Tree-based method did not detect, but the face detection algorithm used by VeriLook 10.0 was able to detect the facial area from the picture.

Figure 5.1: The image that was not detected by the Tree-based method.

Although the Tree-Based method seems to perform perfectly with images taken in the unconstrained environment, there are several issues that need to be taken into consideration. For example, the time required to detect the facial area of an image is roughly 50 seconds, while the time needed to detect the same image with Viola-Jones is close to 0.5 seconds. When considering the application of face detection for mobile verification, the time required to authenticate the user on the device is one of the main

acceptability issues that biometric technology has to overcome (as mentioned in Section 3.2). The balance between timings and performance is fundamental in this context.

A second problem that needs to be considered is whether the images that have been detected are truly facial areas. For instance, when considering the Viola-Jones method, after a visual examination it was discovered that within 107 images the algorithm classified a "detected" face not aligned to an actual facial area. In the database collected for this study, 1.23% of the images detected by Viola-Jones were not facial areas. An example of this is shown in Figure 5.2. The incorrectly detected images were not considered as FTDs since the detection system labelled them as "detected images", but quality assessment was carried out to identify the variations that could have led to a misplace of the facial area.

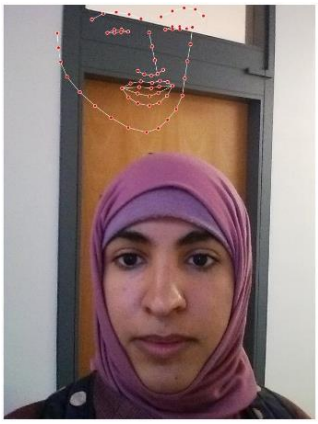


Figure 5.2: An example of an image where the face had been detected but did not correspond to the user's facial area.

The frequency of FTDs was investigated for the whole database under different aspects. One of the considerations made was to check whether there are differences in FTDs in the occurrences between the time separated donation sessions. These results are shown in Figure 5.3. The Tree-Based method was not included in the analysis since it did not report a sufficient number of FTD to enable a comparison between the images.

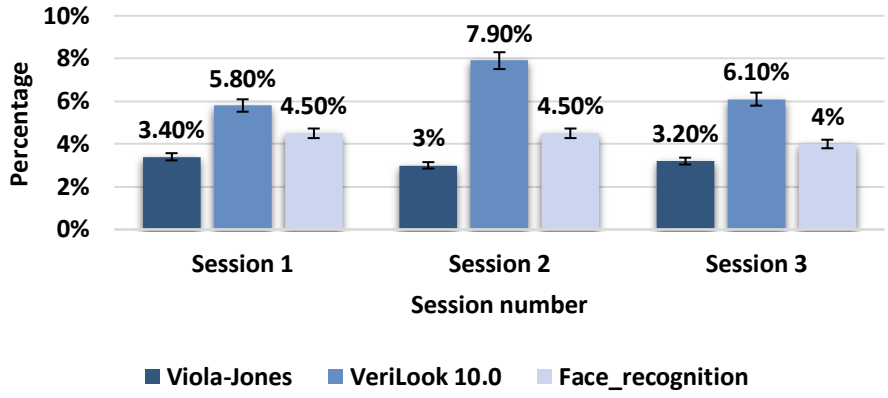


Figure 5.3: Percentage of FTD across sessions.

From the bar chart it is possible to note that the FTD frequency is not as would be expected. The algorithms' outcomes reported three different trends that leads to the hypothesis that the test subjects' learnability of taking images with a given smartphone did not affect the face detection outcome. It would be expected, in fact, that the number of FTDs that occur decreases over the sessions, but this is valid only for the Face_recognition algorithm. Possibly, the time between sessions within the collected dataset is insufficient to reach a definite conclusion on this aspect.

Furthermore, the participants reported on multiple occasions to the operator that they found it more difficult to take good outdoor sample images when the weather condition were adverse. This could be a possible explanation of the unpredictable trend across the sessions. This highlights the importance of assessing the difference between environmental conditions.

5.3 Environment analysis

This section presents an analysis of face detection outcomes depending on the location types in which the images were taken. These results could help to understand the different FTDs recorded in each scenario and could provide information on the surroundings in which the image was taken. The analysis proposed in this work assessed not only the differences according to the location type, but also the information that could be obtained about an image from its background throughout a texture analysis. Furthermore, the presence of other subjects within an image was investigated to check whether this "noise" in the background could have an impact on the detection of the user's facial area, which should contain the only face that needs to be verified.

Overall, the percentages of FTDs are higher for the images taken outdoors. The face detection outcomes for each scenario are shown in Table 5.2. There are only a few images (one when using the Viola-Jones method and two when VeriLook 10.0 was used) that recorded an FTD when the images were taken using the smartphone camera by the participants in the experimental laboratory.

Table 5.2: Face detection according to the different scenarios.

Scenario	Viola-Jones	VeriLook 10.0	Face_verification
Scenario 1	0.0%	0.0%	0.0%
Scenario 2	0.4%	0.8%	0.0%
Scenario 3	2.4%	5.7%	3.3%
Scenario 4	3.8%	7.4%	5.2%

A Chi-square test was performed to check whether there were significant associations between the outcome of the algorithms and the different scenarios. The location types have a statistically significant relationship with the Viola-Jones ($\chi^2(2) = 19.38$, $p < 0.001$), VeriLook 10.0 ($\chi^2(2) = 25.02$, $p < 0.001$) and Face_recognition ($\chi^2(2) = 30.23$, $p < 0.001$) methods. A Cramer's V test was performed as a post-test to determine the strengths of

the association between the variables. Cramer's V value varies between 0 and 1, where 1 indicates a strong association. For each algorithm the constant value was around 0.05, indicating that despite having a significant association between the detection outcome and the different scenarios, the strength of the association is weak.

A similar test was also performed to check for a possible significant relationship between the outcomes of the detection systems and the locations on maps A, B, and C used during the data collection. The test did not report any significant result, meaning that the FTD recorded were not affected by the differences of locations within the maps used or the order of the locations in which the participants stopped to take the images. This outcome shows that the results have value in any situation and do not depend on those selected for the data collection.

5.3.1 Background complexity

A texture analysis was performed to understand the role that the background of facial images has on face detection approaches. The background texture was assessed not only to investigate the differences in terms of complexity of the background but also to understand how the surrounding environment varies between location types.

Each image in the database was segmented into different parts. An example of the segmentation of an image is shown in Figure 5.4. Every part of an image that does not include the user's facial area was considered as background, including hair and the clothes of the person as these can be considered an element of "noise" for the detection algorithm and possibly be mistaken as a face.

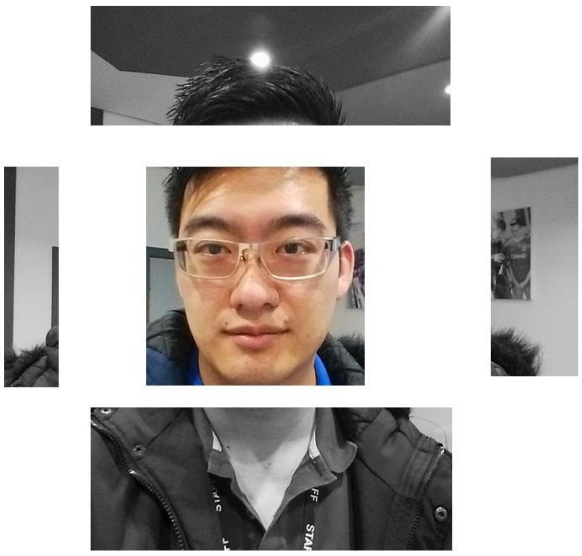


Figure 5.4: Example of segmentation of the image background.

The background complexity of all the images was calculated in grayscale considering three different metrics (that were described in Section 4.4.1.3): Texture Range, Texture

Standard Deviation and Texture Entropy. Each value was calculated locally for each segmented part and combined as an average for the whole background. To enable a comparison between the three metrics, the values obtained were normalised to a range between 0 and 5. The variation for each metric across the database is explored in the following subsections.

5.3.1.1 Texture Range

The Texture Range gives us information of the local variability of each segmented background image. Variation within the different environmental locations can be seen in Table 5.3.

Table 5.3: Variation of local Texture Range values across the different location types.

Image background section	Experimental lab			Indoors			Outdoors		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Right	0.04	1.28	0.22	0.03	1.85	0.51	0.00	5.00	1.45
Left	0.05	0.88	0.18	0.00	1.90	0.50	0.00	5.00	1.42
Bottom	0.28	2.39	0.94	0.05	3.75	0.74	0.00	5.00	1.15
Top	0.19	1.06	0.49	0.17	2.13	0.59	0.00	5.00	1.68
Overall	0.10	1.18	0.40	0.04	1.81	0.61	0.00	5.00	1.78

Considering the four segmented parts of the background separately, it can be noted that the area selected below the face reported higher values, except from the images that were taken outdoors. It appears that the variations in Texture Range of the surrounding environment are higher in this location than in those corresponding to participant's clothing segments. While the Texture Range for the background of images taken in the experimental laboratory vary between 0 and 2.50, indicating a smoother background, and the background for indoors images presented variations from 0 to 3.75, the images taken outdoors have variations that cover the whole range from 0 to 5. The mean values for the indoors locations (including the experimental laboratory) present mean variations within the Texture Range that change between 0 and 1. Outdoors locations present higher variation resulting in a mean interval from 1 to 2.

The chart shown in Figure 5.5 shows the mean values for each location type. The differences observed between the location types could be used to obtain information on where the biometric presentation is occurring. For instance, the background section containing the user's clothing could be removed to enable a distinction between the location types with an adequate selected threshold. Depending on the variations of the other texture metrics, it could be possible to combine all the information and estimate if an image has been taken indoors or outdoors.

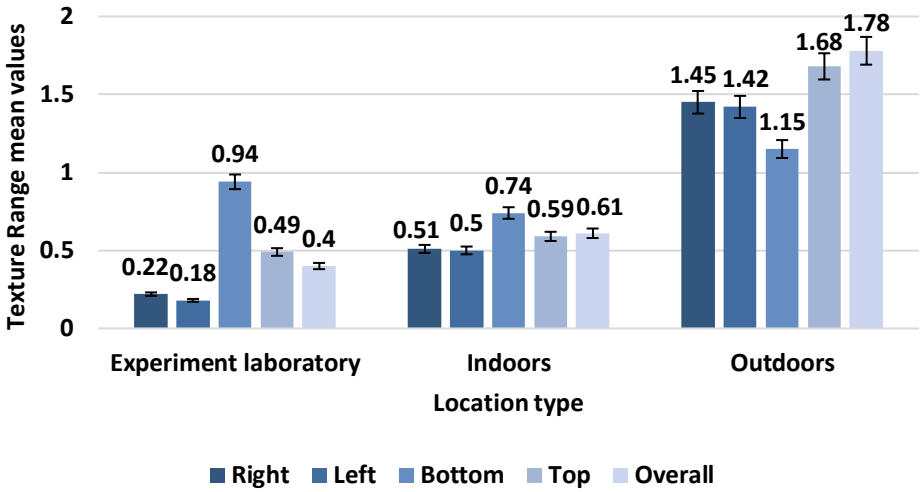


Figure 5.5: Mean of Texture Range across location types.

5.3.1.2 Texture Standard Deviation

Similarly, the Texture Standard Deviation was calculated as the local standard deviation within a 3-by-3 window around the considered pixel. The texture metric was calculated for each segmented part of the background. Table 5.4 describes the variations across different location types.

Table 5.4: Variation of Texture Standard Deviation across the different location types.

Image background section	Experimental lab			Indoors			Outdoors		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Right	0.05	1.25	0.23	0.04	2.02	0.52	0.00	5.00	1.44
Left	0.06	0.86	0.19	0.00	1.94	0.51	0.00	5.00	1.43
Bottom	0.28	2.52	0.96	0.05	3.82	0.75	0.00	5.00	1.17
Top	0.19	0.99	0.46	0.16	1.97	0.56	0.00	5.00	1.61
Overall	0.11	1.16	0.40	0.05	1.86	0.62	0.00	5.00	1.78

The Texture Standard Deviation values have similar results as to those observed for the Texture Range variability. Higher values of Texture Standard Deviation correspond to a larger local standard deviation within the original image. When observing this texture metric, the minimum and maximum pixel intensity recorded do not vary from the mean pixel intensity values of the image. The segmented part that contains clothing is the background segment that reported higher values, except from when considering the images that were taken outdoors, where the average variability of the surroundings increases.

In Figure 5.6, the chart reports the differences between the mean values recorded by the Texture Standard Deviations according to each location types. As similarly observed for Texture Range, the values for indoors locations vary between 0 and 1, while the

outdoors variations are between 1 and 2. It appears that Texture Standard Deviation can also be used as information to distinguish between the three location types by removing the background section that includes the user's clothing selecting a threshold.

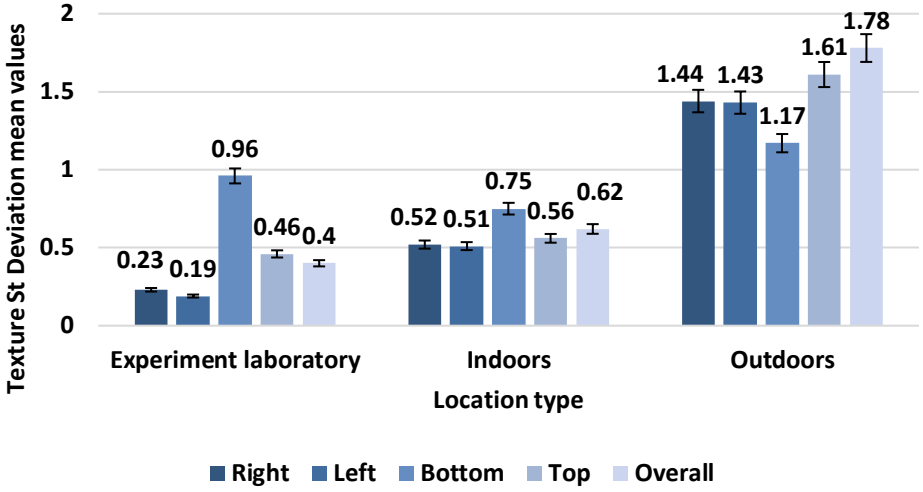


Figure 5.6: Mean of Texture Standard Deviation across location types.

5.3.1.3 Texture Entropy

The Texture Entropy describes the randomness of the image background. When compared to the previous texture metrics, Texture Entropy values seem to change more within the range across each location type. The variations can be observed according to the environmental locations in Table 5.5.

Table 5.5: Variation of Texture Entropy across the different location types.

Image background section	Experimental lab			Indoors			Outdoors		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Right	1.20	3.70	1.94	0.29	4.36	2.88	0.00	5.00	3.84
Left	1.31	3.50	1.97	0.01	4.45	2.90	0.00	5.00	3.84
Bottom	1.33	4.34	2.81	0.00	4.65	2.68	0.02	5.00	3.02
Top	2.22	3.79	2.95	1.04	4.60	3.18	0.00	5.00	3.60
Overall	1.17	3.33	2.05	0.81	4.38	2.74	0.00	5.00	3.60

The background segments recorded in the experimental room reported Texture Entropy values between 1 and 4 (or even higher when considering the section below the facial area), while the images that were taken indoors and outdoors change within the whole spectrum of values, from 0 to 5, so it is more difficult to estimate a distinction as could be achieve for the Texture Range and Texture Standard Deviation.

There are also fewer differences when considering each segmented part. The section above and below the facial area recorded higher mean Texture Entropy values in the experimental laboratory, but for the other two location types the differences are not so

evident. The clothing sections of the image follow the trend as for the Texture Range and Texture Standard Deviation of having less variations than the remained of the images when taken outdoors (Figure 5.7).

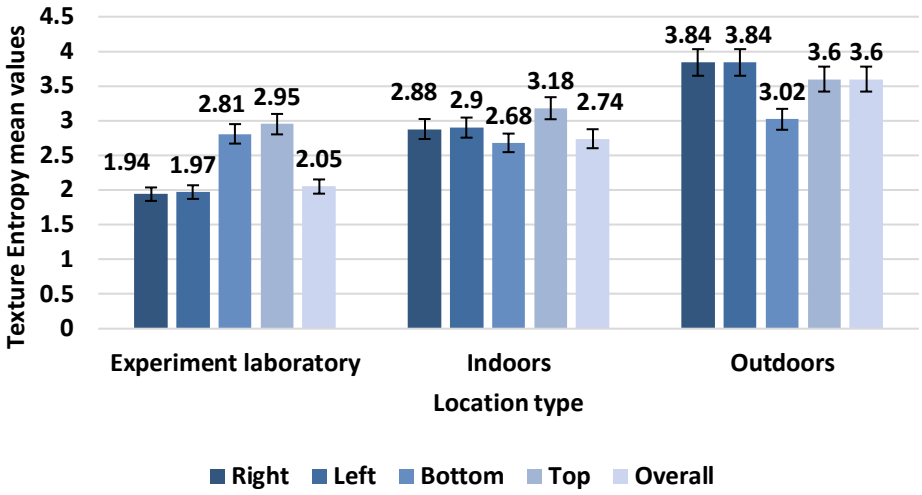


Figure 5.7: Mean of Texture Entropy values across location types.

5.3.1.4 The effect of the background on face detection

After analysing the variation that background texture metrics reported in the different location types, an investigation was carried out to understand if it is possible to establish the location type in which the image had been taken depending on the values reported from the background. Figure 5.8 gives an example on texture metrics in two extreme cases.

Texture Range = 0.22
Texture St Deviation = 0.24
Texture Entropy = 0.81

Texture Range = 4.13
Texture St Deviation = 4.09
Texture Entropy = 5

Figure 5.8: Examples of a low level of background texture (left hand-side) and high level of background texture (right hand-side).

A logistic regression model was designed where each of the texture metrics was considered as single contributors to establish whether an image was taken indoors or outdoors. We compared three logistic models to determine which of the three selected

texture metrics would be more significant in distinguishing the location in which the image was taken.

The first model considered Texture Range as a contributor:

$$logit(p) = -4.50 + 4.95 * \text{Texture Range}$$

The second model considered Texture Standard Deviation:

$$logit(p) = -4.49 + 4.88 * \text{Texture Standard Deviation}$$

Finally, the third considered as predictor the Texture Entropy:

$$logit(p) = -7.40 + 2.47 * \text{Texture Entropy}$$

Both the model that used Texture Range and Texture Standard Deviation explained between 52% (Cox & Snell R^2), and 71% (Nagelkerke R^2) of the variation and they were both able to correctly classify 87.5% of the cases. The model that considered as a contributor the Texture Entropy was only able to explain between 34% (Cox & Snell R^2) and 46% (Nagelkerke R^2) of the variation and to classify 78.4% of the cases correctly. These results indicate that Texture Range is the best metric describing the image background that can be employed to establish whether an image has been taken indoors or outdoors. Texture Standard Deviation can be alternatively used as a predictor to fulfil this purpose, while Texture Entropy did not show the same accuracy in completing this task, according to the logistic model. A summary of the logistic regression models for each contributor can be seen in Table 5.6.

Table 5.6: Statistical values for the logistic regression model predicting location types.

Background texture metric	B	S.E.	Wald test	df	p	Odds Ratio
Texture Range	4.95	0.11	1940.33	1	0.000	141.57
Texture St Deviation	4.88	0.11	1961.67	1	0.000	132.04
Texture Entropy	2.47	0.05	2142.19	1	0.000	11.86

Similarly, logistic regression was performed to ascertain the effect that the background of the image has on the face detection outcome. The model was designed to check whether it was possible to predict the detection of the facial area in the image by removing the background section that considers the clothing of the user and evaluating the image background complexity.

The first two models, designed with Texture Range and Texture Standard Deviation, respectively, did not result significant when considering Viola-Jones and Face_recognition as detection algorithms (Table 5.7). VeriLook 10.0 reported significant results for both the models. The reason for this difference could be explained by the nature of the two algorithms that do not allow a proper estimation for the detection of the facial area through the texture background metrics.

Table 5.7: Statistical values for the logistic regression model predicting Viola-Jones facial areas detection.

Viola-Jones	**Texture Range**	$\chi^2(1) = 0.003$, p = 0.954
	Texture St Deviation	$\chi^2(1) = 0.001$, p = 0.979
Face_recognition	**Texture Range**	$\chi^2(1) = 0.003$, p = 0.954
	Texture St Deviation	$\chi^2(1) = 0.001$, p = 0.979

The first model for VeriLook 10.0 considering Texture Range as a contributor was:

$$logit(p) = 2.79 - 0.126 * \text{Texture Range}$$

While the model for VeriLook 10.0 considering the predictor Texture Standard Deviation was:

$$logit(p) = 2.8 - 0.132 * \text{Texture Standard Deviation}$$

Despite the significant results, in both cases, the model could only explain 0.2% (Nagelkerke R^2) of the variation.

The regression model designed to estimate the detection outcome using Texture Entropy showed better outcomes and reported significant results for all the detection algorithms.

When considering Viola-Jones, the model explained between 3% (Nagelkerke R^2) of the variation:

$$logit(p) = 1.32 + 0.67 * \text{Texture Entropy}$$

For VeryLook 10.0 the model was able to explain 6% (Nagelkerke R^2) of the variation:

$$logit(p) = 1.76 + 0.27 * \text{Texture Entropy}$$

Finally, the model using Texture Entropy as a predictor for the detection outcome of Face_recognition was able to explain 2.7% (Nagelkerke R^2) of the variation:

$$logit(p) = 1.14 + \ 0.618 * \text{Texture Entropy}$$

It is possible to conclude that the analysis of the background texture reported results that explain the impact that the image background has on facial detection systems. While Texture Entropy was the metrics that reported less information to determine the location in which the image was taken, it resulted in being the metrics that suggest more significantly the prediction of facial area detection in an image. Results will improve when using a database with the same number of images detected and FTD. We can conclude that acknowledging information of the background could be useful both to detect whether the image had been taken indoors or outdoors and to predict the face detection outcome.

5.3.2 Other faces detected in the same image

It might happen that, in an image, there are other faces that appear in the background, or that there are some objects that could be mistaken as a face.

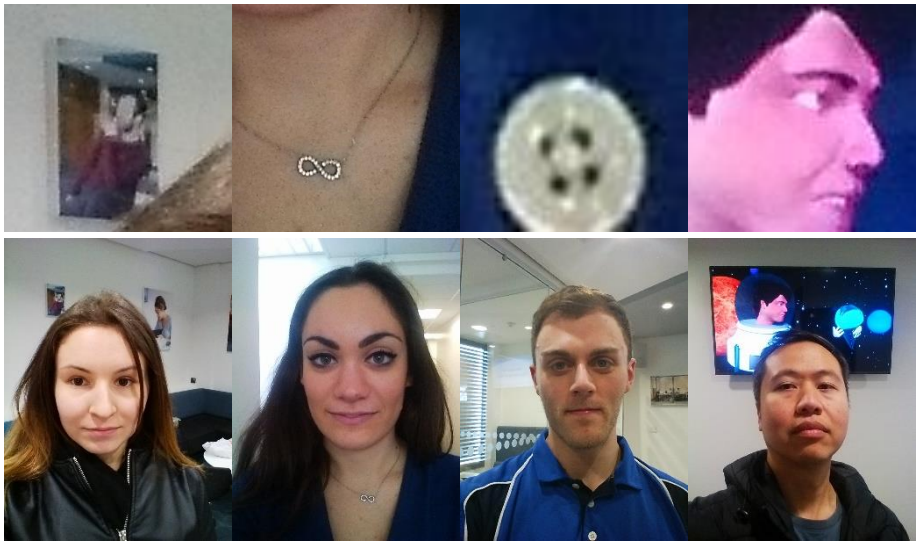


Figure 5.9: Examples of objects (images above) that were mistakenly detected as facial areas on the respective facial images (shown below).

Considering the Viola-Jones algorithm, the number of faces detected in each image was saved to assess if there was a relationship with the images with more than one face detected and the ones that were detected incorrectly by Viola-Jones. Out of the 9103 images, discarding the 289 images where a face has not been detected, the number of faces detected in the remaining images can be up to 7. Some were actual real faces of other subjects that were passing by, but others were objects that show or have the shape of a face, as shown in Figure 5.9.

The percentage of images that detected more than one facial area was 17%: 14.5% presented two faces, while 2% of the images presented three or more. When the system detects more than one face, it could erroneously select the facial area that does not correspond to the face of the user. The Viola-Jones algorithm presented 1.23% of the images where the face was not or was partially included in the segmented facial area. The images were assessed to check whether having more than one detected face in the image affected the detection system.

Despite the fact that having other people's faces on the background could create an element of "noise" for the detection system, from the results it appeared that detection performance was not affected. When facial detection is applied on mobile devices, the facial area to be considered is the one closer to the frontal camera, as it can be easy to filter out extra facial areas detected simply by applying a filter over the facial area dimensions. The larger the facial area detected, the closer to the camera.

5.4 User interaction

Face detection algorithms can be influenced by the users' aspect and interaction. As noted in Chapter 2, the ISO/IEC 29749-5 Technical Report [7] describes the user's characteristics with respect to device interaction and divides them into two categories: static and dynamic. The work presented in this section will be an analysis of the effect that user's demographics and the static and dynamic characteristics have on the detection of the facial area, as well as an assessment of the user's experience during the data collection.

5.4.1 Demographics

From the frequency of FTD, it can be seen from the chart in Figure 5.10 that male users had more images where their face was not detected.

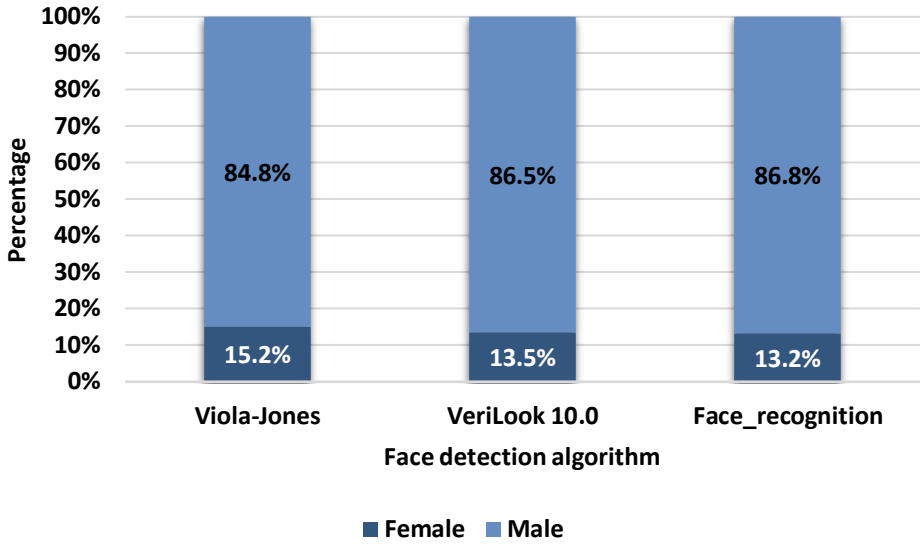


Figure 5.10: FTD in respect to sex for the different face detection algorithms.

As confirmed by the Chi-Square test, there is a significant negative association between the participants' sex and face detection outcome within Viola-Jones ($\chi^2(1) = 150.54$, $p < 0.001$), VeriLook 10.0 ($\chi^2(1) = 357.28$, $p < 0.001$) and Face_recognition $\chi^2(1) = 233.13$, $p < 0.001$). The Phi coefficient was calculated as it measures the strength on an association between two binary variables.

Phi values ranges between 0 and 1; a small effect is considered for Phi values of 0.1, a medium effect for values of 0.3 and a large effect for 0.5. According to the Phi values calculated for the detection systems, the strength of the association is not strong, as it presents values of -0.13 for Viola-Jones, -0.2 for VeriLook 10.0 and -0.16 for Face_verification.

When investigating the relationships that FTD has with age groups, the older the participant, the more frequent the occurrence of FTD, as shown in Figure 5.11. The 24-40

age group reported the highest number of FTDs across all the three algorithms used, while VeriLook 10.0 recorded a highest number for ages between 31 to 35.

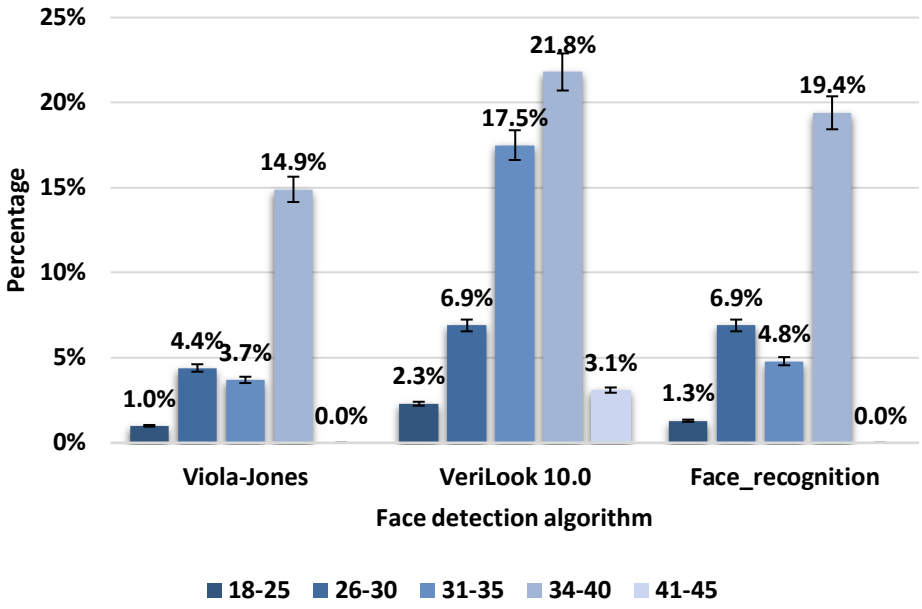


Figure 5.11: FTD recorded by the algorithms. Each percentage is calculated in respect to the total images taken per each age group.

A Chi-square test was performed to check whether there was a significant association between the variables. Viola-Jones algorithm reported a weak (Cramer's V: 0.22) but significant association with $p < 0.001$ and $\chi^2(4) = 446.13$. Similar results were obtained for VeriLook 10.0 ($\chi^2(4) = 658.01$, $p < 0.001$) and Face_recognition ($\chi^2(4) = 588.21$, $p < 0.001$), and the strength of the association was reported with the Cramer's V constant of 0.27 and 0.25 respectively.

The difference between age could be explained in different ways. It could be possible for instance that the participants that have a younger age are more used to take self-portrait images with the smartphone and that could be an explanation for the smaller number of FTDs. It could also be considered the age factor as a variable that affected the detection of the image.

Differences were also investigated considering the participants ethnicity. The relationship with the detection outcome was assessed using a Chi-Square test. The statistical test revealed that there is a weak but significant association between the ethnicity groups considered and the outcome of face detection algorithms. Viola-Jones reported a Cramer's V constant of 0.147 for $\chi2(3) = 197.45$, $p < 0.001$. The Cramer's V constant reported for VeriLook 10.0 is 0.126 for $\chi2(3) = 143.62$, $p < 0.001$. Face_recognition reported a significant association for $\chi2(3) = 266.09$, $p < 0.001$, with Cramer's constant of 0.171. The differences observed across the ethnicity groups can be seen in Figure 5.12.

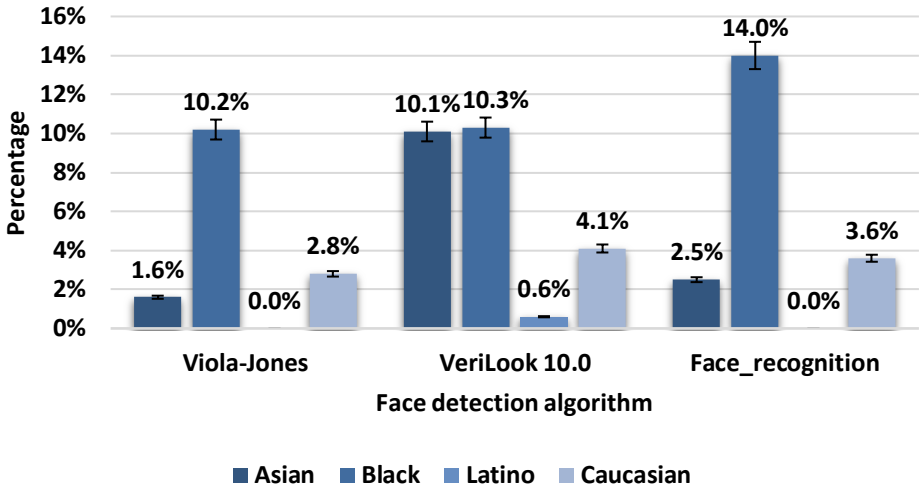


Figure 5.12: FTD recorded by the algorithms. Each percentage is calculated in respect to the total images taken per each ethnicity group.

The level of education was also considered as a variable that could affect the detection of a facial area in the image. Participants that reported a higher level of education, also reported the higher percentages of FTD, as shown in the bar chart in Figure 5.13.

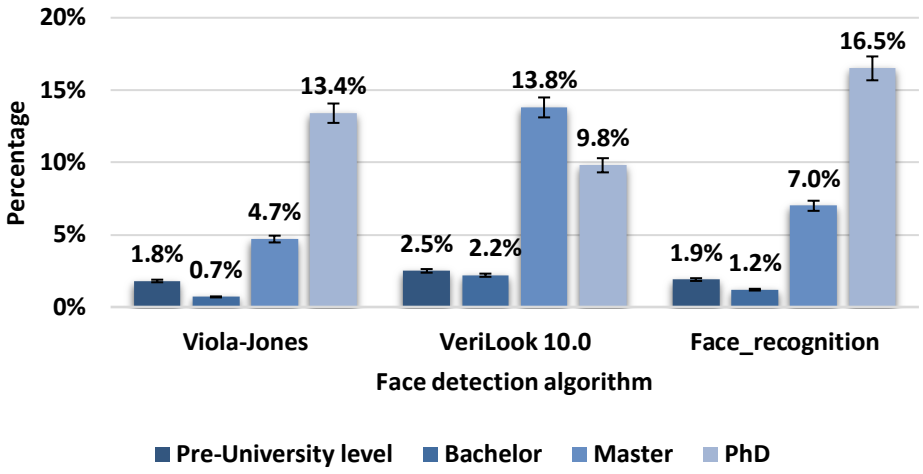


Figure 5.13: FTD recorded by the algorithms. Each percentage is calculated in respect to the images taken per each completed education group.

The Chi-Square test reported significant association for Viola-Jones ($\chi^2(3) = 315.29$, $p < 0.001$ with Cramer's V: 0.186), VeriLook 10.0 ($\chi^2(3) = 420.51$, $p < 0.001$ with Cramer's V: 0.215) and Face_recognition ($\chi^2(3) = 381.55$, $p < 0.001$ with Cramer's V: 0.205), but the Cramer's V constant showed that the strength of the association is weak.

A possible explanation to the high percentage of FTD observed for higher level of education could be considering that in our database the participants that collected the images were mainly university members and as such the level of education is strictly connected to the age of the participants and to the results obtained for age groups. There

were not cases for instance of a lower degree of education for older participants, and this aspect could have influenced the analysis, resulting in a high number of FTD recorded for PhD participants.

Finally, the images were divided in groups considering the operating system used by the participants on their own devices, that were recorded as Android, iOS or both operating systems. The percentage of FTD calculated for each group across the face detection algorithms can be seen in Figure 5.14.

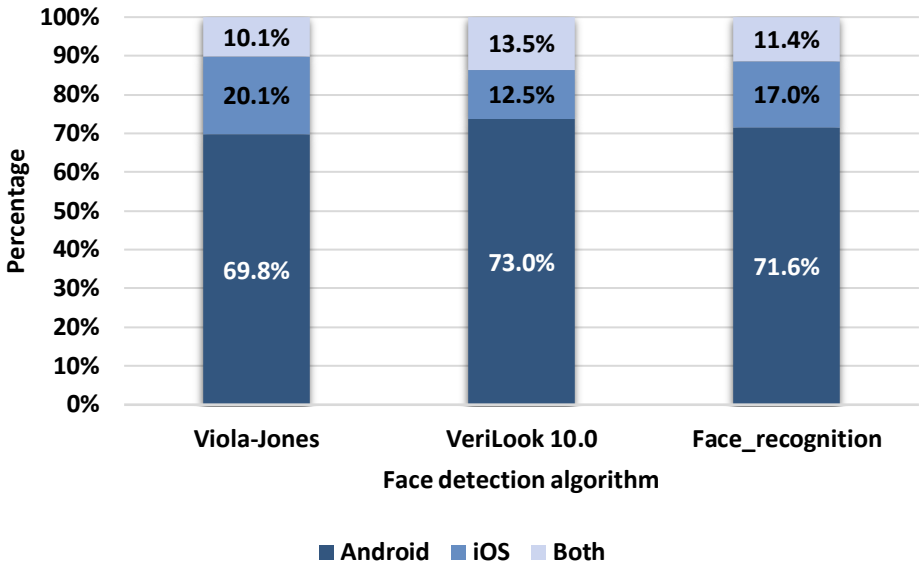


Figure 5.14: FTD in respect to operating system used for the different face detection algorithms.

It appears that participants that own an Android device recorded more FTDs than participants using iOS as operating systems. The associations are significant according to the Chi-Square test: the significant values and the respective Phi coefficients are reported on the Table 5.8. However, the strength of the association between all groups resulted weak as can be seen when observing the Phi values.

Table 5.8: Statistically significant associations between FTDs and Operating System used by the participants.

Algorithms	Operating System	Chi-Square	p	Phi
Viola-Jones	Android	$\chi2(1) = 18.09$	$p < 0.001$	-0.045
	iOS	$\chi2(1) = 35.2$	$p < 0.001$	0.062
VeriLook 10.0	Android	$\chi2(1) = 106.07$	$p < 0.001$	-0.108
	iOS	$\chi2(1) = 115.2$	$p < 0.001$	0.113
Face_recognition	Android	$\chi2(1) = 40.47$	$p < 0.001$	-0.067
	iOS	$\chi2(1) = 59.14$	$p < 0.001$	0.081

When considering prior experience that participants declared to have with biometric systems for mobile authentication there were no significant differences observed between the groups.

In conclusion, it can be seen from the results that there is a significant association amongst all the variables considered for demographics and the number of FTDs that occurred when using the detection algorithms, although some demographic groupings have a stronger association than others.

5.4.2 Static characteristics

The ISO/IEC 24947-5 Technical Report describes the different characteristics of the user that are considered static. The characteristics considered for this analysis are described in detail in Chapter 4.

To check whether there were glasses (including sunglasses) in the facial image, the Neurotechnology VeriLook 10.0 SDK was used as it enables an automatic detection of glasses within an image. A visual inspection of the database was also performed to confirm whether the outcome of the detection of glasses in the image was correct. Clearly, the assessment of this characteristic was only possible for those images that VeriLook 10.0 was able to detect a face, so this algorithm was excluded from the analysis when comparing the algorithms FTDs and the characteristic relationship.

A total of 2198 images where detected using VeriLook 10.0 containing glasses with a further 494 containing dark glasses. Although the participants did not present dark glasses at the beginning of the session, there were cases of participants where they used dark lenses when unsupervised in the outdoor environment.

There were also two participants that were wearing glasses with photochromic lenses that darkened during the session. There were also some dark glasses, as indicated in Table 5.9, that were erroneously detected in the experimental room that could not possibly been used because the session was supervised and the operator would have reported it.

Table 5.9: Number of images where glasses or dark glasses were detected and the percentages according to camera type.

Static feature	SLR	Smartphone camera
	Percentage	Percentage
Glasses	22.6%	25.8%
Dark glasses	5.7%	5.8%

A Chi-Square test was performed to check whether there was a significant association between the images that presented a user wearing glasses and the outcome of the detection system. The test reported a significant association for the images detected when using Viola-Jones. The negative association for $\chi^2(1) = 4.825$, $p = 0.028$ indicated that the images where the static characteristic was present reported lower FTDs, and the

strength of the association was weak according to the Phi value of -0.024. Face_recognition report significant results for $\chi^2(1) = 0.48$, $p = 0.489$, but the strength of the association is closer to 0, as the Phi coefficient resulted -0.008.

These results appear to be encouraging, as the presence of glasses in the image resulted to have a significant but small effect on face detection. These results indicate that users to not need to remove their glasses every time they need to authenticate on their devices as this will have a huge impact on the acceptability of the system. In Figure 5.15 the chart illustrates the percentages of FTDs for participants where glasses were and were not detected.

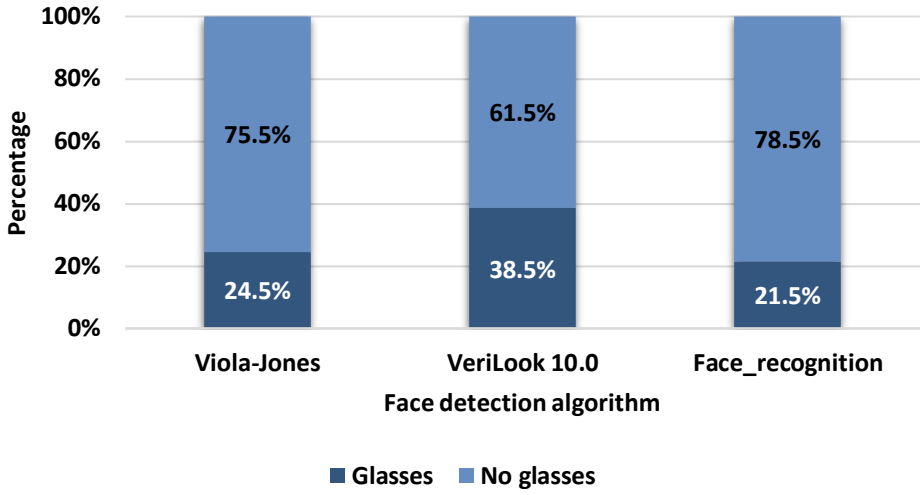


Figure 5.15: Percentages of images among the FTDs in which participants wear glasses for each detection algorithm.

The images where participants owned a beard recorded a 6.3% of FTD with Viola-Jones, 8.6% with VeriLook 10.0 and 8.9% with Face_recognition. The Chi-Square test performed revealed a number of significant associations, the values are reported in Table 5.10. The test was performed to check whether the presence of a beard in the image could affect the detection outcome.

Table 5.10: Statistically significant associations between FTD and images where participants presented a facial image including a beard.

Algorithms	Chi-Square	p	Phi
Viola-Jones	$\chi^2(1) = 53.58$	$p < 0.001$	-0.081
VeriLook 10.0	$\chi^2(1) = 10.86$	$p < 0.001$	-0.037
Face_recognition	$\chi^2(1) = 83.31$	$p < 0.001$	-0.101

There were 5 participants in particular that presented differences in the presence of a beard in the images they presented during the data collection. Results shows that despite

the significant results, the Phi coefficient shows a small strength of association between the variables, indicating a small effect in the differences between the two groups of images.

Across images that presented heavy make-up, the percentages of FTDs with Viola-Jones was 4.4%, while 1.6% and 4.4% were the percentages of FTD reported by VeriLook 10.0 and Face_recognition respectively. When performing the Chi-Square test, no significant associations were found in this case.

The results indicate that the presence of make-up that participants wore during the data collection, replicating a realistic scenario, did not contribute in adding a variation in the image that significantly affect the detection of the facial area.

5.4.3 Dynamic characteristics

The Neurotechnology VeriLook 10.0 SDK also enabled the detection of those features that are considered as dynamic characteristics which included blink and open mouth but also participants' facial expressions, as previously described in detail in Section 4.4.2.2.

In Table 5.11 are reported the percentages of the images taken with the smartphone detected by VeriLook 10.0 that presented a dynamic feature. For comparison, the Table also reports the percentages of the same features detected when the images where taken with the SLR.

Table 5.11: Percentages of images where VeriLook 10.0 detected a specific dynamic characteristic.

Dynamic characteristic	SLR	Smartphone camera
	Percentage	Percentage
Blink	0.6%	9.6%
Mouth open	24.2%	19.7%

From the analysis, when users were blinking in an image, the system had more difficulties to detect a facial area when this characteristic was present. This was confirmed by the Chi-Square tests: there was a significant association with FTD images, for $\chi^2(1) = 90.95$ when using Viola-Jones and $\chi^2(1) = 123.02$ when using Face_recognition. The Phi values reported a negative association for -0.105 and -0.123 for Viola-Jones and Face_recognition respectively. Despite the Phi coefficients indicate a small strength of association within the variables, the effect size can have a bigger impact when considering a larger dataset of images.

Results underlines that this variable should be taken into account on real life applications when considering the accuracy of the facial area detection, and that is effect could be significant when considering a larger population of data, since this characteristic was present on around the 10% of the images in our database.

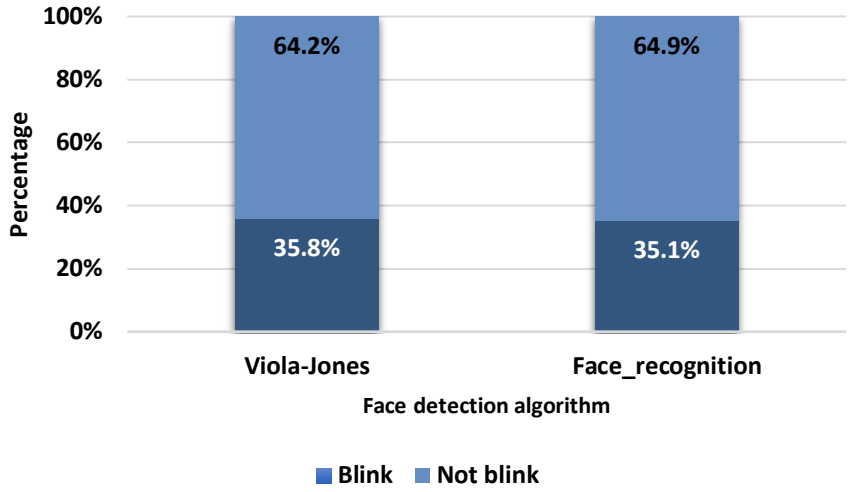


Figure 5.16: Percentage of image where blink was detected amongst the FTDs.

In Figure 5.16 it is possible to see the percentage of images that presented blink amongst the FTDs recorded by the two face detection algorithms. A similar case resulted when considering the dynamic characteristic of mouth open. Whether the users had their mouth open (or not) is reported as a weak but significant association with Viola-Jones FTD images ($\chi^2(1) = 18.46$, $p < 0.001$ Phi: -0.048) and Face_recognition ($\chi^2(1) = 47.63$, $p < 0.001$ Phi: -0.077).

The percentages of mouth open amongst the FTDs can be seen in Figure 5.17. Despite the significant results, the strength of the associations between the images in this case is smaller than for those observed for blink, indicating that the effect of a subject's mouth open characteristic has a small size effect on the detection outcome.

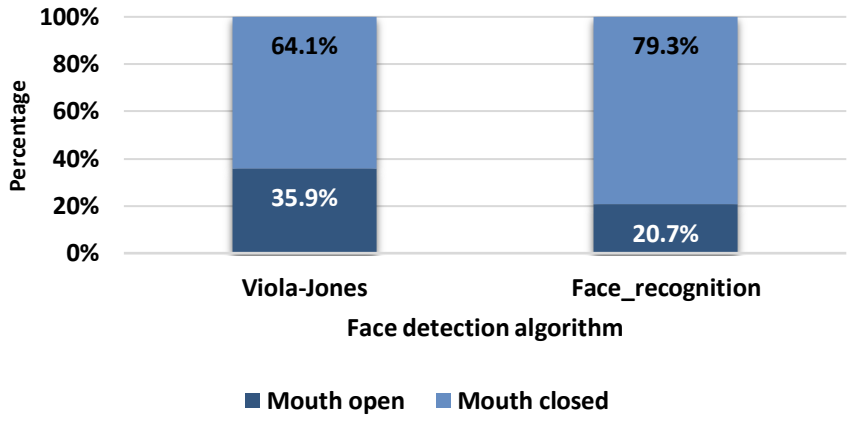


Figure 5.17: FTDs where mouth open was detected in the facial image.

5.4.3.1 Facial expression

VeriLook 10.0 detected 7 different facial expressions from the users' images (Figure 5.18).

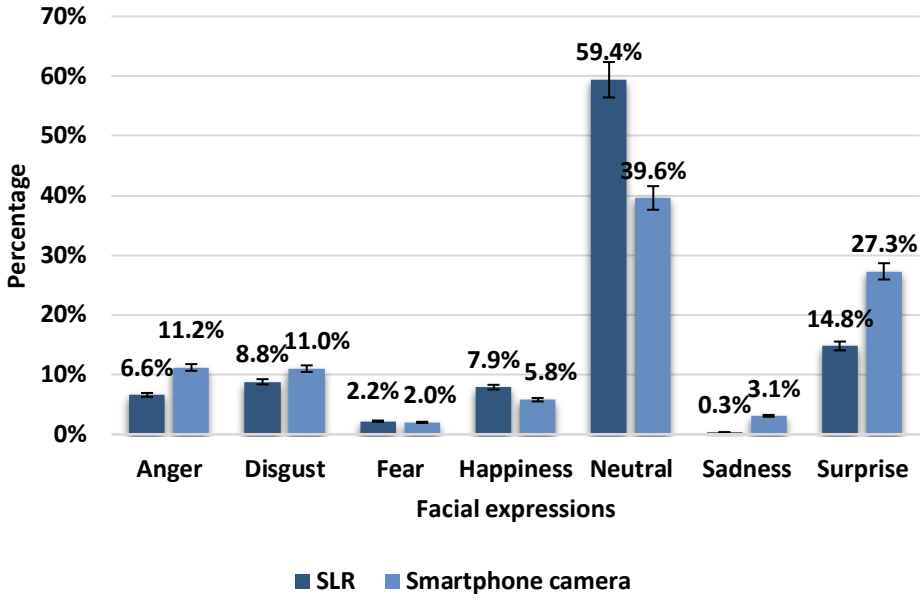


Figure 5.18: The percentages of facial expressions recorded when the images were taken with the SLR and the smartphone camera

Neutral is the most popular facial expression, as expected, even more so when the images were taken with the SLR. From the Chi-Square test, the associations between the facial expressions and the FTD with the detection algorithms presented a significant association but it was not particularly strong. Viola-Jones ($\chi^2(6) = 162.12$, $p < 0.001$) presented in fact a Cramer's V constant of 0.138 while Face_recognition ($\chi^2(6) = 226.78$, $p < 0.001$) reported Cramer's V = 0.163.

It can be seen that, despite a significant association between the facial expression presented in the image and the detection of the subject's facial area, the size effect that this dynamic characteristic has on the outcome of the detection system is not particularly strong.

5.4.3.2 Users' pose

The head pose that the user presents towards the camera has been an important area of study as it has shown that it can have an impact on the performance of the system [95]. In the Standard ISO/IEC 19794-5 Biometric Data Interchange Formats – Part 5: Face Image Data [8], user pose has been presented in terms of angular rotations:

- Yaw angles are the rotations in degrees about the vertical axis (y), similarly to a "head-shaking" movement. Yaw angles are positive when the head is facing left, and negative when facing right.
- Pitch indicates the rotation angles about the horizontal axis (x) like a nodding movement. Positive degrees angles are presented when the person is looking down, negative when is looking up.

- Roll angles are rotations about the horizontal back-to-front axis (z). Positive angles are representative of a head tilt toward the right shoulder and negatives when the head is tilted towards the left shoulder.

The pose angles described in the Standard are illustrated in Figure 5.19 with a frontal pose that presents a reference point at the (0,0,0) rotation angles.

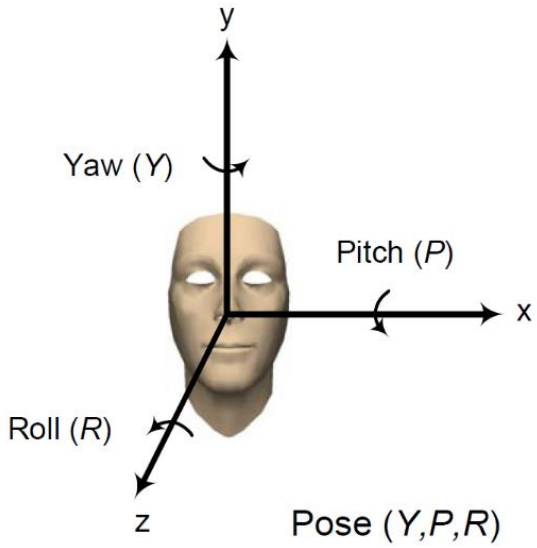
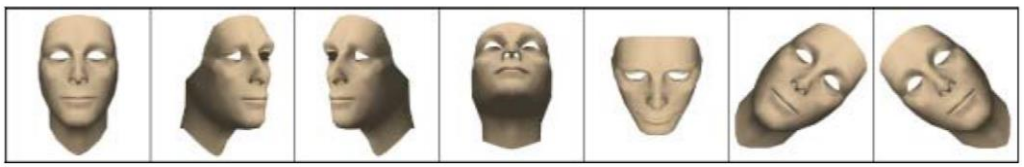


Figure 5.19: Pitch, Yaw and Roll angles indicated for the user's face in frontal pose.

According to the Standard ISO/IEC 19794-5, to enhance the performance of an automated facial recognition system, the user's pose should follow the following requirements: pitch and yaw should not present a rotation that is more than ±5 degrees from the frontal reference, while the rotation of the head should show not more than ±8 degrees from the frontal reference for roll. Examples of different rotations is shown in Figure 5.20.



(0, 0, 0) **(+45, 0, 0)** **(-45, 0, 0)** **(0, -45, 0)** **(0, +45, 0)** **(0, 0, -45)** **(0, 0, +45)**

Figure 5.20: Examples of different pose angles (Y, P, R).

The Discriminative Response Map Fitting method (DRMF 2013), as presented in [81], was used to calculate the facial landmarks and the estimations of the angular rotations, as mentioned in Section 4.4.2.2.

Pose angles recorded presented a normal distribution, and the mean and standard deviation values across all images are presented in Table 5.12.

Table 5.12: Mean and standard deviation of poses across different camera sensors.

Rotation angle	SLR camera				Smartphone camera			
	Min	Max	Mean	St deviation	Min	Max	Mean	St deviation
Pitch	-14.72	3	-6.50	3.51	-27.78	14.80	-8.97	4.60
Yaw	-19.69	17.52	0.26	3.70	-29.04	40.65	-1.06	3.90
Roll	-10.62	11.98	-2.80	2.84	-19.52	20.23	-2.98	2.44

If the indications provided by the Standard ISO/IEC 19794-5 were followed, the vast majority of our images, even the ones taken with an SLR, would not conform with the requirements. Table 5.13 shows the percentages of the images that conformed with the angular pose requirements.

Table 5.13: Percentages of images that were compliant with the Standard ISO/IEC 19794-5 image acquisition requirements for user pose.

ISO\IEC 19794-5 user's pose compliance:	SLR camera	Smartphone camera
Compliant	32%	12.6%
One pose not compliant	64.2%	76.9%
Two poses not compliant	3.8%	9.8%
None of the poses is compliant	0%	0.7%

When collecting images with the SLR, the static requisitions for the collection of passport images were considered to adjust the camera fixing and the image background. Dynamics requirements, such as the user's head pose and facial expressions, were assessed according to the participants' interaction to simulate a more realistic scenario. For this reason, even if the conditions of images taken with the SLR were similar to a passport enrolment scenario, 68% of the images were not compliant with the angular pose requirements. Yaw is the head angular rotation that presented the highest percentage of not compliance (66.2%), while pitch and roll angles were not compliant only for the 3.5% and 2.2% of the images.

Higher percentages of not compliant images were observed for the smartphone camera: yaw angles were not compliant for the 83.4% of the images, pitch for the 12% and roll for the 3.1%. A logistic regression model was designed for each detection algorithms to assess whether the face detection outcome could be estimated acknowledging the three angular rotations.

The test reported significant results for the VeriLook 1.0 ($\chi^2(8) = 208.34$, $p < 0.001$) and the Face_recognition ($\chi^2(8) = 34.684$, $p < 0.001$) systems, but not for the Viola-Jones algorithm ($\chi^2(8) = 13.39$, $p = 0.099$). The regression model designed for VeriLook 10.0%

explained between 4% (Cox & Snell R^2) and 13% (Nagelkerke R^2) of the variance and was able to classify correctly 93.5% of the FTD cases.

The equation for the VeriLook 10.0 regression model can be read as follows:

$$logit(p) = 2.25 - 0.14 * \text{Pitch} + 0.02 * \text{Yaw} + 0.17 * \text{Roll}$$

Similarly, the regression model designed for Face_recognition was able to estimate correctly 95.7% of the FTDs, but yaw angles did not present significant results as contributor. The Model explained between 3% (Cox & Snell R^2) and 9.8% (Nagelkerke R^2) of the variance and the formula can be read as follows:

$$logit(p) = 2.23 - 0.15 * \text{Pitch} + 0.07 * \text{Roll}$$

A summary of the significant values for is presented in Table 5.14.

Table 5.14: Statistical values for the logistic regression considering head angular rotations as contributors across the face detection algorithms.

Detection algorithm	Pose angle	B	S.E.	Wald test	df	p	Odds Ratio
VeriLook 10.0	Pitch	-0.140	0.009	255.499	1	0.000	1.150
	Yaw	0.021	0.010	4.348	1	0.037	0.979
	Roll	0.169	0.016	113.007	1	0.000	0.845
Face_recognition	Pitch	-0.155	0.010	234.161	1	0.000	1.167
	Yaw	0.007	0.012	0.392	1	0.531	0.993
	Roll	0.070	0.019	13.648	1	0.000	0.932

If the requirements from the ISO/IEC 19794-5 Standard were followed, it could be possible to estimate around 80% of the FTD images and discard them accordingly (Table 5.15). However, if the images where angular poses do not conform with the requirements were rejected, around 88% of the images where a face was correctly detected would be erroneously discarded.

Table 5.15: Percentages of FTD and detected images according to the user's head pose compliance.

ISO\IEC 19794-5 user's pose compliance:	Viola-Jones		VeriLook 10.0		Face_recognition	
	FTD	Detected	FTD	Detected	FTD	Detected
Compliant	21.8%	12.3%	18.6%	12.1%	22.3%	12.1%
One pose not compliant	54.9%	77.7%	55.1%	78.5%	54.1%	78%
Two poses not compliant	19.1%	9.5%	20.8%	9%	19.5%	9.4%
None of the poses are compliant	4.2%	0.5%	5.5%	0.4%	4.1%	0.5%

The requirements for user's pose defined in the ISO/IEC 19794-5 Standard should be adjusted to adapt to the variability that the head angular rotations present over smartphone images. From the results, yaw angles presented the highest percentage of not compliance. The requirements specify that there should not be a variation of ±5 degrees from the reference system; applying a more permissive variation of movements, the performance could be improved, as for the example shown in Table 5.16.

Table 5.16: Percentages of FTD according to yaw angles with different degrees of compliance.

Yaw angles requirements	Viola-Jones	VeriLook 10.0	Face_recognition
±5 degrees	14.31%	10.4%	7.69%
±10 degrees	5.56%	3.54%	2.59%

The Table shows the percentages of FTD that occur when considering the images that presented yaw angles within the range of ±5 degrees as compliant with the Standard compared to yaw angles within the range of ±10. The angles requirements could be adjusted according to the application of the detection system and the algorithm used.

5.4.4 User's opinions and experience

The user's point of view is important when performing biometric authentication but it is an aspect that is often overlooked. If the person is not feeling comfortable during the biometric presentation to the sensor, the quality of the sample can be lowered which can have an influence on the performance of the verification system.

This aspect is more frequently considered in behavioural biometrics, but in the case of facial verification, if the users are not feeling at ease when taking the facial images, it could have an impact in the verification score. Users could feel uncomfortable in taking facial images, for instance, in a specific location type or if the users feel that they are being observed by other people. In some cases, the verification image could appear blurred because it is taken in a hurry, or the users could show an unusual pose because they did not want to show to others that they were taking a picture of themselves.

To enable an analysis of the relationship between participants' experiences that they had during the data collection, questions were asked of the participants with the aim to compare whether they encounter difficulties in the different situations in which they were taking the images, either indoors or outdoors and either with or without the presence of other people. At the end of each session of the data collection, participants were asked to complete a questionnaire indicating their experience of the whole session on the Likert Scale [55] from 1 (strongly disagree) to 5 (strongly agree). An overall analysis of the questionnaires answers was described in Section 4.3.4.

A statistical analysis was carried out to check whether the experience that the participants reported in the questionnaire could have had an impact in the way they were

taking the images with the smartphone and consequently in the face detection outcome. A Chi-Square test reported significant results in the FTD recorded when participants were collecting the facial images in presence or not of other people during the acquisition of the image ($\chi^2(4) = 140.03$, $p < 0.001$). From the results it seems that when nobody was around during the image capture, participants that felt more confident were less likely to report an FTD.

When considering different location types, the chances of an FTD increases as the users expressed less confidence that they were taking good sample images, as happened when considering images from outdoors location, and this was assessed with a Chi-Square test that reported significant differences between the group of images ($\chi^2(4) = 197.4$, $p < 0.001$). On the contrary, when the participants reported more confidence that they provided good image sample for face verification, the likelihood of having an FTD decreases, as happens when the images were taken indoors.

Finally, a Chi-Square test was performed to confirm the increased likelihood to result in an FTD as participants found it harder to present facial images to the smartphone, either because they did not know how to present themselves to the camera or how to position the device ($\chi^2(4) = 247.36$, $p < 0.001$).

5.5 Quality metrics across FTD images

One of the reasons for which a detection algorithm could fail to locate the user's face is the quality of the image. If, for instance, the image contains excessive blurriness or brightness, the detail of the user's face might not be evident enough for the algorithms to identify the necessary features that could enable the detection of the facial area.

Assessing the quality of images that resulted in an FTD could be useful in understanding aspects that have an influence on the detection outcome and maybe predict it when similar situations are presented. The quality assessment was carried out calculating the Facial Image Quality (FIQ) metrics for the facial area of each image, after it was manually segmented using the **'imcrop'** command in MATLAB. The FIQ metrics that were considered are described in detail in Section 4.4.3.1 and include: Brightness, Contrast, Global Contrast Factor (GCF), Blurriness and Exposure. The values were normalised to be within a range between 0 and 5 to enable a comparison between the FIQ metrics.

From the results it appears that the quality metrics of images reporting an FTD presented an approximate normal distribution. Figure 5.21 shows the mean values for each of the FIQ metric calculated for the FTD reported by each detection algorithm and for the images that Viola-Jones reported as "detected" but for which the facial area did not correspond to the participant's face.

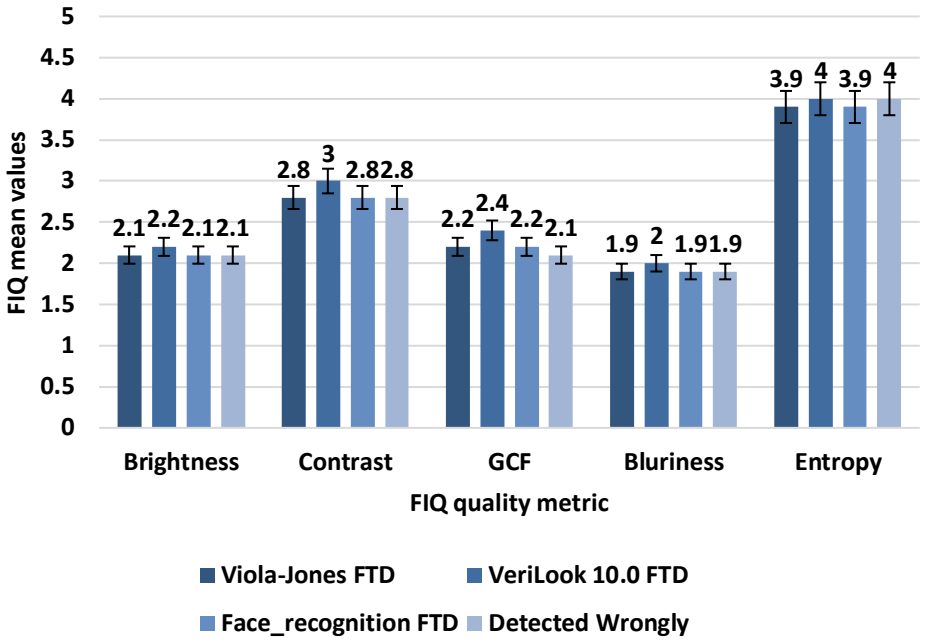


Figure 5.21: Mean values of FIQ metrics calculated for the FTDs reported by the detections algorithms and for the wrongly detected facial images.

A logistic regression was performed to check whether it could be possible to estimate the detection outcome knowing the FIQ metrics of the images. The model was designed for each detection algorithm and reported statistically significant results in each case. When considering Viola-Jones, the significance was observed for $\chi^2(8) = 42.61$ $p < 0.001$ and the model explained between the 5% (Cox & Snell R^2) and the 22% (Nagelkerke R^2) of the variance, predicting the facial image detection for 96.7% of the cases.

Not all the FIQ metrics contributed to the estimation of the detection outcome. As can be seen in Table 5.17, Brightness was not a significant contributor to the model. The logistic regression equation for this detection algorithm can be read as follows:

$$logit(p) = -2.72 + 2.266 * \text{Contrast} + 0.867 * \text{GCF} + 0.378 * \text{Blurriness} - 0.852 * \text{Exposure}$$

Table 5.17: Logistic regression predicting the detection of a facial area when using Viola-Jones.

FIQ metric	B	S.E.	Wald test	df	p	Odds Ratio	95% C.I. for Odds Ratio	
							Lower	Upper
Brightness	-0.198	0.162	1.499	1	0.221	0.820	0.597	1.126
Contrast	2.266	0.203	124.047	1	0.000	9.636	6.468	14.357
GCF	0.867	0.112	60.309	1	0.000	2.379	1.912	2.961
Blurriness	0.378	0.109	11.973	1	0.001	1.460	1.178	1.809
Exposure	-0.852	0.245	12.108	1	0.001	0.427	0.264	0.689

Similarly, the model designed for VeriLook 10.0 reported significant results for $\chi^2(8) = 37.13$ $p < 0.001$. For this detection algorithm the model explained between the 8.5% (Cox

& Snell R^2) and the 22% (Nagelkerke R^2) of the variance, predicting correctly 94.3% of the cases. The FIQ metrics were considered as contributors, but GCF in this case did not report significant results, as can be seen in Table 5.18.

Table 5.18: Logistic regression predicting the detection of a facial area when using VeriLook 10.0.

FIQ metric	B	S.E.	Wald test	df	p	Odds Ratio	95% C.I. for Odds Ratio	
							Lower	Upper
Brightness	-1.153	0.117	96.384	1	0.000	0.316	0.251	0.397
Contrast	1.794	0.149	145.359	1	0.000	6.016	4.494	8.054
GCF	-0.123	0.076	2.631	1	0.105	0.884	0.761	1.026
Blurriness	0.391	0.077	25.579	1	0.000	1.478	1.270	1.720
Exposure	1.201	0.183	43.100	1	0.000	3.322	2.321	4.754

For VeriLook 10.0, the logistic regression equation can be read as follows:

$$logit(p) = -6.24 - 1.153 * \text{Brightness} + 1.794 * \text{Contrast} + 0.391 * \text{Blurriness} + 1.201 * \text{Exposure}$$

Finally, a logistic regression model was designed for Face_recognition reporting significant results for $\chi^2(8) = 21.21$ $p = 0.007$. The model was able to estimate 95.7% of the cases, explaining between the 7.4% (Cox & Snell R^2) and the 24.6% (Nagelkerke R^2) of the variance. Similarly, as observed for the other detection algorithms, the model was designed with the five FIQ metrics as contributors, but in this case image Exposure did not report significant results. The regression equation can be read as follows:

$$logit(p) = -4.54 - 0.448 * \text{Brightness} + 2.11 * \text{Contrast} + 0.542 * \text{GCF} + 0.531 * \text{Blurriness}$$

The variables for the logistic regression model designed for Face_verification can be seen in Table 5.19. Knowing the variations amongst the selected quality metrics it was possible to estimate the facial detection outcome. These results were valid for all the detection systems, confirming that a prediction of an FTD using the quality of an image can be applied in realistic mobile scenarios. Some differences were observed between the contributors for each designed logistic model, although it could be explained by considering the different methods used within each detection algorithm considered in this analysis.

Table 5.19: Logistic regression predicting the detection of a facial area when using Face_recognition.

FIQ metric	B	S.E.	Wald test	df	p	Odds Ratio	95% C.I. for Odds Ratio	
							Lower	Upper
Brightness	-0.448	0.142	9.990	1	0.002	0.639	0.484	0.843
Contrast	2.110	0.179	138.650	1	0.000	8.249	5.806	11.720
GCF	0.542	0.096	32.113	1	0.000	1.719	1.425	2.074
Blurriness	0.531	0.097	29.917	1	0.000	1.700	1.406	2.056
Exposure	-0.146	0.219	0.447	1	0.504	0.864	0.563	1.326

5.6 Face detection: overall observations

Face detection is a fundamental first step for facial verification. If the facial area of the image is not detected correctly, the biometric verification system can be impacted on its performance. The analysis undertaken and described in this Chapter investigates the different variables that could occur when detecting faces in mobile facial recognition and is aimed to identify the aspects that are valid in the mobile scenario.

We have used several state-of-the-art algorithms to perform face detection. In particular, one of the algorithms assessed in this analysis used a tree-based method and seemed able to detect all the images that were collected in the proposed database except for one. Despite the resilience that this algorithm showed when detecting images that were taken in different types of environments, there are still some aspects that needs to be considered. In fact, the time required to detect facial areas using this method was longer than the other algorithms assessed in this study, and the time necessary to verify the users on their mobile devices is an extremely important acceptability issue, as explained already in Chapter 3. Furthermore, the detection accuracy should be assessed to ensure that the estimation that the system makes to locate the face in the image is actually correct.

From the results presented in this Chapter, it appeared that the larger number of FTDs occurred when the images were taken in outdoors locations. There was not a clear improvement in detecting the faces when considering subsequent sessions for the data collection. An explanation of this could be the unpredictable environmental elements that can change across days in outdoors locations. The environment effect was analysed in this work and included a background analysis that reported interesting results. The results showed not only that background information could be useful to predict the outcome of a face detection algorithm, but it was also determined that knowing the texture of the background it could be used to estimate whether the images were taken in indoors or outdoors locations.

The results also showed that even if other faces are present in the images, they did not have a strong influence in the outcome of face detection. A filter could be applied that would remove erroneous facial areas in images considering the dimensions of the area and the distance between the smartphone and the user.

Demographics appeared to have a statistically significant association with images where a face was not detected. Participants' sex and ethnicity reported stronger associations, but significances were also found for age. The user's educational background and the prior experience with the biometrics did not report strongly significant associations.

When considering the static characteristics of the users, the significant associations found were not particularly strong. More interesting results were obtained when studying the associations between the images that failed to have a face detected and the head

angular poses that the participants presented to the smartphone camera. It was possible to estimate the outcome of the detection system knowing the head pose angles of the users. The ISO/IEC 19794-5 Standard for passport images resulted difficult to be applied in a mobile scenario and should be reconsidered and adapted to the variations of pose angles that the subjects presented in the database, in particular for yaw angles.

This study also highlights the importance that users' opinion can have on mobile face authentication. According to the answers that the participants provided in the questionnaires at the end of each experimental session, there were associations found with the FTDs occurrences. From one side, it is important to educate the users on how to use the technology to avoid, for instance, non-compliant head poses or other known elements that can affect the detection of the face in the image. On the other side, mobile developers need to take in consideration the opinions of the users and understand what experiences they encounter that could affect the presentation of the biometrics.

Finally, a quality assessment revealed that there were statistically significant associations between the FIQ variables and the FTD outcomes, and that it could be possible to estimate the outcome of the detection image by assessing the FIQ metric values recorded for the facial image.

To have an enhanced perspective of the quality assessment of these images, the results should be compared with the FIQ metrics calculated for the detected faces in the database. A complete quality assessment will be presented in the following chapter.

The Verification Process: Quality Assessment

6.1 Introduction

Quality assessment is an integral part of biometric facial authentication. By investigating the quality of a facial image, it is possible to contemplate whether a rejection from the verification system is caused by an impostor or by a genuine user that is presenting a poor-quality image. Uncontrolled light exposure, user interaction and poor resolution of the camera are a few examples of elements that could influence the quality of a facial image and hence reduce the biometric performance, particularly in a mobile context.

This study contained in this Chapter embraces a quality assessment of facial images taken within a mobile context. The aim is to define requirements and observations to estimate the quality of an image and adjust them to ensure higher mobile facial verification performance. The Facial Image Quality (FIQ) metrics, defined in Section 4.4.3.1 were calculated for the whole database formed of 9,103 images taken with a smartphone camera and a further 318 images taken with the Single Lens Reflex (SLR) camera. The metrics were calculated on previously segmented facial areas. Where it was not possible to automatically detect them, facial regions were manually cropped using MATLAB. Five quality metrics were chosen from those proposed by the ISO/IEC 29794-5 Technical Report [7]: Brightness, Contrast, Global Contrast Factor (GCF), Blurriness and Exposure. The FIQ metrics were normalised to range from 0 to 5 to enable a comparison between metrics.

The analysis presented in this Chapter describes the variations recorded in quality across the different scenarios. Image quality was then investigated with respect to the environment, the user interaction and camera characteristics.

6.2 Assessing image quality across scenarios

The analysis used the experimental data collection as defined in Section 4.2.3. The analysis across the different scenarios within the dataset replicates several aspects to consider when implementing biometric facial recognition on mobile devices. Initially, this study addresses the differences in quality between images taken with two different camera types. Table 6.1 shows the mean values and the standard deviations calculated for the SLR and smartphone camera in the first and second scenarios. The mean values were calculated from images taken under similar conditions: the movements from the participants were limited since they were sitting on a chair and the artificial light source was the identical in all the instances. Despite the similar context, it appears that images

taken with a smartphone camera reported higher FIQ values, of approximately one unit more than those extracted from the SLR images.

Table 6.1: Mean and standard deviation for the quality metrics in Scenarios 1 and 2.

FIQ metrics	SLR		Smartphone camera	
	Mean	St deviation	Mean	St deviation
Brightness	1.46	0.80	2.56	0.50
Contrast	3.20	0.97	3.63	0.44
GCF	2.69	1.12	3.26	0.71
Blurriness	1.98	0.86	2.53	0.48
Exposure	3.43	1.05	4.46	0.29

A comparison between the two camera types can be seen in Figure 6.1; the histogram shows the distribution for the images taken with the SLR (in green) that appears to range manly between 0 and 3, and the images taken with the smartphone camera (in blue) under similar environmental conditions.

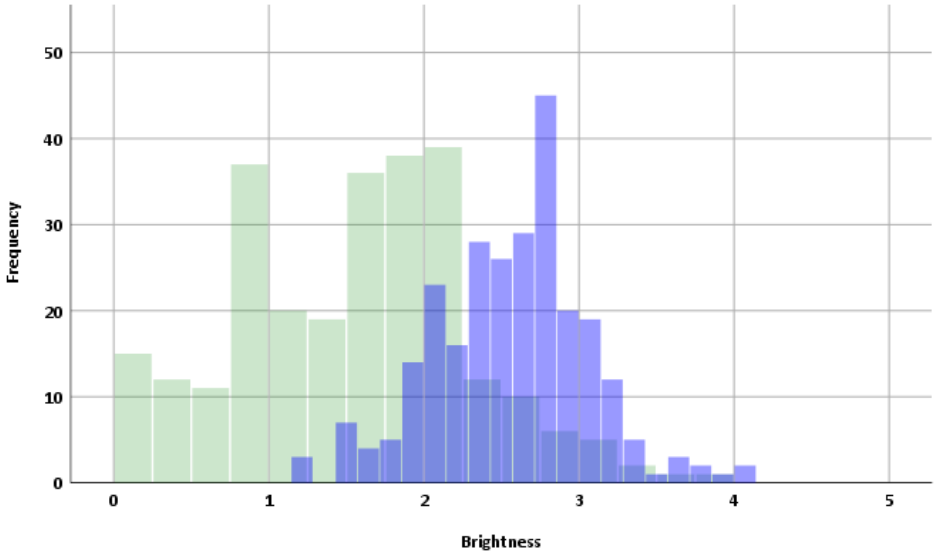


Figure 6.1: Histogram for Brightness for SLR (in green) and smartphone (in blue) images.

Differences are noticeable for all FIQ metrics. The distributions for image Contrast can be seen in Figure 6.2. The trend is similar to Brightness: there is a higher number of SLR images that showed low Contrast values ranging between 0 to 2.5 compared to the images taken with the smartphone. Lower values of GCF (Figure 6.3) were also observed for SLR images when compared to those images taken with the mobile device's camera, although the differences are less visible as the values resulted spread within the whole range. Interestingly, smartphone camera presented Blurriness levels centred between 1 and 4 (Figure 6.4). Images taken with the SLR camera presented values closer to 0 indicating a sharper image. Finally, Exposure histogram can be seen in Figure 6.5. While the Exposure distribution for this quality metric appear skewed to higher values, there is a clear difference between smartphone camera images that range between 3 and 5, compared to the images taken with SLR camera, that presented also lower values.

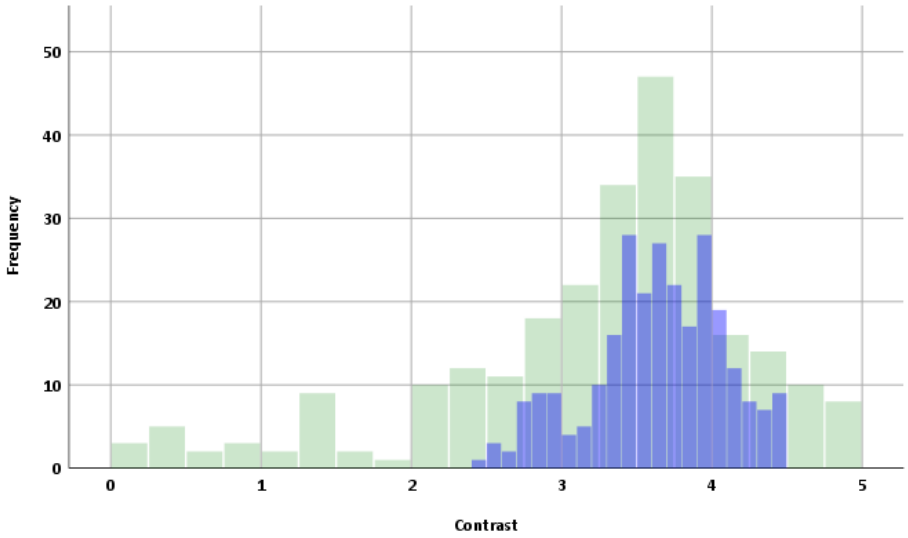


Figure 6.2: Histogram for Contrast for SLR (in green) and smartphone (in blue) images.

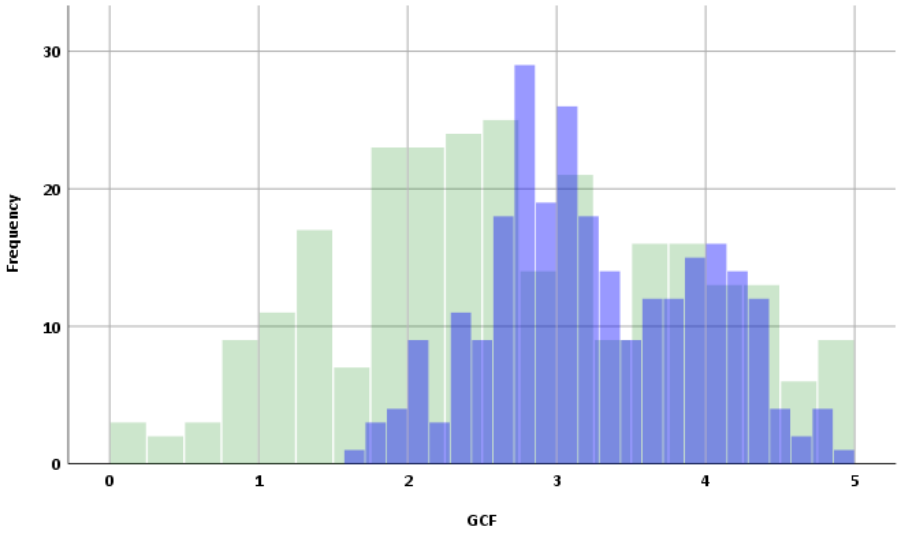


Figure 6.3: Histogram for GCF for the SLR (in green) and the smartphone (in blue) images.

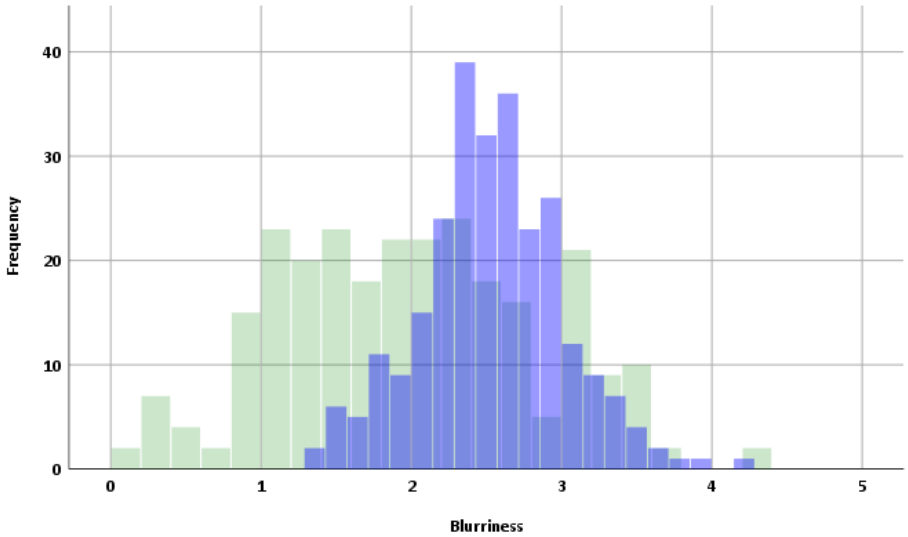


Figure 6.4: Histogram for Blurriness for SLR (in green) and smartphone (in blue) images.

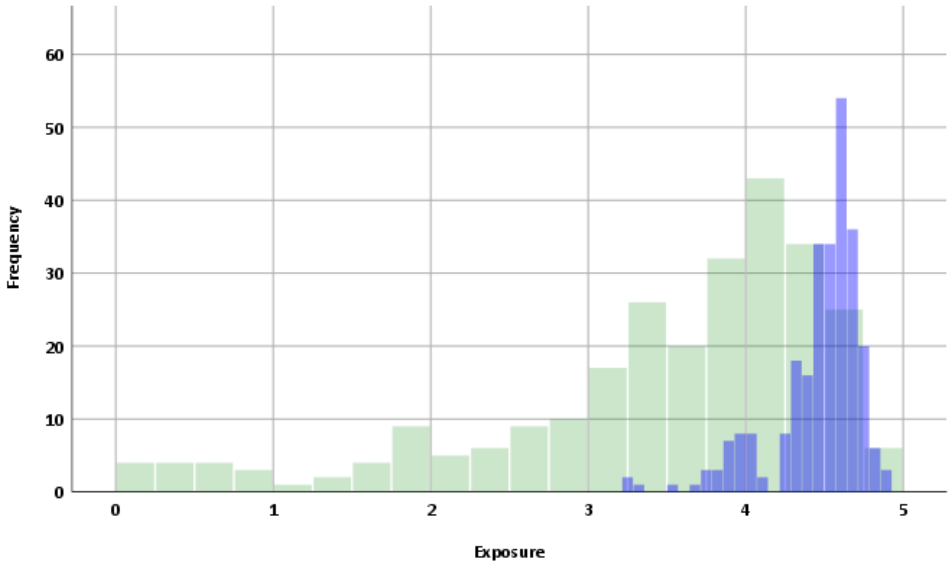


Figure 6.5: Histogram for Exposure for the SLR (in green) and the smartphone (in blue) images.

Despite the images were taken under similar conditions, the camera types reported different FIQ metrics distributions. The results highlight that the quality values required to result in a better verification performance would have different effect depending on the camera type. Furthermore, since the distributions differ from the SLR images even if they were recorded under similar passport enrolment conditions, FIQ requirements should be formulated specifically for a mobile scenario. An independent-samples t-test was performed to check whether these differences have statistical significance. The statistical test compared the mean scores recorded within the two independent groups: images taken with the SLR and images taken with the smartphone. Each of the considered quality metrics reported a significant statistical difference, as shown in Table 6.2.

Table 6.2: Independent-samples t-test significant results comparing camera types.

Quality metrics	t-test	p	Mean difference	Magnitude of the difference	95% Confidence Interval of the Difference	
					Lower	Upper
Brightness	t(541) = -20.1	p < 0.001	-1.1	0.414	-1.21	-0.99
Contrast	t(460) = -7.2	p < 0.001	-0.44	0.083	-0.56	-0.32
GCF	t(546) = -7.4	p < 0.001	-0.57	0.087	-0.72	-0.42
Blurriness	t(511) = -9.7	p < 0.001	-0.55	0.141	-0.66	-0.44
Exposure	t(372) = -16.9	p < 0.001	-1.04	0.333	-1.16	-0.92

The magnitude of the difference was calculated using Eta Squared:

$$\frac{t^2}{t^2 + (N_1 + N_2 - 2)}$$

where t is the t-value from the statistical test and N_1 and N_2 are the number of images in each group. The magnitude was interpreted following the guidelines proposed by Cohen [96] that indicates a small effect for values less than 0.1. From the results,

Brightness and Exposure are the two metrics that reported a higher magnitude of distance between the two groups.

The differences observed for the two camera capture systems highlight that quality requirements specified for images taken using an SLR camera might not have the same applicability for a different specification of camera.

Scenarios 3 and 4 consider images that were taken with only the smartphone camera but under different environmental conditions. Each quality metric was assessed to understand the variations across all images. Brightness presented an approximately normal distribution centred around a mean of 2.53 and a standard deviation of 0.635 as shown in Figure 6.6.

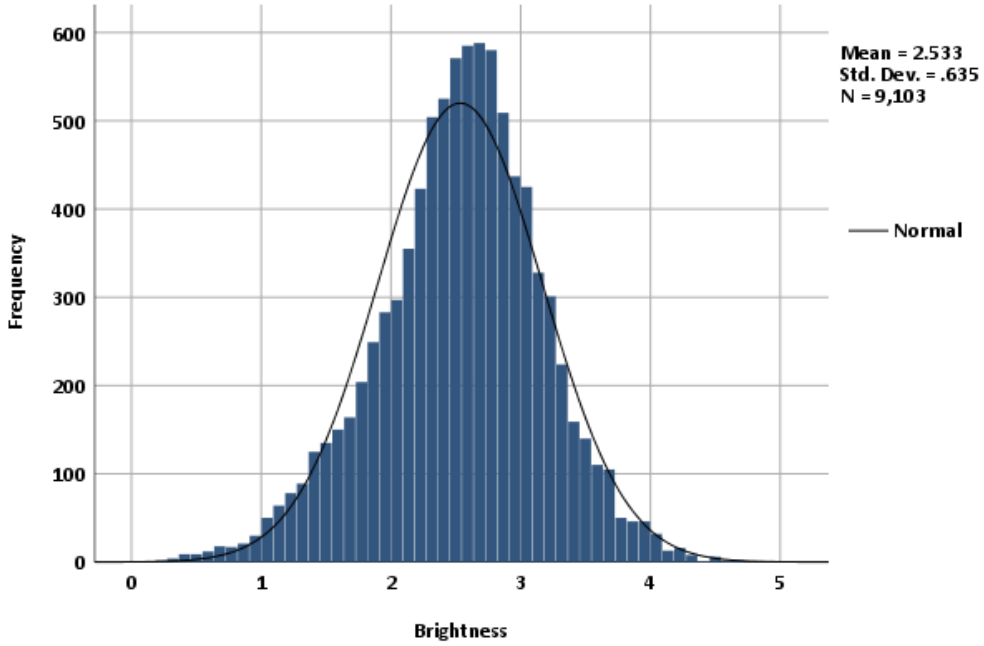


Figure 6.6: Brightness histogram distribution for smartphone images.

It appears from the histogram that across all of the smartphone images, only a few that reported an extreme level of light (1.4% images between 4 and 5) or darkness (0.9% images between 0 and 1) over the facial area.

Similarly, the distribution for Contrast, presented in Figure 6.7, reported an approximately normal distribution centred at 3.5 and presenting a standard deviation of 0.52. Compared to Brightness, the level of Contrast that was reported within the smartphone images is shifted to higher values, with only 1.5% of images presented low Contrast between 0 and 2. Higher values of Contrast indicate that the images' facial area contains significant differences to the background area of the image, but also contains some regions in the face that could present shadows or light points that could influence biometric performance.

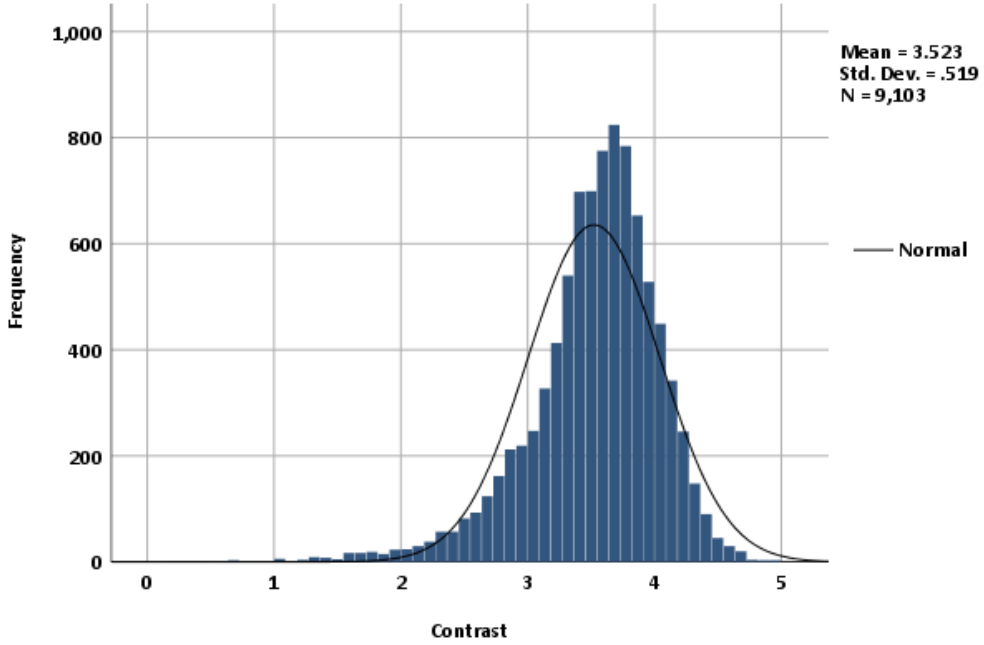


Figure 6.7: Contrast histogram distribution for smartphone images.

The GCF approximates a normal distribution (Figure 6.8) showing that the values from the images were ranging mostly between 1.50 and 4, where a higher level of GCF represents a more detailed local contrast examination of the facial area.

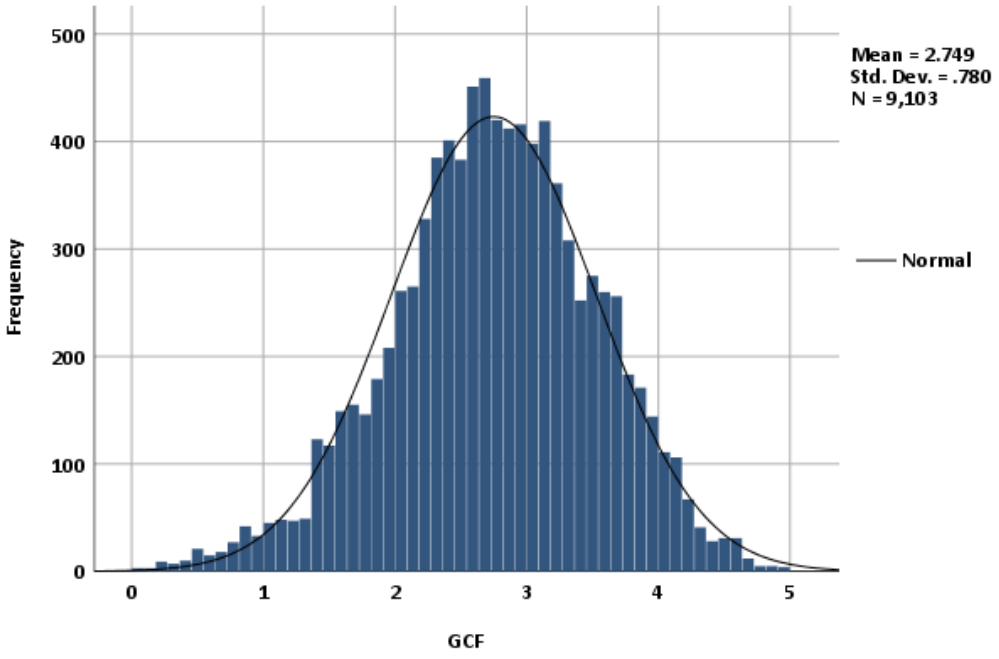


Figure 6.8: Global Contrast Factor histogram distribution for smartphone images.

Similar observations as the ones described for the previous FIQ metrics can be seen for the approximate normal distributions reported when observing Blurriness over the smartphone images (Figure 6.9).

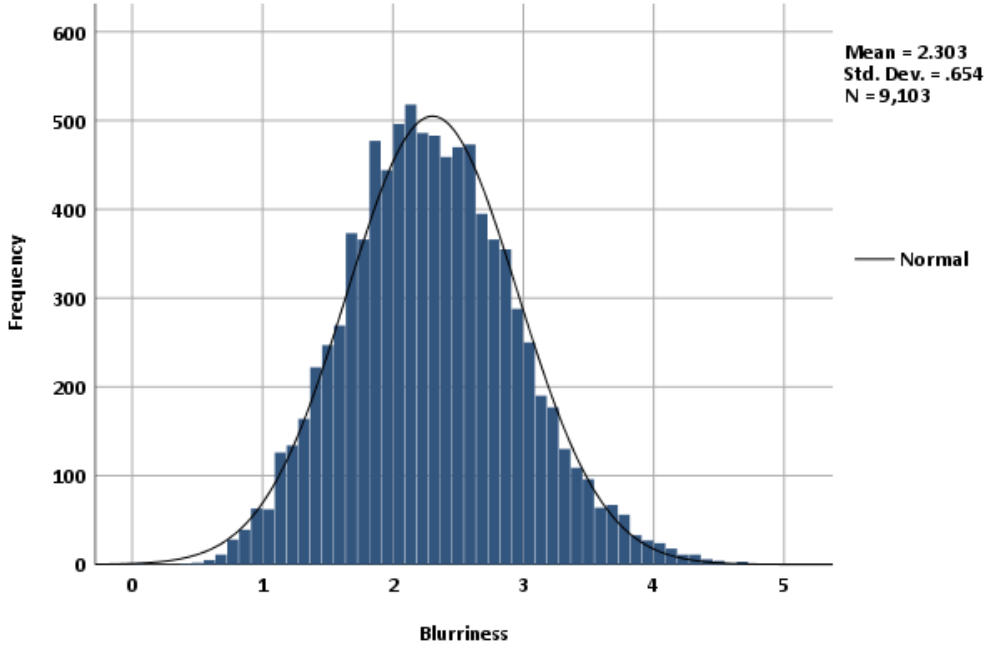


Figure 6.9: Blurriness histogram distribution for smartphone images.

The only distribution that differs from the previous metrics is the one calculated for Exposure. The distribution, in this case, is skewed to higher values with a peak around the mean value at 4.47 and a standard deviation of 0.40 (Figure 6.10).

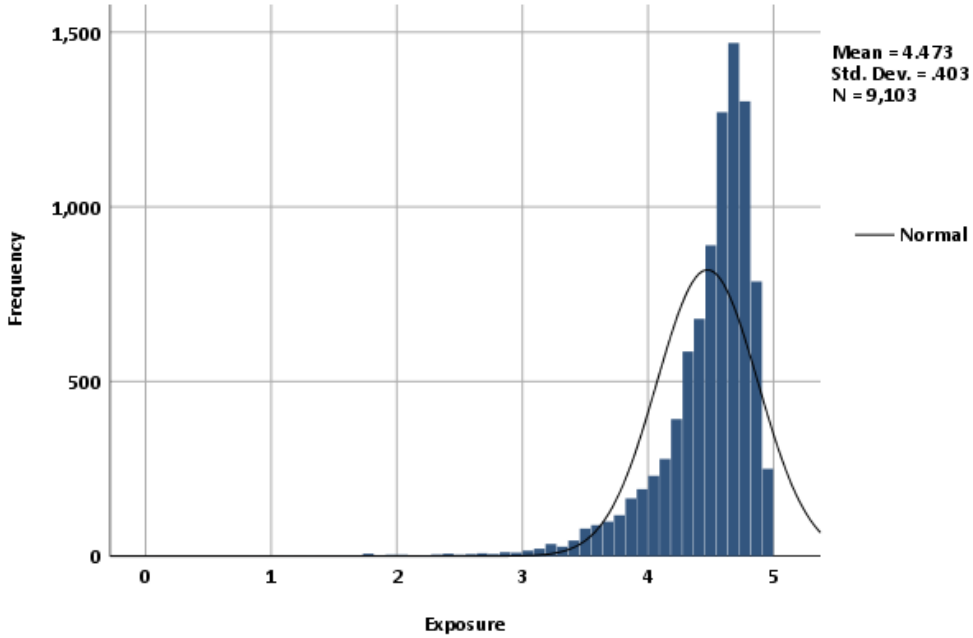


Figure 6.10: Exposure histogram distribution for smartphone images.

It can be observed from Figure 6.10 that smartphone images reported mainly high Exposure values that were recorded between 3 and 5. It could be hypothesised that, since the users took images following the same procedure for three consecutive sessions, habituation will lead to a trend in the FIQ scores obtained, with the scores reported from

the smartphone images taken in the last session would be different than the ones reported in the first one. Figure 6.11 shows a chart where the FIQ metrics seem to have approximately the same mean values across the three sessions.

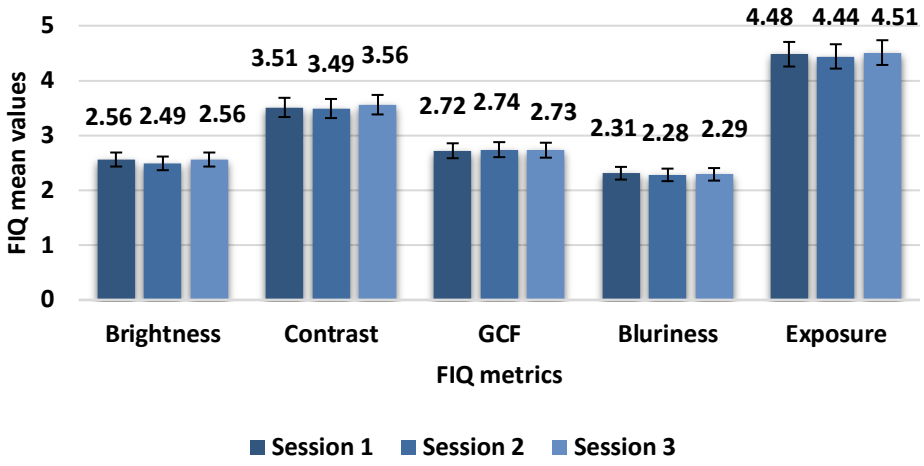


Figure 6.11: Mean values recorded for the FIQ metrics across Sessions 1, 2, and 3.

The images were divided into three different groups depending on which session they were taken. The means for the three independent groups were assessed using a one-way between-groups analysis of variance (ANOVA) to explore whether the small differences observed within the three sessions were statistically significant. Two of the FIQ metrics, GCF ($F(2,5882) = 0.305$, $p = 0.737$) and Blurriness ($F(2,5884) = 1.44$, $p = 0.236$), did not report significant results, but there were significant differences with $p < 0.001$ between the means recorded for Brightness ($F(2,5885) = 11.52$), Contrast ($F(2,5860) = 17.48$) and Exposure ($F(2,5836) = 22.86$). Post-hoc comparisons were performed considering the Tukey's Honest Significant Distance (HSD) as a test to understand for which of the groups the differences occurred. The summary of the multiple comparisons can be seen in Table 6.3. It is interesting to notice from the Table that Brightness, Contrast and Exposure reported entirely different trends among the variables:

1. The images collected in Session 2 have Brightness means significantly different from the ones recorded in Sessions 1 and 3. The images taken in Session 2 reported a significantly lower level of Brightness compared to the other two sessions.
2. The Contrast levels recorded for images taken in Session 3 presented significantly higher means compared to those collected in Sessions 1 and 2.
3. Finally, the Exposure means resulted in significantly different in each of the group comparisons. The level of Exposure calculated for images in Session 1 dropped significantly in Session 2 and reached a higher level in Session 3 that is significantly different from the previous two sessions.

It can be considered from these results that there is not a unique trend followed within the three sessions, but rather that the image quality could be affected by external factors that can cause these differences across the sessions.

Table 6.3: Post Hoc comparisons obteined using the Tukey HSD test.

Dependent Variable	(I) session	(J) session	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Brightness	1	2	0.070*	0.017	.000	.031	.108
		3	-0.001	0.017	.997	-.040	.038
	2	1	-0.070*	0.017	.000	-.108	-.031
		3	-0.071*	0.017	.000	-.110	-.032
	3	1	0.001	0.017	.997	-.038	.040
		2	0.071*	0.017	.000	.032	.110
Contrast	1	2	0.021	0.014	.252	-.010	.053
		3	-0.054*	0.014	.000	-.086	-.022
	2	1	-0.021	0.014	.252	-.053	.010
		3	-0.076*	0.014	.000	-.107	-.044
	3	1	0.054*	0.014	.000	.022	.086
		2	0.076*	0.014	.000	.044	.107
Exposure	1	2	0.040*	0.011	.000	.015	.064
		3	-0.032*	0.011	.007	-.057	-.007
	2	1	-0.040*	0.011	.000	-.064	-.015
		3	-0.072*	0.011	.000	-.097	-.047
	3	1	0.032*	0.011	.007	.007	.057
		2	0.072*	0.011	.000	.047	.097

*The mean difference is significant at the 0.05 level.

6.3 The environmental effect on image quality

The images recorded with the smartphone by the participants outside the experimental laboratory were divided into two independent groups depending on if they were taken indoors (within a building) or outdoors. The FIQ mean values for the two different conditions can be observed in Figure 6.12.

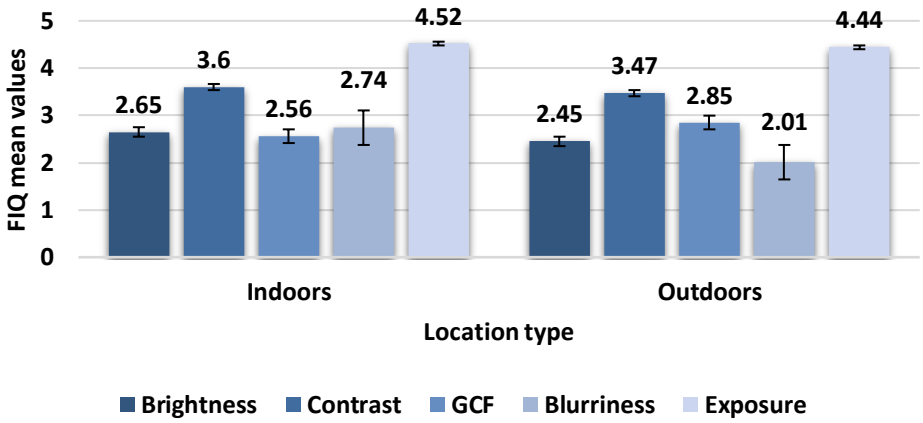


Figure 6.12: FIQ mean values recorded in the two different environment types.

Through a visual examination, it can be observed that all the metrics recorded outdoors, except GCF, appear lower in mean values than those taken indoors. It could be

possible to distinguish the type of locations from the FIQ values recorded if the differences are significant enough. A t-test was performed to assess this hypotesis. The five metrics considered have significant differences with $p < 0.001$ within the two groups of images (indoors and outdoors). A summary of the statistical values is reported in Table 6.4.

Table 6.4: Independent Sample Test for images groups collected in different location types.

Quality metrics	t-test	p	Mean difference	Magnitude of the difference	95% Confidence Interval of the difference	
					Lower	Upper
Brightness	t(7896) = 14.7	p < 0.001	0.20	0.024	0.17	0.22
Contrast	t(8008) = 10.06	p < 0.001	0.11	0.011	0.09	0.13
GCF	t(8197) = -17.69	p < 0.001	-0.28	0.034	-0.32	-0.25
Blurriness	t(7298) = 60.02	p < 0.001	0.73	0.289	0.70	0.75
Exposure	t(8689) = 10.11	p < 0.001	0.08	0.011	0.07	0.10

Blurriness is the FIQ metric that presented the highest magnitude of difference, calculated in Eta Squared, but the other quality metrics reported a small effect size according to the magnitude values. Nevertheless, these encouraging results lead to the assumption that an appropriate threshold on quality could be used to estimate whether the users were in an indoor or outdoor location during the biometric presentation.

With this premise, a logistic regression model was designed containing all the five metrics as predictors. The test reported a significant result for $\chi^2(8) = 52.17$ $p < 0.001$, indicating that the model was able to distinguish the images taken in the indoors and outdoors location groups. The model explained between 37% (Cox & Snell R^2) to 51% (Nagelkerke R^2) of the variance and correctly classified 81% of the cases. As shown in Table 6.5, each of the independent variables contributed to the model. The equation can be read as follows:

$$logit(p) = 9.898 + 0.596 * \text{Brightness} + 0.843 * \text{Contrast} + 1.610 * \text{GCF} - 2.974 * \text{Blurriness} - 2.518 * \text{Exposure}$$

According to the results, GCF is the stronger predictor reporting an Odds Ratio of 5, meaning that for every unit of GCF recorded from the image, the likelihood of that image belonging to those taken outdoors increases a magnitude of 5 times.

Table 6.5: Logistic regression predicting the likelihood that an image was taken indoors or outdoors.

FIQ metrics	B	S.E.	Wald test	df	p	Odds Ratio	95% C.I.for Odds Ratio	
							Lower	Upper
Brightness	0.596	0.086	48.431	1	0.00	1.815	1.534	2.146
Contrast	0.843	0.100	70.795	1	0.00	2.323	1.909	2.827
GCF	1.610	0.060	726.836	1	0.00	5.003	4.450	5.624
Blurriness	-2.974	0.068	1918.303	1	0.00	0.051	0.045	0.058
Exposure	-2.518	0.149	287.425	1	0.00	0.081	0.060	0.108

These results lead to several observations. First of all, that if the system needs to satisfy specific image FIQ requirements for mobile verification, those requirements would need to consider that quality varies depending on the location. Secondly, image quality could provide relevant information to estimate in which location type the facial images were taken using a smartphone camera.

6.4 User interaction

The surrounding environment is not the only element that influences the presentation of facial images. The users themselves are a variable that needs to be taken into account, particularly in a mobile context, for two main reasons: the subjects' physical aspect can influence the FIQ values, and furthermore they are the ones that take the facial images. If, for instance, the person is walking while verifying on the smartphone, the facial image presented for the authentication might appear overly blurred. An analysis was carried out to understand the relationship between the quality metrics calculated over the facial images and user interaction. Results in this Section are presented in terms of subject's demographics, and static and dynamic characteristics of the users.

6.4.1 Demographics

Differences were studied considering the demographic information that could affect the physical aspect of a user. Initially, the images were divided into two groups to assess the effect on FIQ scores with respect to whether a user is a male or a female subject. Differences amongst the means can be observed in Figure 6.13 where the values recorded from female participant's images are higher than those recorded from male subjects.

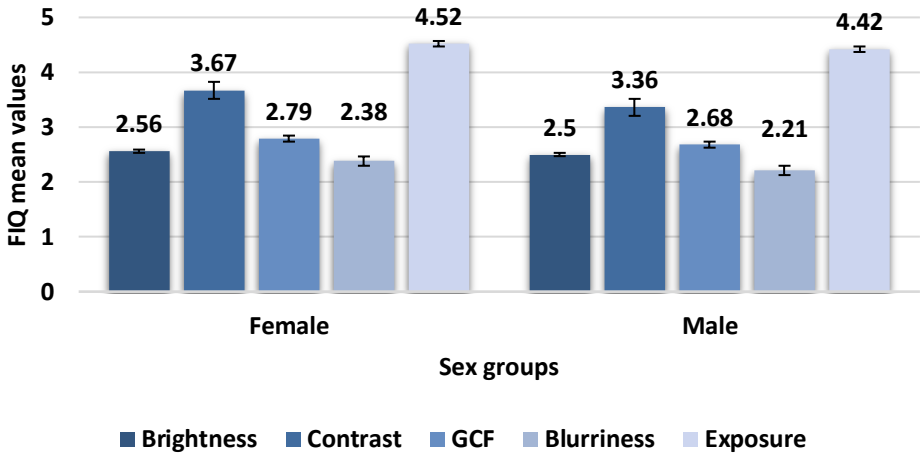


Figure 6.13: Mean differences between sex recorded across all smartphone images.

An independent-sample t-test was performed to assess these differences, and significant results were observed for all the quality metrics. However, the magnitude of the differences between the two groups is small as observed when calculating the values with Eta Squared. For instance, Contrast is the variable that reported the highest

difference between groups, but only 0.8% of the variance in the quality metric is explained by sex for this metric.

Table 6.6: Independent Sample Test results performed on image quality between males and females.

Quality metrics	t-test	p	Mean difference	Magnitude of the difference	95% Confidence Interval of the difference	
					Lower	Upper
Brightness	t(8488) = -4.5	p < 0.001	-0.061	0.002	-0.09	-0.03
Contrast	t(8107) = -28.49	p < 0.001	-0.303	0.084	-0.32	-0.28
GCF	t(8836) = -6.76	p < 0.001	-0.111	0.005	-0.14	-0.08
Blurriness	t(8836) = -12.39	p < 0.001	-0.172	0.017	-0.2	-0.14
Exposure	t(7313) = -12.29	p < 0.001	-0.106	0.017	-0.12	-0.09

Even if the effect size is small, these differences can still have an impact when the size of the dataset increases, especially considering mobile biometric applications that can involve a large population.

The subject's age was considered to check whether there were differences between the FIQ metrics. Using the age groups as defined in Section 4.3.1, statistical differences were found amongst the groups by performing a One-way analysis of variance with post-hoc comparisons using the Tukey HSD test to check which groups presented significant differences. Figure 6.14 shows the mean FIQ metric values for each age group.

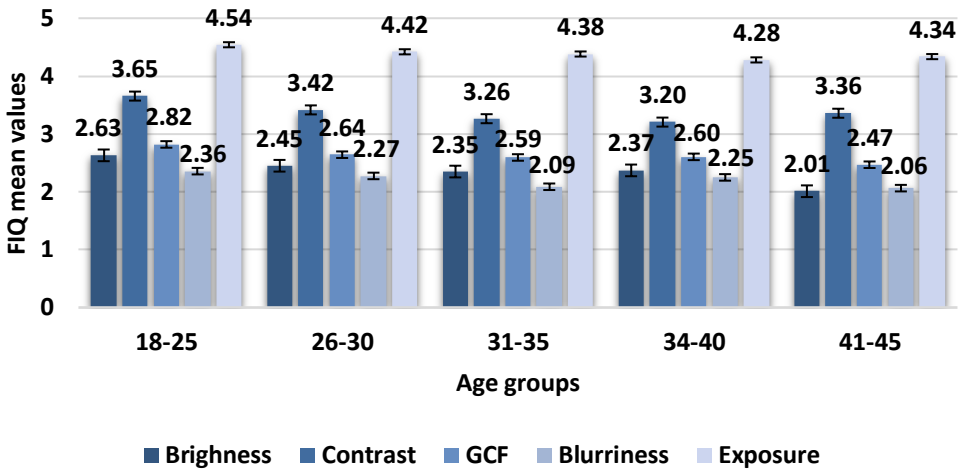


Figure 6.14: Mean differences in quality metrics calculated for each age groups.

Since the FIQ values presented different trends, it is important to address each FIQ metric individually:

- The level of Brightness in the facial images appears to decrease as the participants' age increases. The differences between age groups reported a statistical significance for F(4,916) = 87.25 at p < 0.001 and the post-hoc comparisons confirmed what can be seen from the bar chart: the mean values

decrease significantly from the youngest group to the oldest, with only one exception as the 31-35 age-group presents similar means as the images belonging to the 34-40 age group.

- Contrast levels follow a decreasing trend from the 18-25 age-group to those subjects aged 34-40. The images included in the 41-45 age-group instead presented mean values close to those observed in the 26-30 group. A significance was observed for $F(4,911) = 244.176$ at $p < 0.001$.
- GCF ($F(4,1026) = 58.43$ at $p < 0.001$) presents high values in the 18-25 age-group that are the only images that are significantly different from the other groups.
- The images recorded from participants aged 18-25 reported the highest values of Blurriness, that were significantly different from images within any of the other groups. Participants aged 26-30 and 34-40 reported similar results for Blurriness that were significantly higher than the values reported from the 31-35 and 41-45 age groups ($F(4,931) = 47.23$ at $p < 0.001$).
- Finally, Exposure values ($F(4,896) = 82.83$ at $p < 0.001$) followed a similar trend that was observed for Contrast, where the level of Exposure decreases as the participants' age increases, with the exclusion of the last group, that presents higher values that are close to the 26-30 age-group.

Although differences are dependent on the metric considered, it can be seen that the overall FIQ trend decreases with age. Interestingly, the youngest participants reported the highest means compared to older participants for all the FIQ metrics, including Blurriness. It would be interesting to check which level of image quality refers to higher biometric performance, to understand if the differences reported within the groups also affect the outcome of the verification system. This analysis of quality levels and biometric matching scores is presented in the following Chapter.

Differences can also be observed when comparing the FIQ metrics across different ethnic groups, as shown in Figure 6.15.

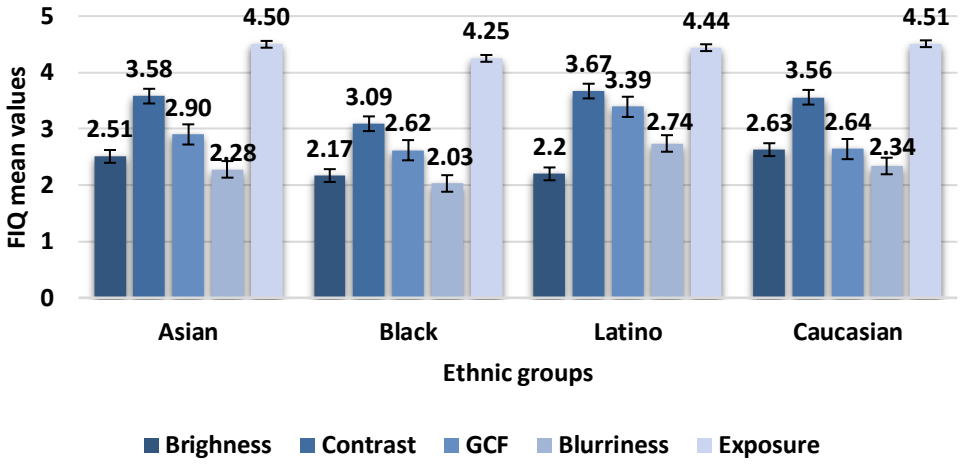


Figure 6.15: Mean differences in quality metrics calculated for each ethnic group.

A One-way ANOVA test was preformed to check whether the differences observed were statistically significant. Post-hoc comparisons using the Tukey test revealed the

relationship between the groups that were significantly different. From the analysis of the differences observed in different ethnic groups, there are some interesting observations. The group that includes the Black or African American subjects reported the lowest values in each metric that were significantly different from the other groups. The images in the Latino group presented significant higher values for GCF and Blurriness compared to the other groups. Asian and Caucasian groups returned approximately the same values for each of the FIQ metrics, with only one exception: there was a significant difference between the two groups in the level of GCF that appears higher for the group of images that belongs to the Asian group. The statistical differences can be seen in Table 6.7.

Table 6.7: One-way ANOVA statistical results reported for each quality metric.

FIQ metrics	One-way between-groups ANOVA
Brightness	$F(3,755) = 186.16$ at $p < 0.001$
Contrast	$F(3,762) = 235.03$ at $p < 0.001$
GCF	$F(3,741) = 128.57$ at $p < 0.001$
Blurriness	$F(3,752) = 110.64$ at $p < 0.001$
Exposure	$F(3,750) = 89.26$ at $p < 0.001$

In summary, the different observations made from this analysis highlight the effect that demographics present on FIQ: it is essential, when rejecting "bad quality" images, that the FIQ requirements take into account the differences that demographic groups can embody (this obviously requires a prior information about the ethnic group of the subject), for instance by adapting thresholds when identifying "poor" quality images.

6.4.2 Static characteristics

The presence of glasses, facial hair and heavy make-up could affect the physical appearance of a facial image and consequently influence the quality values returned. As described in Section 4.4.2.1, the smartphone images were divided into groups depending on these static characteristics to check whether they influence the quality.

The mean results returned for images where users are wearing glasses or heavy make-up tend to show a similar trend in that they returned higher FIQ values for each metric. An example is reported for the mean values of images where the user is wearing glasses (Figure 6.16).

The group of images where a beard was present, conversely, reported lower values than those without a beard. The differences between the groups were also assessed using independent t-tests. The statistical tests confirmed that there were significant differences between the variations in quality and the three characteristics considered, but the magnitude of the differences reported was small (on the scale of 0.01). The presence of glasses in the image affects the mean values recorded for GCF ($t(8234) = 18.67$ for $p < 0.001$).

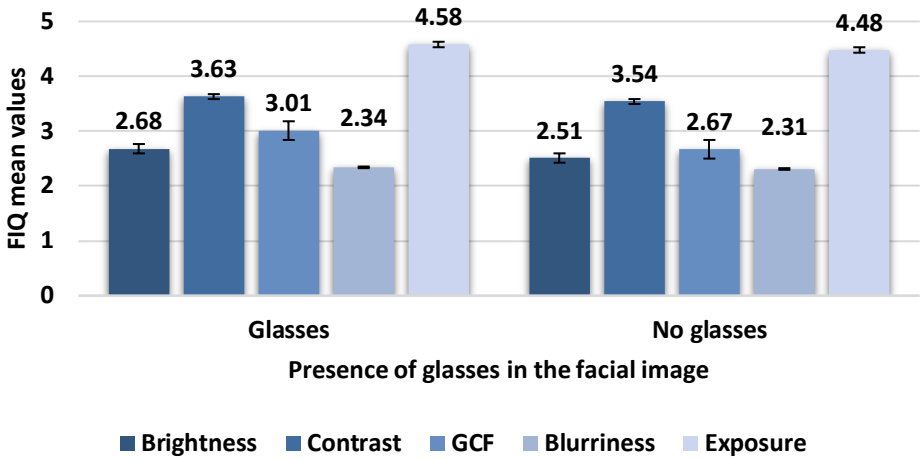


Figure 6.16: Mean differences between participant that were and were not wearing glasses during the acquisition of the facial image.

In Figure 6.17 is reported an example of a person with and without glasses and the respective GCF values. The presence of make-up and a beard instead returned the smallest distance for GCF calculated in the two groups, but they both reported significant differences for the Contrast level: make-up ($t(429) = 21.99$ for $p < 0.001$) when present in the image results in a higher Contrast value, while a beard ($t(2096) = -18.59$ for $p < 0.001$) have lower values compared to those that did not present a beard.

Figure 6.17: Examples of two images from the same participant with (GCF = 3.30) and without (GCF = 3.10) glasses during the biometric presentation.

These interesting results confirmed not only that there are quality metrics that are affected by the presence of these user's characteristics, but also which ones are affected the most so that it could be possible to monitor them in a specific context. For instance, an application could automatically adjust the requirements for GCF knowing in advance that the user will be wearing glasses in the image.

6.4.3 Dynamic characteristics

Quality can also be affected by the users' movements during the presentation. The Android ActivityRecognition API [74] provided an estimation of the type of activity that the user is performing (as described in Section 4.4.2.2). The percentages of images across

the different user's activities can be seen in Figure 6.18. The system could not indicate a specific activity for a considerable part of the database (24.20%), and errors were returned (e.g. participants could have possibly been in a vehicle) which indicates that improvements are still needed to use this information for assessing user interaction.

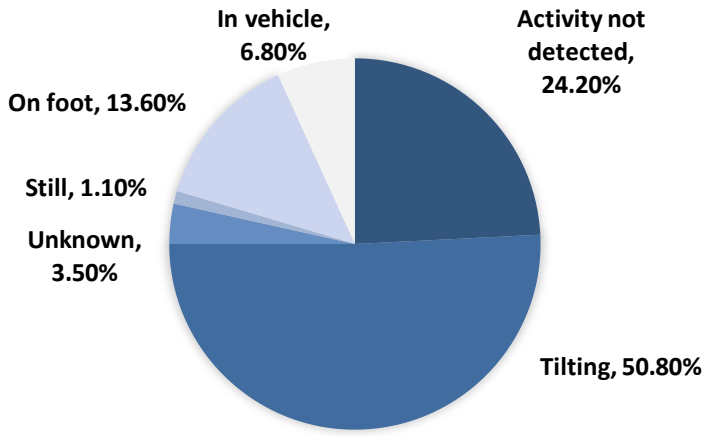


Figure 6.18: Percentages of activity detected and unable to detect from the smartphone images.

Ideally, users should stop walking and remain still while taking the image for the authentication. Assessing the accelerometer data could give an estimate of the smartphone movements, but it was not possible to have a clear distinction between the movements caused by the participants walking during the facial image acquisition and the device's movements caused by the users to position the camera. From the Figure, it can be seen that the images taken from an estimation of still users are only 1.10%, but we cannot rely on the accuracy of this information. Therefore, it was not possible to perform an analysis on the variation of quality metrics according to the estimated activity. Perhaps, future research will be able to use an enhanced version of this Android API to recognise the user activity to assess whether the user is moving the device or is walking during image presentation.

6.5 The camera sensor

In mobile scenarios, one of the main challenges when assessing image quality is the variability of smartphones available in the market: the embedded camera sensors can provide different settings and resolution levels depending on the device. Often the resolution of the front-mounted camera is different from the one located on the rear of the device. The aim of the two sensors is substantially different; the rear camera is mainly considered for capturing landscape images and has been typically designed to ensure better quality and resolution on the newest smartphone models. However, recently, the trend to take self-portrait images, or "Selfies", has contributed to shifting the attention on the front-mounted camera performance [97].

The analysis presented in this study was conducted on facial image quality considering the static characteristics of the Nexus 5 camera used during the data collection. The variations were recorded for each image from the camera ISO and the Light Value. Dynamic camera characteristics were also considered. The accelerometer recordings provided features that allowed the estimation of the magnitude of movements during image capture process. The extraction of accelerometer features for this analysis was presented in Section 4.4.2.5.

6.5.1 Static characteristics

Camera settings can be adjusted to regulate the quality of an image. The settings recorded from the smartphone frontal camera were preset using software settings and were not manually adjusted during the capture process. Hence, setting information was recorded for each image to understand how the settings automatically varied across the database and what relationship they have with the FIQ metric chosen for the analysis. Information was recorded from the camera ISO and Shutter Speed, two main settings that regulate the light exposure of an image, together with Aperture.

The camera ISO setting regulates the light sensitivity of the camera: higher values of ISO result to a "noisy" image but allow pictures to be taken in low-light conditions, for instance in indoors locations; while lower ISO levels result in less "noise". Acknowledging the variations in camera ISO across the images may allow a comparison with other cameras and enable an adjustment in quality depending on the context.

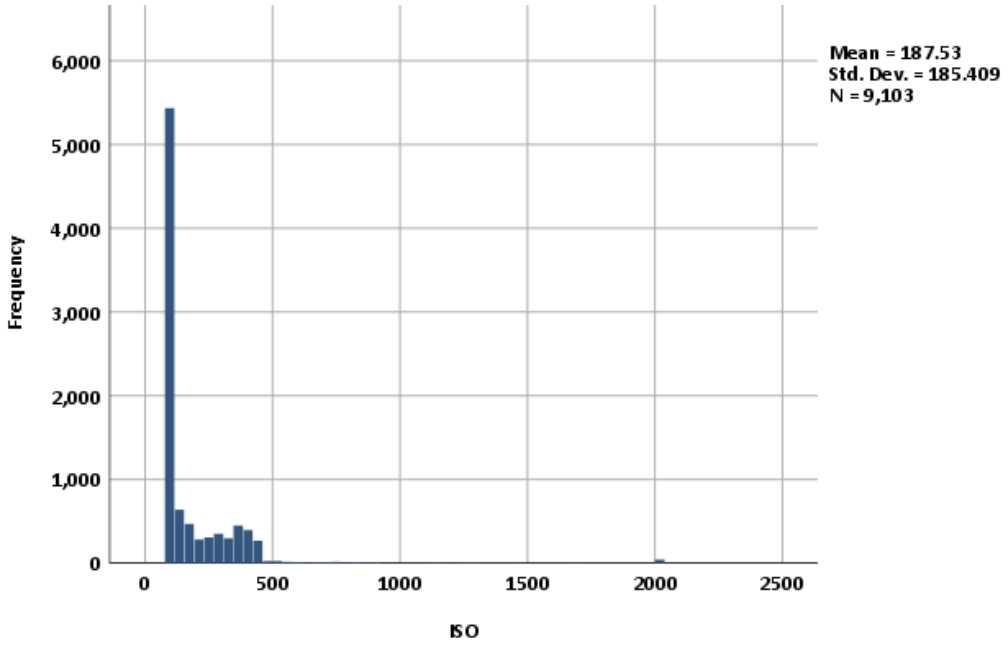


Figure 6.19: Histogram for ISO values across the smartphone images in Scenarios 3 and 4.

Throughout all the smartphone images in the database, camera ISO values were recorded from 100 (low sensitivity) to 2000 (high sensitivity), but the occurrences were higher for images taken with ISO levels between 100 and 110, as shown in Figure 6.19.

The images were divided into five different groups depending on their ISO value (Table 6.8) This division enabled a comparison on the effects that camera ISO has on image quality.

Table 6.8: ISO groups frequencies across the smartphone images.

Camera ISO groups	Percent
ISO 100-200	72.1%
ISO 200-400	20.1%
ISO 400-800	6.8%
ISO 800-1600	0.5%
ISO 1600-2000	0.5%

Both Shutter Speed and Aperture control the amount of light that enters the camera sensor. The Aperture regulates the size of the camera blades and had a fixed value in the database of f/2.9 across all the images. The Shutter Speed indicates the time while the shutter is open, regulating the amount of light entering the sensor. The Shutter Speed values recorded over the database can be assessed within their relationship with the Light Value, that is the reference of the Exposure Value (EV) considering an ISO 100. The relationship between the Exposure Value, Aperture and Shutter Speed can be seen in Figure 6.20.

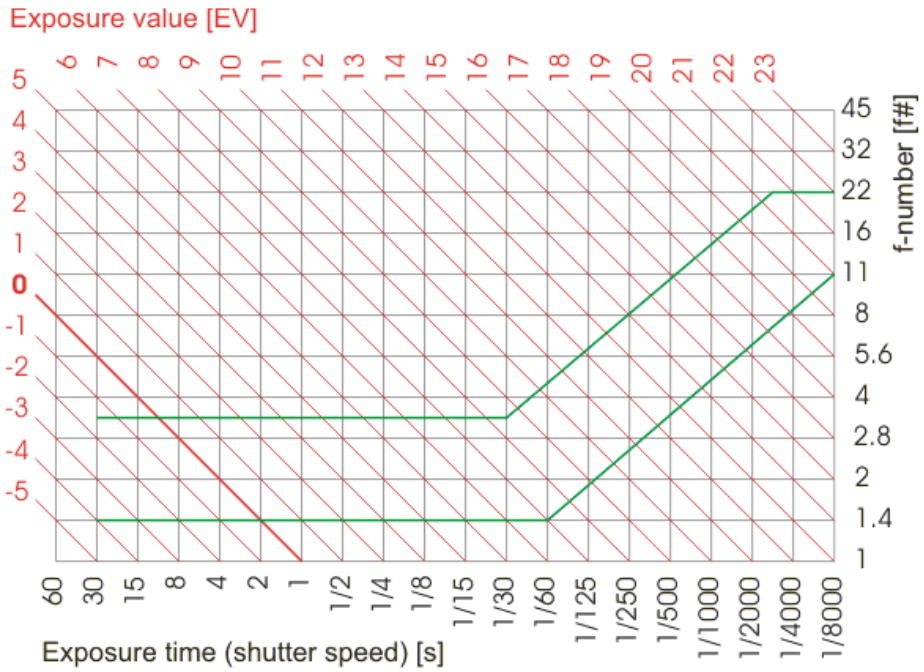


Figure 6.20: Exposure program chart [98].

The Light Value recorded over the database ranged between 2.10 EV and 16.90 EV as can be seen in Figure 6.21.

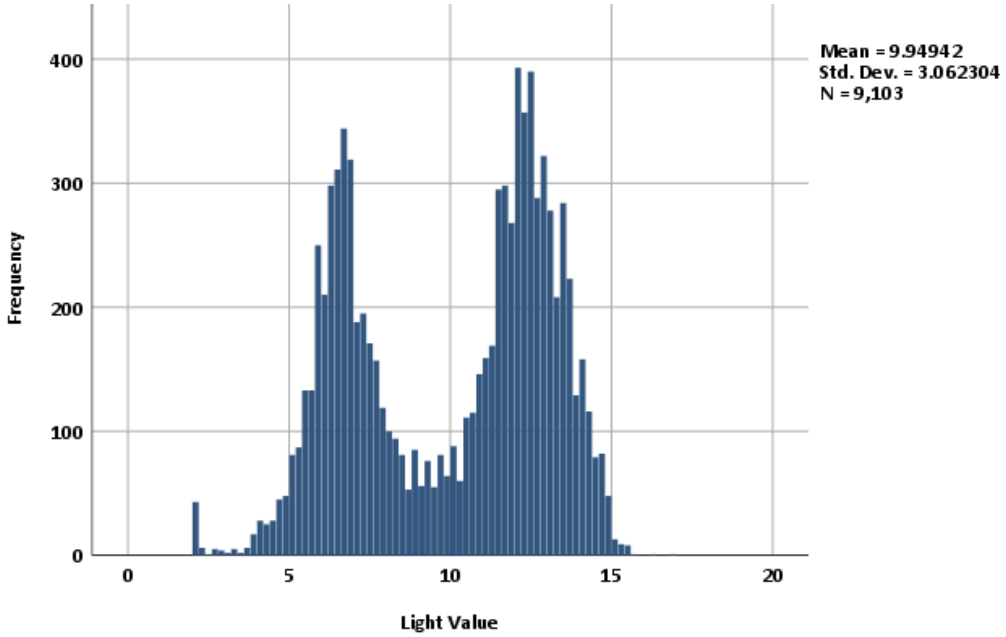


Figure 6.21: Histogram for Light Value across smartphone images.

Faster Shutter Speed corresponds to lower Light Value in the image [99]. The values for image Shutter Speed were also divided into groups, as shown in Table 6.9.

Table 6.9: Light Value [EV] groups frequencies across the smartphone images.

Light Value group	Percent
2.10-5.00	3.2%
5.10-8.00	33.5%
8.10-11.00	14.0%
11.10-14.00	44.8%
14.10-16.90	4.5%

Knowing the values of camera ISO and Shutter Speed needed to obtain the required FIQ level for face verification could lead to an enhancement in the performance of the biometric system.

The variation within the camera ISO groups is shown in Figure 6.22. The quality metric that presented the most extensive variation is Blurriness that substantially increases with the level of camera ISO used. GCF was high when ISO was between 100 and 200, for then decrease until dropping to the lowest values for ISO of 1600 or over. Brightness, Contrast and Exposure also seem affected, mainly resulting in lower values on the highest level of ISO reported.

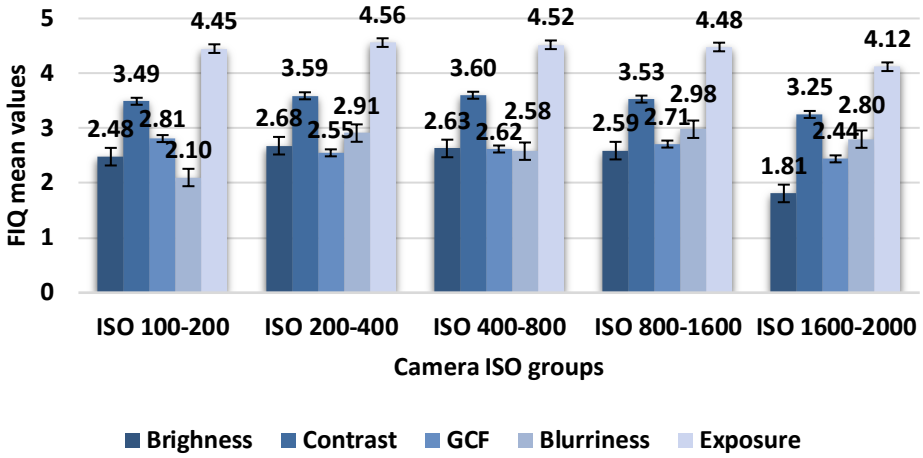


Figure 6.22: Mean quality values recorded on the images depending on the camera ISO groups.

As the Light Value changes, Blurriness is again the metric that reported the highest differences across the images but, contrary of what is seen with camera ISO, the level of Blur decreases with the level of Light Value that is recorded in the image (Figure 6.23). Similarly, GCF increases with Light Value. Image Brightness instead decreases when the Light Value are considered between the 11.10 and over.

The results above describe the relationship that each quality metric has with the two static camera characteristics. Depending on which quality metrics would work better in a specific situation, the level of camera ISO and Light Value (considering the relationship with Aperture and Shutter Speed) could be adjusted to obtained the quality needed.

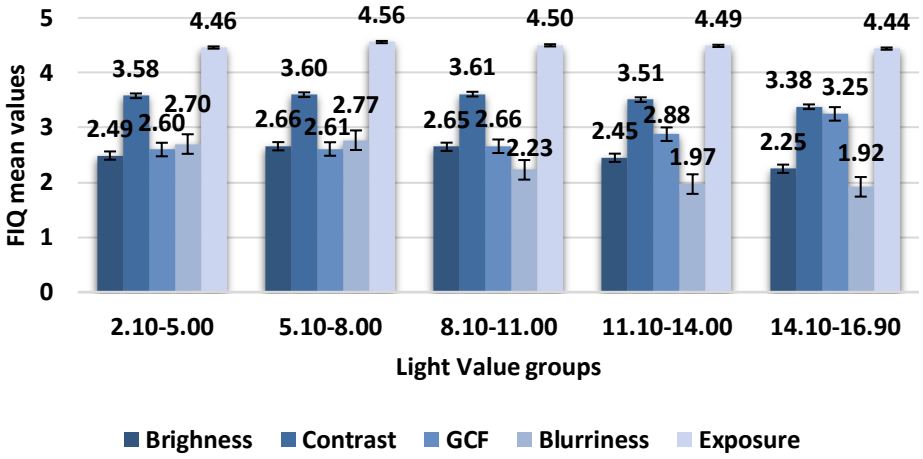


Figure 6.23: FIQ means recorded on the images depending on the Light Value groups.

6.5.2 Dynamic characteristics

In a mobile system, not only the users are moving to pose in front of the camera, but also the camera sensor is moving with them as it is held in the hand. The challenge is to understand which movements recorded from the smartphone were caused by the user

moving in general and which were caused by the user moving the device/camera to acquire the image.

Features were extracted from the magnitude of the accelerometer signals, that were obtained by combining the information from each of the three-axial accelerations recorded in m/s^2 as described in Section 4.4.2.5. The differences between the magnitude of the peaks represents the extent of the movement. Figure 6.24 shows the different in magnitude between two peaks.

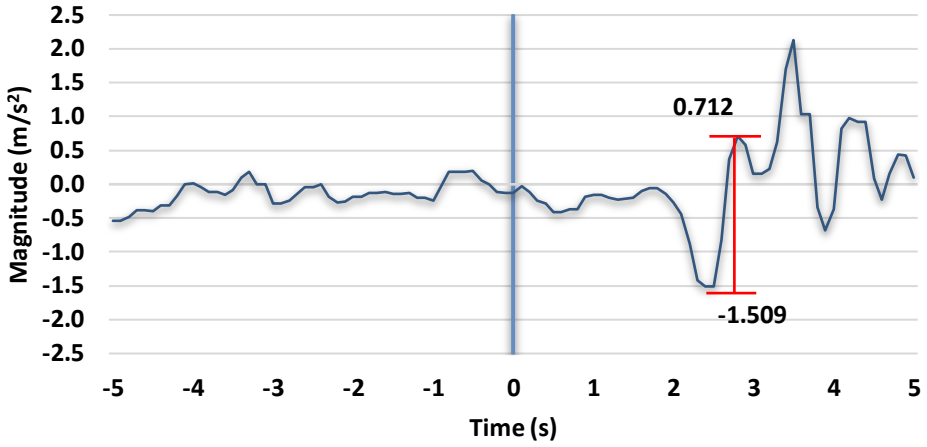


Figure 6.24: Magnitude signal recorded for 5 seconds before and after taking the picture.

Features considered the number of movements that presented a magnitude difference over two different empirically selected thresholds: 1.5 m/s^2 and 2 m/s^2. Occurrences of images that presented peaks over the threshold of 1.5 m/s^2 are shown in Figure 6.25.

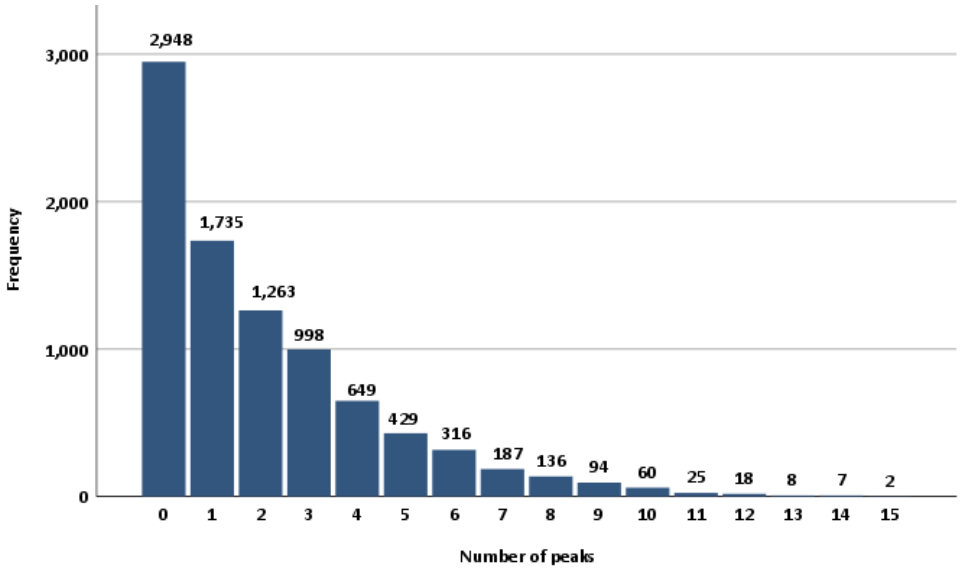


Figure 6.25: Frequency peaks number recorded over the threshold of 1.5 m/s^2.

The images that presented numerous peaks are also the ones where the camera movements were recorded more often within the 5 seconds window in which the image

was taken. Different thresholds were selected considering that the images presenting peaks distances over a higher value of threshold restrict the focus only to those movements that were higher in magnitude. With a threshold of 2 m/s^2 the number of images that did not present a peak over the threshold increased to 5,408 as shown in Figure 6.26.

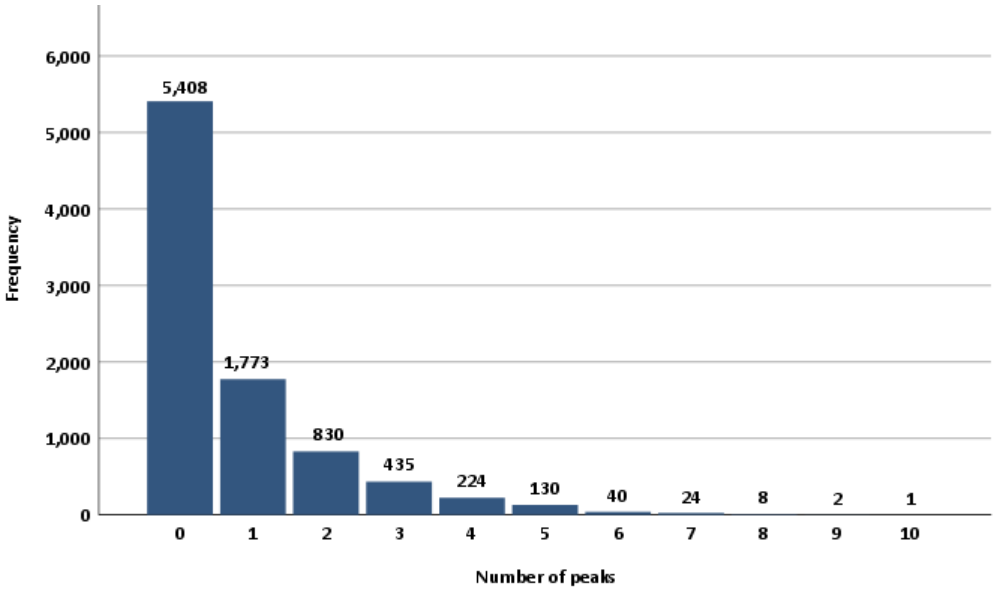


Figure 6.26: Frequency peaks number recorded over the threshold of 2 m/s^2.

The images were divided into five groups, as indicated in Table 6.10, to understand the effect that recorded camera movements have over quality and to compare the magnitude of the movements depending on the selected threshold.

Table 6.10: Frequencies and percentages for peak groups.

Group	Peaks over 2 m/s^2		Peaks over 1.5 m/s^2	
	Frequency	Percent	Frequency	Percent
No peaks	5408	59.4%	2948	32.4%
1 peak	1773	19.5%	1735	19.1%
2 peaks	830	9.1%	1263	13.9%
3 peaks	435	4.8%	998	11.0%
4 or more peaks	657	7.2%	2159	23.7%
Total	9103	100%	9103	100%

According to the results, Contrast and GCF were not affected by movements that were wider than 1.5 m/s^2. Brightness, when no peaks or 1 peak was reported, presented values that are significantly ($F(4,3629) = 4.11$, $p < 0.001$) lower for those recorded with 2 peaks, and decreased when 3 or more peaks were recorded. The level of Blurriness increases significantly ($F(4,3624) = 23$, $p < 0.001$) with the number of peaks, while the level of Exposure decreases significantly ($F(4,3617) = 4.02$, $p < 0.001$) with the number of peaks.

When considering larger movements, all the quality metrics were significantly affected (Table 6.11). Blurriness is the only metric that resulted in a different trend since the values

reported from the images significantly increase with the number of peaks. The remaining metrics all presented a similar trend: the groups of images that presented one, or no peaks have similar values that decrease when the number of detected peaks is between 2 or 3. Finally, when the images recorded 4 or more peaks, the FIQ values increased to maximum values recorded.

Table 6.11: One-way ANOVA statistical results reported for each quality metric.

FIQ metrics	One-way between-groups ANOVA
Brightness	$F(4,1636) = 6.82$ at $p < 0.001$
Contrast	$F(4,1628) = 4.08$ at $p < 0.001$
GCF	$F(4,1656) = 7.99$ at $p < 0.001$
Blurriness	$F(4,1649) = 18.05$ at $p < 0.001$
Exposure	$F(4,1611) = 2.87$ at $p < 0.001$

The observations made in this analysis are important because they show a first practical approach to detect user interaction with the accelerometer data. Furthermore, knowing how the FIQ metrics vary depending on the movement recorded provides valid information that can be useful to estimate the "noise" on an image and predict the biometric outcome, or could be used to ask the user for a second facial presentation.

6.6 Quality assessment: overall observations

Assessing quality on facial images taken with smartphones is not an easy task. The number of variations that the surrounding environment, the user and the camera can add to the system is what makes this task a challenge. The aim of this analysis was to detect the effect that these variations have on FIQ metrics and define general observations that need to be considered when adopting face verification on a mobile device.

From an analysis of the results obtained when comparing two different camera types, it was observed that the specific quality requirements adopted for passport scenarios need to be adapted for mobile devices. When smartphone images were taken in an environment similar to the one recommended for passport image acquisition, the values for quality recorded were significantly different from those with the SLR. Results also highlight that quality should be assessed from different types of smartphones camera to ensure the best FIQ requirements for mobile devices.

The quality metrics assessed across all smartphone images reported approximate normal distributions with the only exception of Exposure that presented a distribution that was skewed towards the highest values. Also, the images reported different trends when compared within the three sessions, especially Brightness, Contrast and Exposure. This also confirmed the previous results obtained for Face Detection, where it was already observed that the different variations provided from the environmental types do not allow improvements within the short time-window considered between the sessions.

There were significant differences from the two different location types observed; Blurriness, in particular, is the FIQ metric that reported the highest values for images that were taken indoors. The analysis demonstrated that it could be possible to distinguish if an image was taken indoors or outdoors using the quality values.

Users affect the image quality in several different ways. This study underlines the importance to consider demographic and user characteristics when designing the quality requirements for facial images taken with smartphones as they presented differences over the images. For instance, the presence of glasses in the image resulted in higher GCF values. The quality requirements for GCF in this specific case should be adjusted considering that this characteristic can be present in the image. Dynamic characteristics instead did not report significant results because the system was not able to accurately detect the user's activity. Perhaps further analysis on the accelerometer could provide in future research an enhanced way to assess the user's movements from those recorded from the camera device.

The camera characteristics were divided between groups so that they could be used to analyse the effect they have on both quality and consequently the performance of the system. Camera ISO and Light Value were examined across the database to understand how they vary depending on the situation. Brightness and Contrast decreased with the increase of ISO and Light Value present in the image, while the level of Blurriness decreased substantially.

Camera movements were also analysed using the accelerometer features. Brightness, Blurriness and Exposure were affected by the camera movements that had a magnitude of more than 1.5 m/s^2. Blurriness in particular increased while Brightness and Exposure decreases. When considering higher movements of peaks over 2 m/s^2 all the quality metrics were affected.

This study illustrates new considerations that should be addressed in future research: the same quality metrics should be investigated using images from different smartphone cameras to identify the best FIQ requirements for facial verification on mobile devices with respect to biometric matching performance.

Finally, after assessing how the FIQ scores vary across the database, this study will focus on how the biometric matching scores relates to the variations over the images, so that it could be possible to identify the best FIQ requirements to obtain enhanced verification performance. This is presented in the following Chapter.

The Verification Process: Biometric Performance

7.1 Introduction

One of the main challenges when assessing mobile biometric performance is to identify and consider the variables introduced by the user and the surrounding environment during facial image presentation. In particular, it should be considered that not only the verification but also the enrolment can take place under unpredictable conditions. Users are unsupervised when interacting with their device, and they can decide to adopt a biometric security system at any time.

The work presented in this Chapter considers four types of enrolment (as described in Section 4.4.3.7) to assess the different conditions that could occur in real-life applications and to compare them with a passport enrolment scenario. The scenarios were selected as follows:

- Enrolment 1 (E1): SLR images taken in the experimental room;
- Enrolment 2 (E2): images taken with the smartphone camera by the users, also in the experimental room;
- Enrolment 3 (E3): images taken with the smartphone in an indoor location;
- Enrolment 4 (E4): images taken with the smartphone in an outdoor location.

The first two enrolment scenarios (E1 and E2) allow a comparison between the two camera sensors when the enrolment occurs under similar conditions as those considered for a passport image. The last two scenarios (E3 and E4) allow an investigation of the different location types: indoors or outdoors.

Each enrolment scenario comprises of 5 separate images. The enrolment images used for each participant were the same for both the verification algorithms considered to assess the biometric performance: Neurotechnology VeriLook 10.0 SDK [80] and Face_recognition [85]. The two algorithms were chosen to compare the obtained results providing general observations that could be applied to any verification system. The verification dataset (7,914) was also the same for both algorithms, and it was formed by those images in Scenarios 3 and 4 that were not utilised for the enrolment (E3 and E4). The verification algorithms reported a binary result, that was recorded as a “Successful” or “Failed” verification, and a matching score, both presented as an average of the comparisons between each verification image and the five used for the enrolment.

This Chapter presents the results obtained for verification performance across the different Scenarios, followed by an assessment of the effect of different variations

introduced to the biometric system by the environment and the user interaction. Finally, the obtained biometric results are presented considering the quality assessment.

7.2 Biometric verification across scenarios

A first analysis between the different enrolment scenarios can be seen in Table 7.1, indicating the percentages of "Successful" and "Failed" verification attempts calculated from the two verification systems considered. The acceptance threshold for both devices was set to the algorithms default values.

Table 7.1: Biometric binary results recorded for the four enrolment scenarios.

Verification system	Binary results	E1	E2	E3	E4
VeriLook 10.0	Successful	91%	96.7%	98.2%	97.1%
	Failed	9%	3.3%	1.8%	2.9%
Face_recognition	Successful	99.8%	99.9%	99.9%	99.8%
	Failed	0.2%	0.1%	0.1%	0.2%

It can be seen that the two verification systems reported different results: False Reject Rate (FRR) is higher for VeriLook 10.0 in all the scenarios compared to Face_recognition. The difference observed is probably caused by the use of two verification algorithms of different nature.

The outliers were mostly detected in the first two types of enrolments, where the image acquisition took place in the experimental laboratory. Face_verification is an algorithm more adaptable to the mobility of a smartphone than the Neurotechnology VeriLook 10.0, mostly known for being used in automated boarding gates. Nevertheless, the outliers observed did not differ substantially from the overall trend. Therefore, also considering the size of the database, we decided to include them in the analysis, bearing in mind that the results could be slightly skewed and opting for a statistical solution more resilient when dealing with outliers. The significant observations and the results described in the analysis, despite the difference observed between the algorithms, can be made for both biometric systems since they both follow tendentially the same trend. First of all, the percentages of "Failed" verifications are higher for the first type of enrolment (E1), and this could be seen in particular for VeriLook 10.0, while the images that were compared to E3 were those that reported a higher acceptance rate.

The matching scores were assessed to understand the relationship between the enrolment scenarios and the FRRs. The matching scores were normalised to range between 0 and 1 to allow a comparison between the algorithms on the same scale, where 0 means no match and 1 means close match. Table 7.2 shows the descriptive statistics depending on the type of enrolment scenario.

Table 7.2: Matching scores descriptive statistics for each biometric verification system across enrolment scenarios.

Enrolment scenario	VeriLook 10.0				Face_recognition			
	Min	Max	Mean	Std. Deviation	Min	Max	Mean	Std. Deviation
E1	0.01	0.32	0.103	0.035	0.10	0.82	0.623	0.065
E2	0.01	0.49	0.123	0.047	0.13	0.84	0.652	0.066
E3	0.01	0.72	0.136	0.054	0.12	0.92	0.681	0.063
E4	0.01	1.00	0.151	0.084	0.14	0.92	0.682	0.075

From these results, it can be noted that, even if normalised to the same scale, the matching scores calculated from the two algorithms considered presented a substantial difference. The values vary because the verification systems have different methods to assess the similarity between the enrolment and verification image. Nevertheless, an analysis over these scores allows the determination of which external variables influenced the performance across verification systems and to formulate general observations. For instance, the values for E1 are lower than those for the other enrolment scenarios when considering both algorithms. This difference could either indicate a lowering of the performance due to the use of two different camera sensors or due to the effect of assessing verification images taken in an environment that was different from the "controlled" enrolment. It could be identified that the closer the conditions in which the enrolment and the verification occur, the higher the similarity scores.

The matching scores were observed across the three experimental sessions, to check whether there were improvements between the first time the users were taking the images and the following two sessions. The mean values calculated with VeriLook 10.0 for each enrolment types are indicated in Figure 7.1.

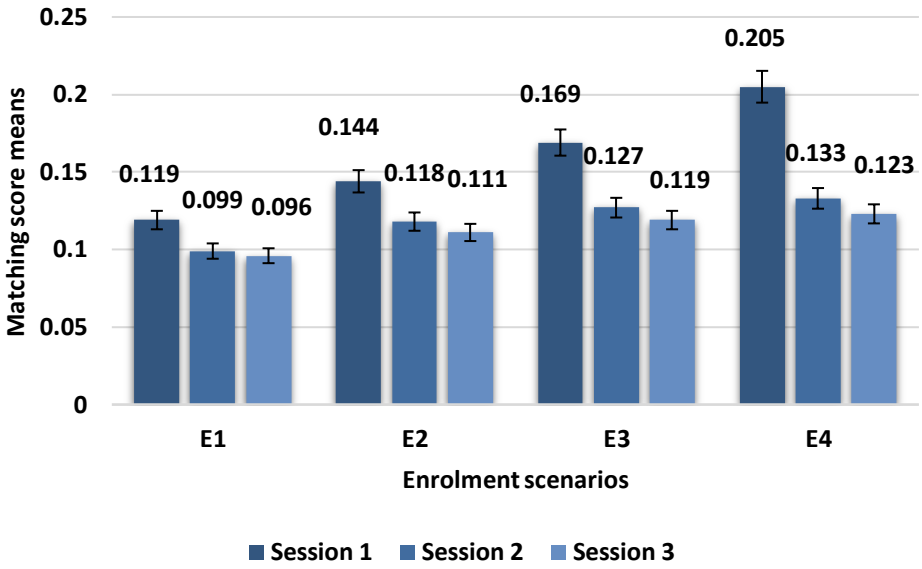


Figure 7.1: Matching score means calculated with VeriLook 10.0 across the three sessions.

A trend can be observed between the matching scores obtained in the first session for each enrolment scenarios and those obtained in the following two sessions. The performance decreases within the sessions, especially between the first and the second sessions that the users participated in within the data collection.

Similar observations can be seen for Face_verification. The mean values for the matching scores obtained with the different scenarios of enrolment can be seen in Figure 7.2. Although the difference between Session 1 and the two following sessions is less evident, it is still possible to note that the performance of the system decreases within the sessions.

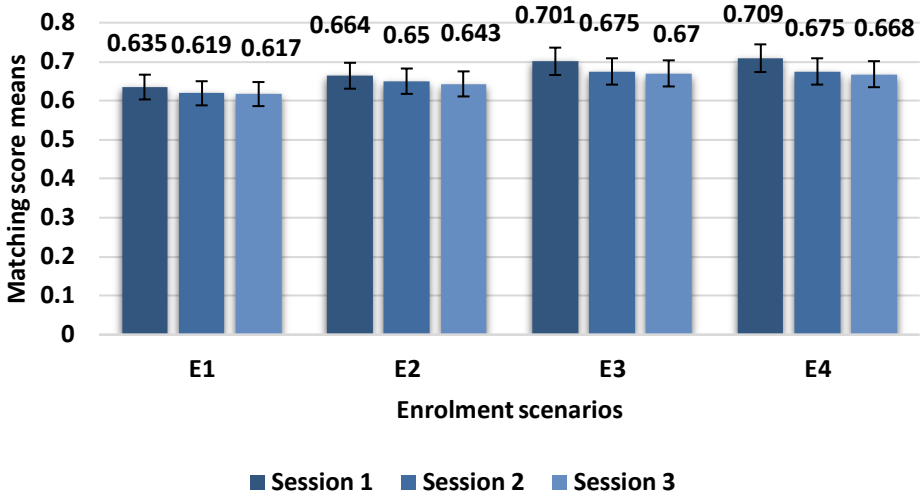


Figure 7.2: Matching score means calculated with Face_recognition across the three sessions.

The FRRs obtained for each session from the two verification systems confirmed these observations, although it could be noted that for E2 and E3 the percentage of "Failed" verification is lower for Session 3 compared to the previous Session 2 (Table 7.3).

Table 7.3: False Reject Rate (FRR) across the three sessions.

Enrolment scenarios	VeriLook 10.0			Face_recognition		
	Session 1	Session 2	Session 3	Session 1	Session 2	Session 3
E1	6.1%	8.9%	11.6%	0.3%	0.1%	0.1%
E2	2.2%	4.3%	3.1%	0.1%	0.1%	0.1%
E3	1%	2.2%	1.9%	0%	0.1%	0.1%
E4	1.6%	2.5%	4.2%	0%	0.2%	0.2%

The differences observed could be affected by the fact that the subjects were performing a repetitive task over a long time. Furthermore, feedback was not provided from the system, as the primary goal for the participants was to take the images for the verifications and not to verify themselves on the device. Participants did not have the

chance to understand how to improve the sample presented over time, and this could result in a lowering of performance over the sessions; it might be that this behaviour is more perceived when using the VeriLook 10.0 method for verification.

These results underline the importance of assessing usability and donation feedback mechanisms on mobile facial verification, since the attitude towards the technology appeared to change within the data collection, suggesting to address these aspects in future research.

7.3 The environmental effect on performance

The surrounding environment, in the particular context of mobile devices, should be assessed considering both the enrolment and verification conditions. For this analysis, the verification dataset was divided into two main groups depending on which location type (indoors or outdoors) the verification image was taken.

The mean values for the matching scores calculated with VeriLook 10.0 are presented in Figure 7.3. The scores presented similar differences to those observed in Section 7.2, indicating that E3 and E4 are the two enrolment scenarios that reported higher matching scores. However, clear differences between verification images taken indoors or outdoors are not evident.

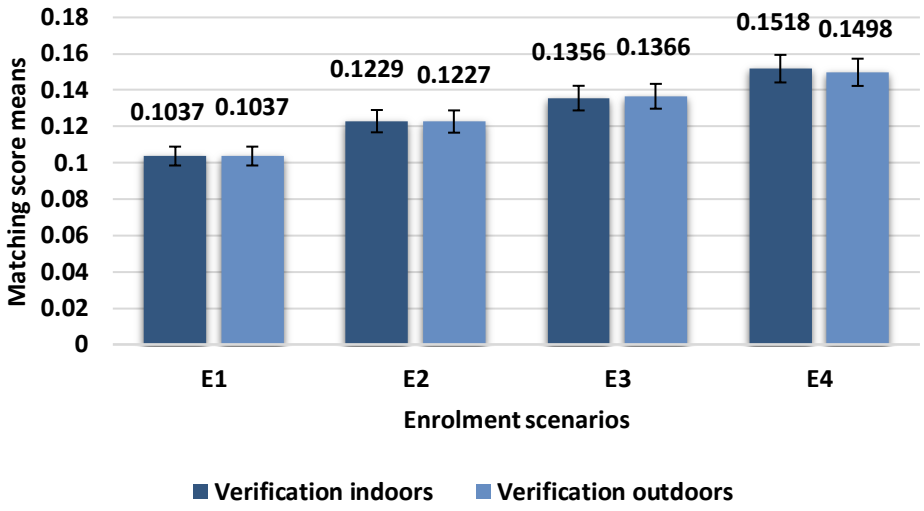


Figure 7.3: Matching score means obtained with VeriLook 10.0 according to the verification location.

The differences between the verification locations are even less evident when observing the mean values obtained using the Face_recognition algorithm (Figure 7.4). The verification location did not appear to have an influence on the results within the same enrolment type.

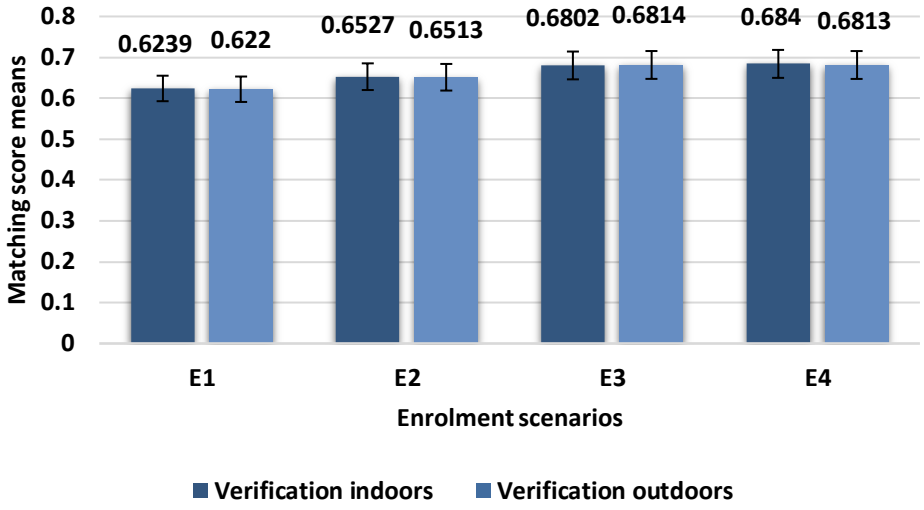


Figure 7.4: Matching score means obtained with Face_recognition according to the verification location.

Statistical tests were also performed to check whether there were significant differences that were not visible from the charts, but the t-test confirmed that no differences were found according to the verification location. However, the closer the environmental condition to which both enrolment and verification occurred, the higher the matching scores. These observations are confirmed for both verification systems considered.

When comparing the FFRs within the verification locations, the two algorithms reported different results. There is a higher number of "Failed" verifications recorded by VeriLook 10.0 when the verification occurred in outdoor locations, as can be seen in Table 7.4, while the opposite case is observed for Face_recognition. For both algorithms, it seems that E3, the enrolment scenarios that included indoors images, could be the best type of enrolment, as it recorded the lower FRR for both verification location types.

Table 7.4: FRRs depending on the locations in which the verification images were taken.

Verification systems	Verification location types	E1	E2	E3	E4
VeriLook 10.0	Indoors	7.9%	3.2%	1.5%	2.4%
	Outdoors	9.8%	3.3%	2%	3.2%
Face_recognition	Indoors	0.3%	0.3%	0.2%	0.3%
	Outdoors	0.1%	0.01%	0.01%	0.1%

From this analysis, it is possible to make several important observations for mobile facial verification. First of all, the assessment of the environmental conditions in which the authentication occurs should include a consideration of the enrolment location scenario. Higher matching scores can be obtained if similar location types were considered for both enrolment and verification. Secondly, it appears that E3 is the type of

enrolment more resilient to the environmental differences added to the system for both algorithms.

7.4 User interaction

Since users provide their biometric characteristic to the mobile device, they are an integral part of the verification process, and as such, they influence the system performance. The work presented in this Section includes an assessment of verification performance with respect to user interaction characteristics. The analysis considers demographics, and static and dynamic characteristics of the subjects, selected as presented in Section 4.4.2.

An investigation of performance was also considered according to the participants' experience and opinions expressed in the questionnaires at the end of each data collection session. However, it was not possible to identify a significant relationship between the variables. The use of questionnaire was designed to help understand the participants perception of their experience during the data collection, but another approach could be considered to directly address this research, namely an open-question interview which would be beneficial for this analysis. This approach will be considered in future research with the lessons learned during the data collection.

7.4.1 Demographics

Biometric verification performance was studied depending on the demographic groups described in Section 4.3.1. Initially, groups of images were compared considering the subject's sex. The means of the normalised matching scores calculated for both verification algorithms are shown in Figure 7.5.

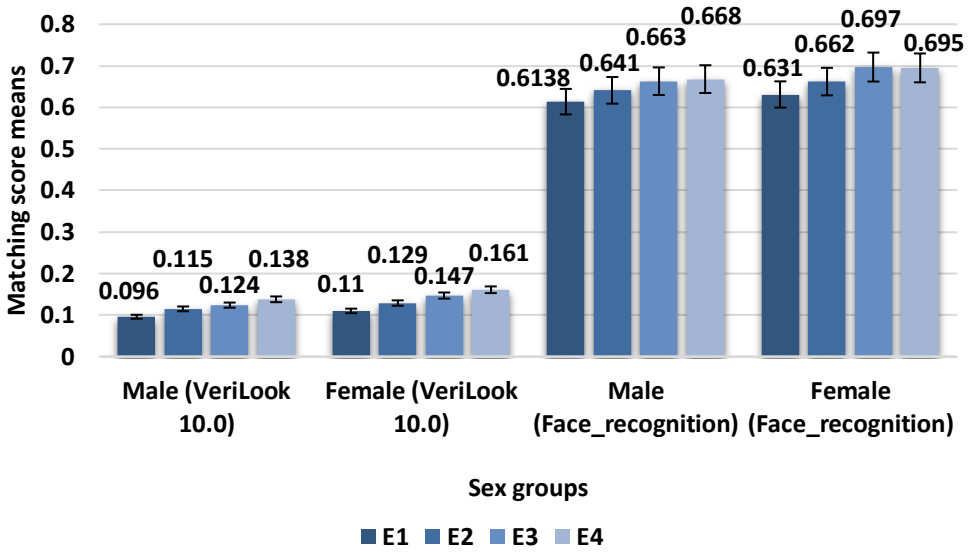


Figure 7.5: Matching score means according to sex groups.

Results showed that there were small differences between the sex groups but that generally male subjects tend to have lower values of verification matching scores compared to female subjects. An independent t-test was performed to check the significance of these differences statistically. The results are shown in Table 7.5, and despite reporting significant results between the groups, the magnitude of the differences, calculated in Eta Squared, was not high.

Table 7.5: Statistical independent t-test performed to compare sex groups.

Verification systems	Enrolment scenarios	t-test	p	Mean difference	Magnitude of the difference
VeriLook 10.0	**E1**	t(7537) = -17.26	p < 0.001	-0.014	0.038
	E2	t(7536) = -13.97	p < 0.001	-0.015	0.025
	E3	t(7537) = -19.42	p < 0.001	-0.023	0.047
	E4	t(7537) = -12.26	p < 0.001	-0.023	0.020
Face_recognition	**E1**	t(7885) = -11.82	p < 0.001	-0.017	0.017
	E2	t(7912) = -13.97	p < 0.001	-0.020	0.024
	E3	t(7909) = -25.36	p < 0.001	-0.034	0.075
	E4	t(7006) = -15.63	p < 0.001	-0.026	0.030

The enrolment scenario E3 reported the highest magnitude of difference across the two considered systems. These differences can have a problematic impact if a larger population of subjects is considered.

A similar analysis was conducted to assess the differences between verification matching scores and age groups. The mean values calculated by VeriLook 10.0 for each enrolment type can be seen in Figure 7.6.

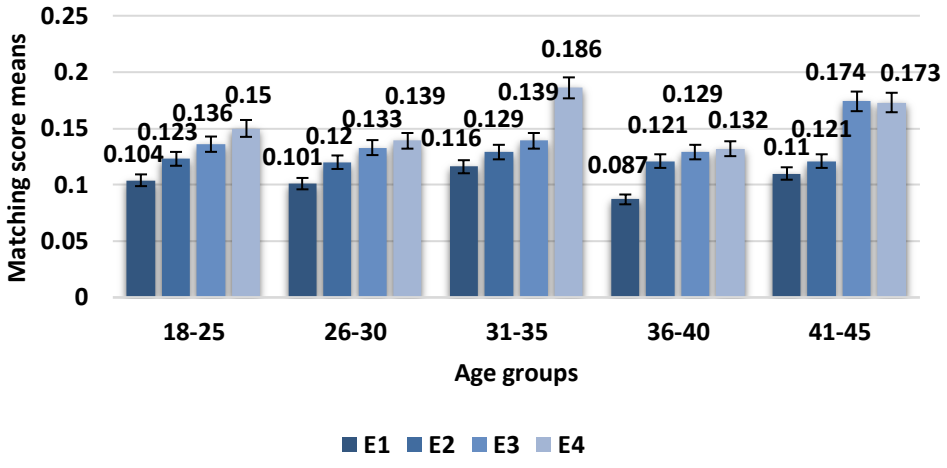


Figure 7.6: Matching score means obtained with VeriLook 10.0 across different age groups.

Participants aged 31-35 and 41-45 presented images that recorded the highest matching scores across the groups, with only two exceptions: the mean scores for the 41-45 group were lower in E2 and the 31-35 group were lower in E3. The group that reported

the lowest matching score values were aged between 36-40. These particular trends can also be seen for the mean matching scores assessed when Face_recognition was used, as presented in Figure 7.7.

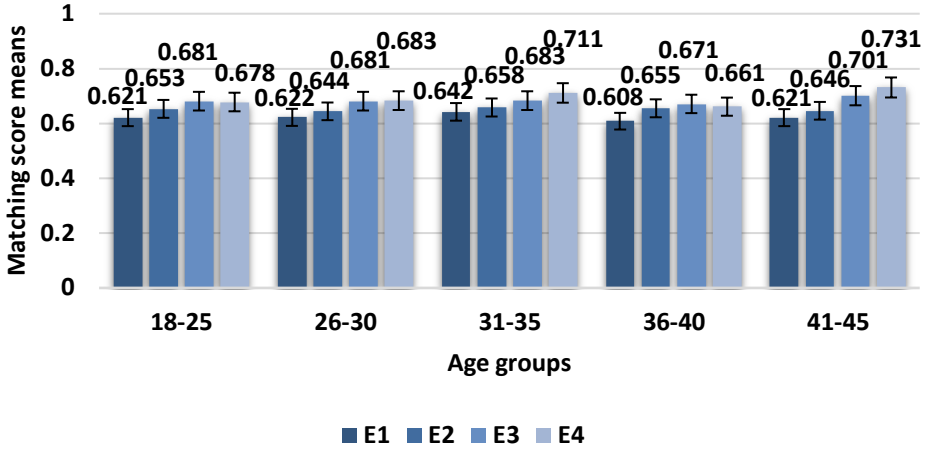


Figure 7.7: Matching score means obtained with Face_recognition across different age groups.

Participant belonging to the 31-35 and 41-45 age groups reported images with the highest scores, while the age group that includes participants aged 36-40, reported the lowest matching scores, with only one exception for the E2. It is interesting to note that the matching scores were higher for older participants but that there is a gap between the group aged 36-40. A deeper analysis over a larger population should enable an understand of what can influence such a difference. The differences amongst the matching scores between the groups were also assessed using a One-way analysis of variance (ANOVA) test that confirmed that there were significant results between each of the groups. The significant differences were observed for the values indicated in Table 7.6. Post-hoc multiple comparisons were assessed using the Tukey Honest Significant Distance (HSD) test indicated and confirmed the relationship that were evident from the charts. The age group 36-40 has values that are significantly lower compared to those calculated for the other age groups.

Table 7.6: One-way ANOVA statistical results across age groups.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	**E1**	$F(4,796) = 51.01$ at $p < 0.001$
	E2	$F(4, 810) = 4.64$ at $p < 0.001$
	E3	$F(4,771) = 17.63$ at $p < 0.001$
	E4	$F(4,791) = 38.46$ at $p < 0.001$
Face_recognition	**E1**	$F(4,888) = 38.12$ at $p < 0.001$
	E2	$F(4, 864) = 10.26$ at $p < 0.001$
	E3	$F(4,844) = 7.64$ at $p < 0.001$
	E4	$F(4,869) = 93.41$ at $p < 0.001$

Although the 36-40 age group presented the lowest matching scores amongst the groups, interesting results can be observed when considering the FRRs calculated for Face_recognition as the percentage of "Failed" verification is higher only for the E4, as can be seen in Table 7.7.

Table 7.7: FRR comparisons across age groups.

Enrolment scenarios	VeriLook 10.0					Face_recognition				
	18-25	26-30	31-35	36-40	41-45	18-25	26-30	31-35	36-40	41-45
E1	7.7%	10.2%	7.8%	22%	2.9%	0.2%	0.1%	0%	0%	0%
E2	3.5%	2.4%	3.8%	4.3%	0%	0.1%	0.1%	0%	0%	0%
E3	1.3%	1.5%	3.7%	4.7%	0%	0.1%	0.1%	0%	0%	0%
E4	2.6%	1.1%	0.7%	14.6%	0%	0. 1%	0.01%	0%	0.4%	0%

These trends should be aknowledged and assessed to understand if there are any external factors that could cause these differences to ensure better verification performance.

Differences were also assessed between different ethnic groups. The mean values for each enrolment are presented in Figure 7.8 for scores calculated with VeriLook 10.0.

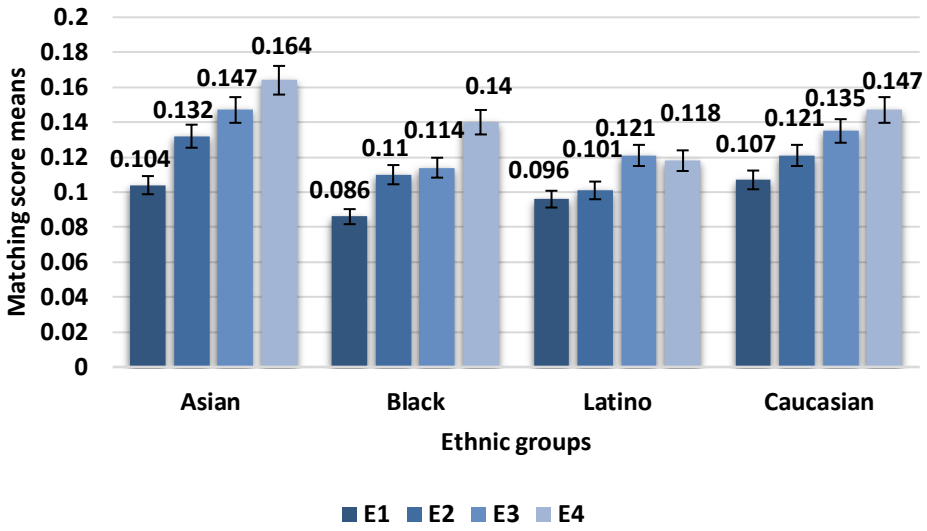


Figure 7.8: Matching score means obtained with VeriLook 10.0 across ethnic groups.

The trends described by the means are similar for all the enrolment scenarios: Asian and Caucasian are the two ethnic groups that recorded the higher values, while Latino and Black ethnic groups reported lower matching scores, although for E4 the means for Black and Caucasian presented similar values. Similar differences were found when checking the scores for Face_recognition as shown in Figure 7.9.

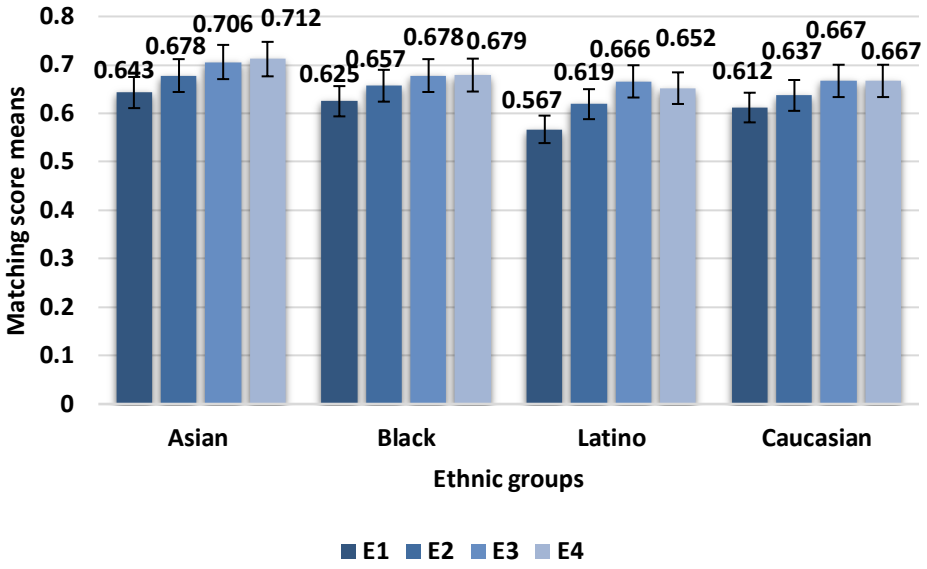


Figure 7.9: Matching score means obtained with Face_recognition across ethnic groups.

The matching scores calculated with Face_recognition reported the highest values for the Asian group in each enrolment scenarios and the lowest for the Latino group. The ethnic groups for Black and Caucasian subjects reported similar matching scores in all the enrolment scenarios, with the exception of E3, that considers indoors enrolment images, where the Caucasian group presented the lowest values that were close to those reported for the Latino group.

These significant differences were also confirmed when assessed through a One-way ANOVA test results of which are reported in Table 7.8.

Table 7.8: One-way ANOVA statistical results across ethnic groups.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	E1	$F(3,719) = 133.53$ at $p < 0.001$
	E2	$F(3, 738) = 88.15$ at $p < 0.001$
	E3	$F(3,665) = 72.68$ at $p < 0.001$
	E4	$F(3,824) = 81.74$ at $p < 0.001$
Face_recognition	E1	$F(3,720) = 239.15$ at $p < 0.001$
	E2	$F(3, 699) = 233.95$ at $p < 0.001$
	E3	$F(3,681) = 223.26$ at $p < 0.001$
	E4	$F(3,709) = 241.45$ at $p < 0.001$

There are differences observed also when assessing the FRRs over the two recognition systems (Table 7.9). These differences reported across ethnic groups need to be considered when implementing facial verification on a mobile device since they can definitely impact the system performance.

Table 7.9: FRR comparisons across ethnic groups.

Enrolment scenarios	VeriLook 10.0				Face_recognition			
	Asian	Black	Latino	Caucasian	Asian	Black	Latino	Caucasian
E1	10.7%	13.4%	4.5%	7.5%	0.1%	0%	0%	0.2%
E2	3 %	4.9%	3.8%	3.1%	0.1%	0%	0%	0.1%
E3	2.2%	3.2%	0%	1.3%	0.1%	0%	0%	0.1%
E4	4.7%	3.8%	0.6%	1.8%	0. 2%	0.%	0%	0.2%

Differences were not identified when assessing the participants educational background. These results do not necessary explain that the user's background does not contribute to the system performance, but rather that it was not statistically possible to estimate within our experimental population how this information can affect the verification performance.

Significant differences were not found within groups for the Operating System used by the participants on their personal devices. These results ensure that the verification performance analysis and the observations made were not affected by the Operating System used for the data collection. However, we cannot be certain that significantly differences would not be observed when using a particular Operating System for facial verification in scenarios that are different from those considered in this work.

Similarly, participants were also divided into groups depending on their previous experience with biometrics on mobile devices. A statistical analysis did not report any significant results between the two groups. However, this could be a consequent of the experimental population that participated in the collected database, as this was obtained in a University environment with a predominantly scientific educational background. For these reasons, general observations cannot be formed from these results over such a specific scenario.

It can be concluded that demographics information should be considered when assessing the system performance, as there are significant differences that can be observed across the groups. In particular, the information that reported significant values are generally connected to the physical appearance of users. There were not significant results observed across groups that defined the user's background or the familiarity with a particular Operating System or the security technology previously experienced in the context of mobile devices.

7.4.2 Static characteristics

Biometric performance was assessed according to the statistic characteristics presented in the images that were selected and described in Section 4.4.2.1. Images were divided depending on the presence of glasses on the participant's face. The matching scores calculated for each verification algorithm for this case are presented in Figure 7.10.

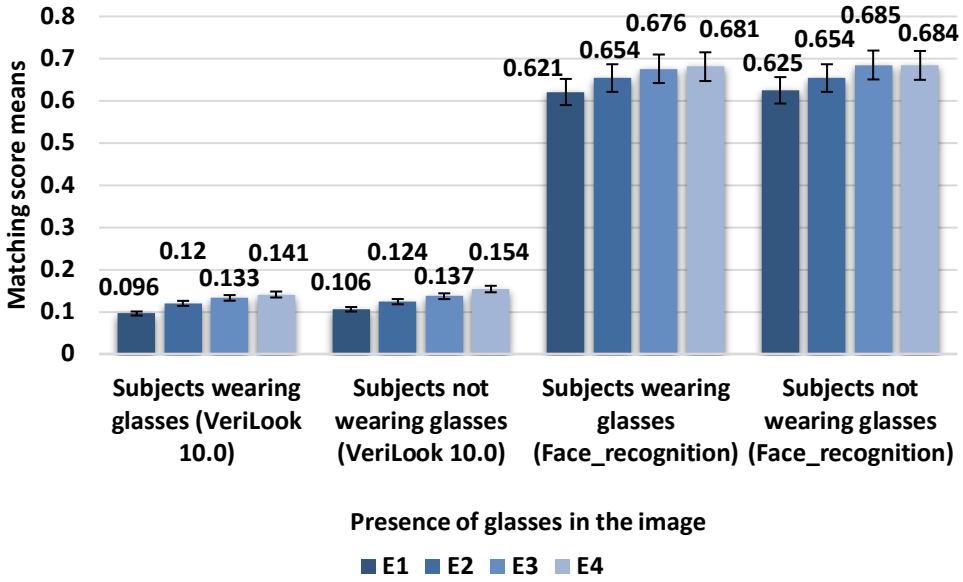


Figure 7.10: Matching scores means across images that presented subjects wearing and not wearing glasses during facial image capture.

Overall, both verification algorithms presented higher matching scores for images where the participants were not wearing glasses. A statistical independent t-test was performed to understand the relationship between the system performance and the presence of the static characteristic on the image, as shown in Table 7.12. Significant differences were observed for each enrolment scenario when the matching scores were calculated using VeriLook 10.0, but only E1 and E3 reported significant results when Face_recognition was used. The differences between the two systems could be explained considering the methods used by the two algorithms; depending on the facial features detected and utilised for the comparisons, the matching score might be more or less affected by the presence of the static characteristic on the image.

Table 7.10: Statistical independent t-test comparing the matching score means of images presenting subjects with or without glasses.

Verification systems	Enrolment scenarios	t-test	p	Mean difference	Magnitude of the difference
VeriLook 10.0	E1	t(3943) = -10.73	p < 0.001	-0.010	0.015
	E2	t(4071) = -3.07	p = 0.002	-0.003	0.001
	E3	t(7534) = -3.20	p = 0.001	-0.004	0.001
	E4	t(3553) = -6.00	p < 0.001	-0.013	0.005
Face_recognition	E1	t(2710) = -2.16	p = 0.031	-0.004	0.001
	E2	t(3282) = 0.38	P = 0.702	0.0001	-
	E3	t(2973) = -5.31	p < 0.001	-0.009	0.004
	E4	t(3.29) = 0.06	P =0.059	-0.0037	-

The Table also shows the magnitude of the difference between the groups, calculated in Eta Squared. It can be seen these differences were not particularly high in every scenario.

The presence in the image of heavy-make or a beard reported similar results to those observed for glasses: there were statistically significant differences between the groups as the matching scores were higher for those images were the static characteristic was not detected (Table 7.11). However, the magnitude of these differences was not significantly high in any of the enrolment scenarios. Differences were not observed only for the E1 and E2 when using Face_recognition. It could be hypothesised that closer matching scores could be obtained whether the enrolment scenarios also present the static characteristics, and this could explain the variations observed in the magnitude of the differences across each scenario.

From the results it could be concluded that the images presenting a user's static characteristic affected the performance of the system. Nevertheless, to better assess their effect, the influence that these images have over quality should also be included in the analysis to have a deeper understanding over the relationship amongst the variables that affected the biometric system.

Table 7.11: Statistical independent t-test comparing the matching score means of images presenting subjects that with heavy-make or a beard.

User static characteristic	Enrolment scenarios	t-test	p	Mean difference	Magnitude of the difference
VeriLook 10.0 Heavy make-up	**E1**	t(404) = -10.31	p < 0.001	-0.014	0.017
	E2	t(340) = -11.96	p < 0.001	-0.035	0.023
	E3	t(6131) = -5.99	p < 0.001	-0.017	0.006
	E4	t(332) = -8.57	p < 0.001	-0.058	0.012
Face_recognition Heavy make-up	**E1**	t(453) = -2.52	p = 0.012	-0.010	0.001
	E2	t(389) = -20.09	p < 0.001	-0.054	0.058
	E3	t(387) = -12.89	p < 0.001	-0.034	0.025
	E4	t(6483) = -5.87	p < 0.001	-0.026	0.005
VeriLook 10.0 Beard	**E1**	t(2241) = -20.49	p < 0.001	-0.019	0.059
	E2	t(2128) = -15.36	p < 0.001	-0.020	0.034
	E3	t(1937) = -13.58	p < 0.001	-0.022	0.027
	E4	t(2157) = -10.41	p < 0.001	-0.024	0.016
Face_recognition Beard	**E1**	t(1996) = -0.50	p = 0.616	-0.001	-
	E2	t(2095) = -0.23	p = 0.816	-0.0004	-
	E3	t(2091) = -7.4	p < 0.001	-0.013	0.007
	E4	t(2081) = -2.84	p = 0.005	-0.006	0.001

7.4.3 Dynamic characteristics

An analysis was carried out to check whether the matching scores provided by the two verification algorithms could be affected by dynamic characteristics of the users; if they were blinking or had their mouth open during the facial image presentation. The verification dataset presented 8.9% of images where the participants were blinking during the facial image acquisition. The matching scores calculated with VeriLook 10.0 were compared between the two groups of images as shown in Figure 7.11. There are noticeable differences in images that presented blink and no blink. Where a dynamic characteristic is not present in the image, the verification matching scores are higher. The trend between the enrolment types is the same in each enrolment scenario: higher values were recorded for the enrolments E3 and E4.

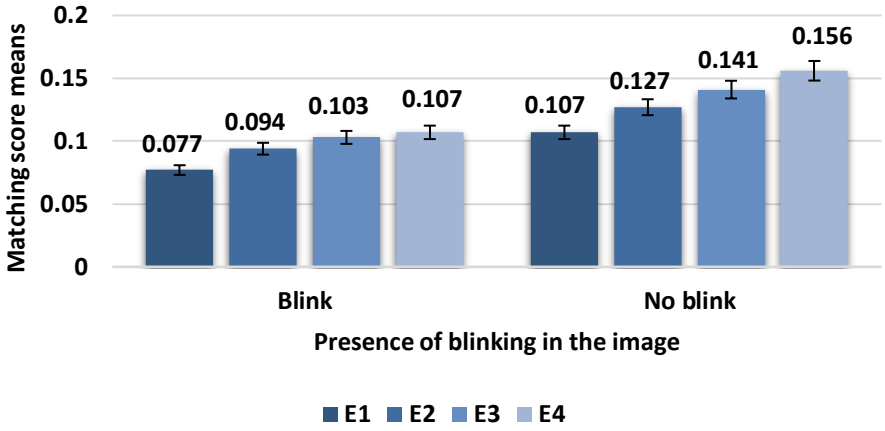


Figure 7.11: Matching score means calculated with VeriLook 10.0 comparing images of subjects that did and did not present blink.

Similar observations can be seen for the matching scores presented by Face_recognition, meaning that these characteristics influence both the verification systems considered (Figure 7.12).

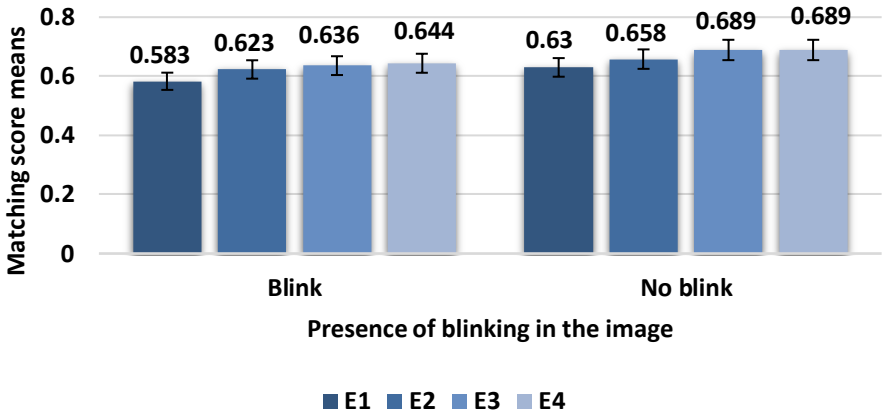


Figure 7.12: Matching score means calculated with Face_recognition comparing images of subjects that did and did not present blink.

A statistical t-test was performed to check the influence of these differences. The results can be seen in Table 7.12, as well as the magnitude of the differences, calculated in Eta Squared. The Face_recognition algorithm seems to be impacted less but, in general, performance is dependent on the type of facial features that are considered in the two algorithms' methods. However, overall the magnitude of differences is not high, but for an extended population, even small differences can have an impact on the performance. Enrolment outdoors seems to be the enrolment type more resilient for both algorithms to this type of variation.

Table 7.12: Independent t-test statistical results comparing matching score means between images that did or did not present a blink during the verification presentation.

Verification systems	Enrolment scenarios	t-test	p	Mean difference	Magnitude of the difference
VeriLook 10.0	**E1**	t(999) = -28.63	p < 0.001	-0.030	0.101
	E2	t(1153) = -26.04	p < 0.001	-0.032	0.085
	E3	t(1091) = -25.21	p < 0.001	-0.038	0.080
	E4	t(1121) = -21.15	p < 0.001	-0.049	0.058
Face_recognition	**E1**	t(7271) = -18.96	p < 0.001	-0.047	0.047
	E2	t(894) = -15.07	p < 0.001	-0.036	0.030
	E3	t(7271) = -22.65	p < 0.001	-0.053	0.066
	E4	t(818) = -13.62	p < 0.001	-0.045	0.025

A similar study was also performed to check the influence that an open mouth image can present to matching performance. There is a higher number of images that presented this characteristic in the verification dataset: 1,447 (18.3%). The differences between the matching scores calculated for the two groups presented similar results as for blink: where the characteristic is present, the images have lower matching scores. However, the effect size of these differences is even smaller than those considered for a blink, as it can be seen in Table 7.13.

Table 7.13: Independent t-test statistical results comparing matching score means of images of participants that did or did not present mouth open during the verification presentation.

Verification systems	Enrolment scenarios	t-test	p	Mean difference	Magnitude of the difference
VeriLook 10.0	**E1**	t(2092) = -5.72	p < 0.001	-0.006	0.004
	E2	t(2481) = -12.09	p < 0.001	-0.015	0.020
	E3	t(2651) = -11.93	p < 0.001	-0.016	0.019
	E4	t(2477) = -5.39	p < 0.001	-0.012	0.004
Face_recognition	**E1**	t(2328) = -5.61	p < 0.001	-0.010	0.004
	E2	t(7211) = -10.75	p < 0.001	-0.020	0.016
	E3	t(7211) = -6.85	p < 0.001	-0.012	0.006
	E4	t(7211) = -5.28	p < 0.001	-0.011	0.004

Even if all the comparisons presented statistical significance, the magnitude of the difference is approximately less than 0.020 overall.

It is possible to detect when users present a dynamic characteristic as blinking or having their mouth open, but it may not be possible to predict when and if the characteristic might occur during an image presentation. Nonetheless, the magnitude of difference observed between images that did and did not present static characteristic is small, indicating that there is a small effect size in the influence that these characteristics have with the biometric performance.

7.4.3.1 Facial expressions

This study aimed to assess the facial expressions presented by the users when images were taken with the smartphone under different conditions. The aim was to understand how these variations affect the matching scores obtained during the verification of the image.

Images were divided into groups according to 7 different types of facial expressions that were detected using Neurotechnology VeriLook 10.0: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. A One-way ANOVA test was performed to check whether there were differences within the groups. Both the verification algorithms reported matching scores values that were significantly different for each of the enrolment scenarios. The values for the statistical significance are shown in Table 7.14.

Table 7.14: One-way ANOVA test across facial expressions groups.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	**E1**	F(6,1050) = 280.60 at p < 0.001
	E2	F(6, 1061) = 256.33 at p < 0.001
	E3	F(6,1046) = 241.94 at p < 0.001
	E4	F(6,1053) = 121.31 at p < 0.001
Face_recognition	**E1**	F(6,1019) = 118.63 at p < 0.001
	E2	F(6, 1022) = 141.76 at p < 0.001
	E3	F(6,1018) = 143.29 at p < 0.001
	E4	F(6,1026) = 75.55 at p < 0.001

The matching scores recorded across the facial expression groups presented a similar trend for both verification algorithms. When E1 was considered, the facial expressions that recorded the highest matching scores were Happiness and Neutral. Anger was the facial expression that reported the lowest values, especially for the matching scores calculated with VeriLook 10.0. Face_recognition identified Sadness as the facial expression that obtained less similarity. The trend of normalised matching scores can be seen in Figure 7.13.

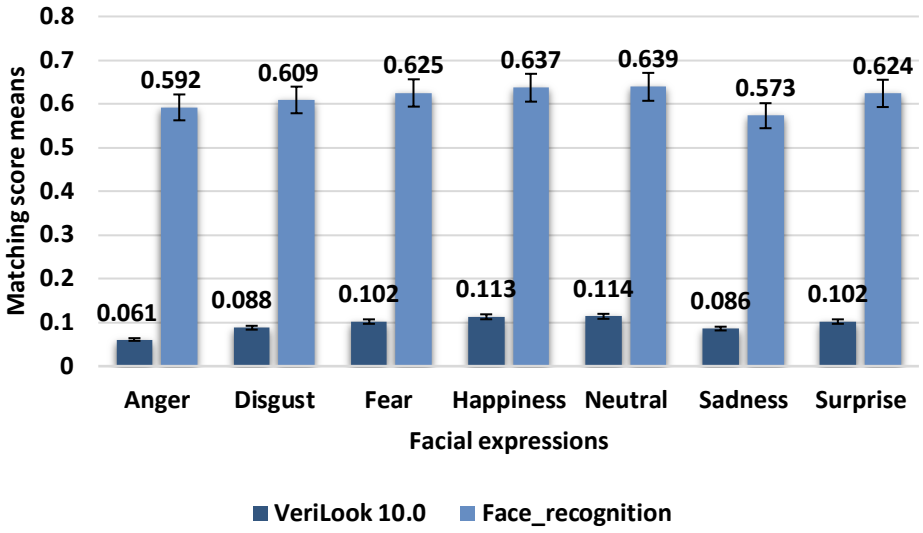


Figure 7.13: Matching score means for the E1 scenario comparing facial expressions detected when the users were collecting facial images.

When the second enrolment scenario E2 was assessed, the trend between the two verification algorithms differed as shown in Figure 7.14. Anger and Sadness are still the two perceived facial expressions that presented lower matching scores, but for VeriLook 10.0 Sadness did not report values lower as when Face_recognition was used. Neutral and Surprise reported the highest matching scores.

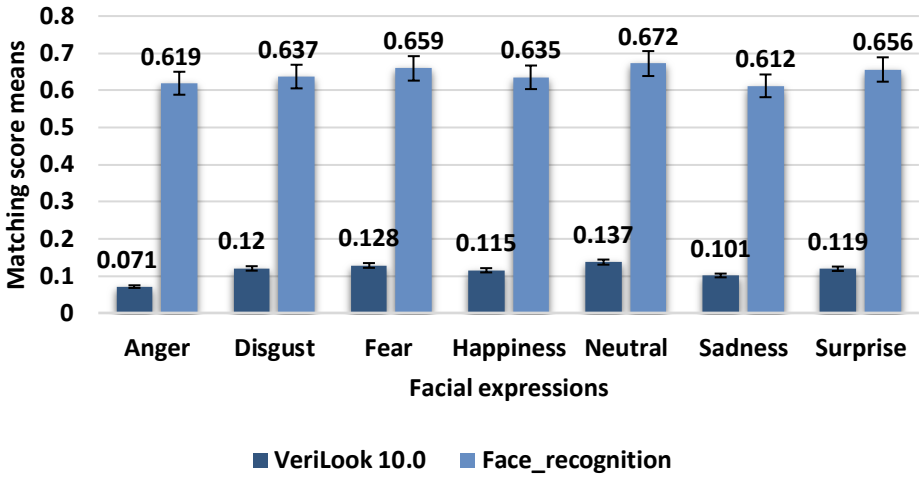


Figure 7.14: Matching score means for the E2 scenario comparing facial expressions detected when the users were collecting facial images.

The third enrolment scenario, E3, considered significantly lower values for Anger, but while Sadness was the facial expression that most affected the matching scores in Face_recognition, VeriLook 10.0 reported lower values also for Disgust. Neutral is the facial expression that resulted in the highest matching scores, but, generally, more positive facial expressions like Happiness and Surprise reported higher matching score

values in both algorithms. The different trends for the matching scores presented in E3 can be seen in Figure 7.15.

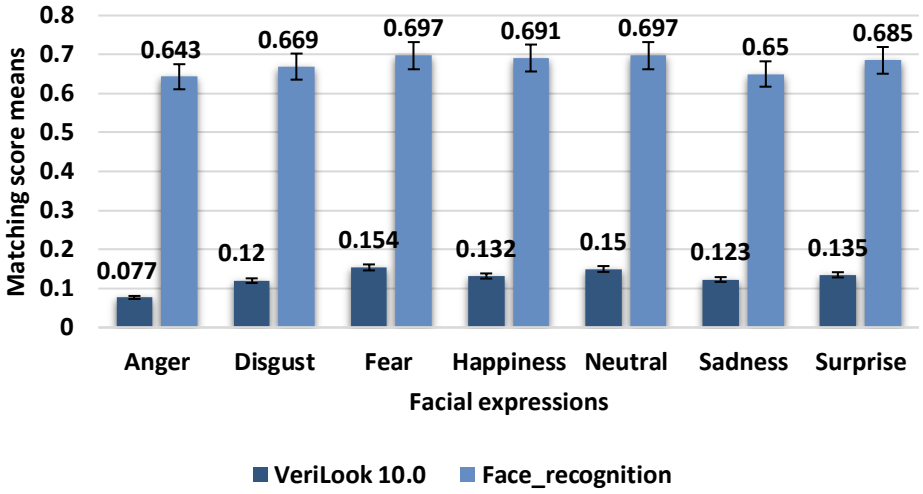


Figure 7.15: Matching score means for the E3 scenario comparing facial expressions detected when the users were collecting facial images.

Finally, when considering the last enrolment scenario, E4, the two verification algorithms reported different trends. Neutral and Happiness are the expressions that reported higher results for Face_recognition, with Sadness reporting the lowest matching scores for the system. When considering the VeriLook 10.0 algorithm, the matching scores that reported the lowest values are those images that presented Anger as a facial expression. These trends can be observed in Figure 7.16.

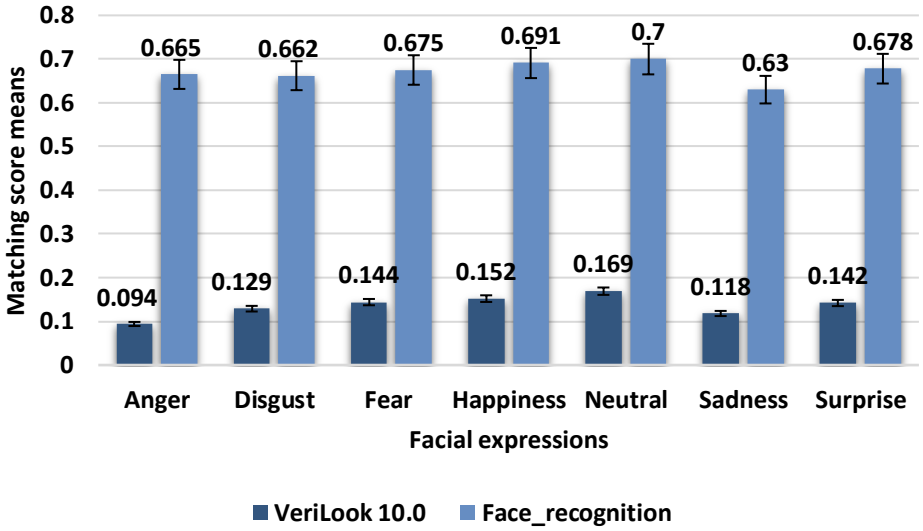


Figure 7.16: Matching score means for the E4 scenario comparing facial expressions detected when the users were collecting facial images.

In summary, despite small dissimilarity between the verification algorithms, it was observed that:

- Anger and Sadness are the two facial expressions that reported the lowest matching scores in all the scenarios.
- When the enrolment was considered for smartphone images taken indoors (E3), lower matching scores were observed when the detected expression was Disgust.
- Generally all the facial expressions that reported "negative" emotions, including Fear, reported lower matching scores than for the images where a more "positive" emotion was detected.
- Neutral is the facial expressions that overall reported the highest matching score in all the scenarios.
- Happiness and Surprise reported high matching scores: images where Happiness was detected presented higher scores for E1 and E3, while Surprise resulted in higher scores especially for E2 and E4.

Often, facial expressions are recorded as an involuntary reaction that the users have with the environmental conditions. Figure 7.17 presents the percentage of occurrence for each facial expression recorded across the different environmental locations, although there were no significant differences recorded between the facial expressions detected over indoors and outdoors images.

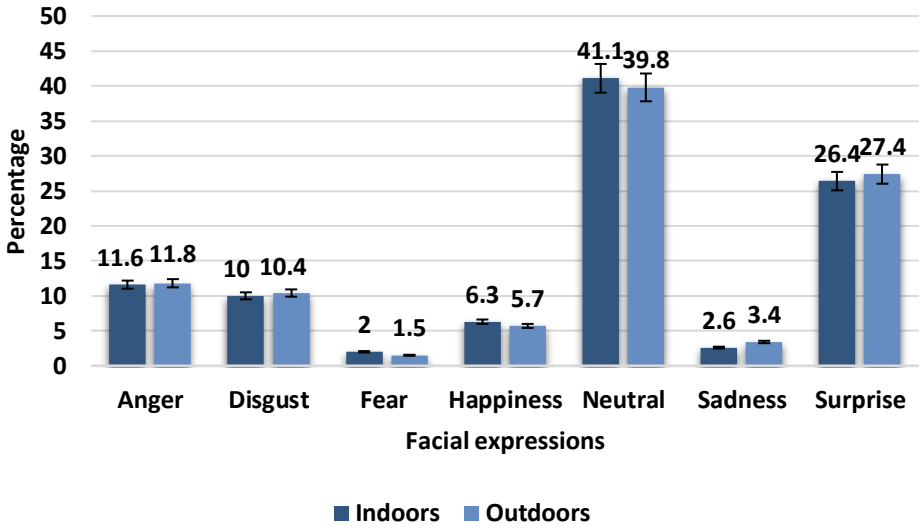


Figure 7.17: Percentages of occurrence of facial expressions in the images between indoors and outdoors locations.

Contrary to the detection of "negative" expressions, such as Sadness and Anger, that reported lower matching performance, Neutral and facial expressions presenting "positive" emotions, like Happiness and Surprise, recorded higher matching scores. These trends were also confirmed by the FRR recorded across the variables as seen in Table 7.15.

Table 7.15: FRR across the different facial expressions detected.

Facial expressions	VeriLook 10.0				Face_recognition			
	E1	E2	E3	E4	E1	E2	E3	E4
Anger	17.7%	6.6%	5.6%	2%	0.2%	0.2%	0.2%	0.2%
Disgust	17.7 %	5.6%	3.8%	9.1%	0.1%	0.3%	0.3%	0.1%
Fear	10.2%	3.1%	2.3%	3.9%	0%	0%	0%	0%
Happiness	4%	2%	1.1%	1.8%	0%	0%	0%	0%
Neutral	3.8%	0.9%	0.6%	1.4%	0.01%	0.01%	0.01%	0.1%
Sadness	10.6%	7.2%	1.7%	8.9%	1.7%	0.9%	0%	0.4%
Surprise	10.6%	4.4%	1.3%	2.4%	0. 01%	0%	0%	0%

These results are promising as they present an analysis over facial expressions that could be potentially used to understand the user interaction and the environmental variations during the biometric presentation, or to provide real-time feedback that could be useful to the users to understand what facial expressions they should avoid to enhance the biometric performance.

7.4.3.2 User's pose

As seen previously in Section 5.4.3.2, users' head pose can be assessed considering the rotation angles of yaw, pitch and roll. The angular rotations of the head can affect the performance of the system. Requirements specified by the ISO/IEC 19794-5 Standard, yaw and pitch angles should not exceed ±5 degrees, while roll angles should not be over ±8 degrees. In the case of a mobile scenario, these restrictions are difficult to implement, and the majority of the images in the verification dataset (88.4%) were not compliant with the Standard, as can be seen in Table 7.16. Yaw angles in particular was not compliant for the 85% of the images, pitch for the 11.4% and roll for the 2.6%.

Table 7.16: Percentages of images that were compliant with the Standard ISO/IEC 19794-5 image acquisition requirements for user pose.

ISO\IEC 19794-5 user's pose compliance:	Percentage
Compliant	11.6%
One pose not compliant	78.3%
Two poses not compliant	9.7%
None of the poses is compliant	0.4%
Total	100%

A One-way ANOVA test was performed to understand the differences between biometric performance and images that presented non-compliant head rotations. The results are shown in Table 7.17. There were significant differences for each of the enrolment scenarios.

Table 7.17: One-way ANOVA test across groups of images according to their compliance in user's pose with the ISO/IEC 29794-5 Standard.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	E1	F(3,160) = 7.87 at p < 0.001
	E2	F(3, 160) = 8.46 at p < 0.001
	E3	F(3,160) = 12.45 at p < 0.001
	E4	F(3,160) = 18.18 at p < 0.001
Face_recognition	E1	F(3,7910) = 8.53 at p < 0.001
	E2	F(3, 7910) = 6.29 at p < 0.001
	E3	F(3, 7910) = 13.45 at p < 0.001
	E4	F(3, 7910) = 19.15 at p < 0.001

Post-hoc multiple comparisons using Tukey tests helped to understand how the differences between each group were observed. The trend is similar for each enrolment scenario: the images that were compliant with the Standard presented higher matching scores and the values decrease when one or two angular positions are not compliant, decreasing to the lowest values when none of the angular rotations were compliant. Matching scores values for the VeriLook 10.0 algorithm are reported in Figure 7.18.

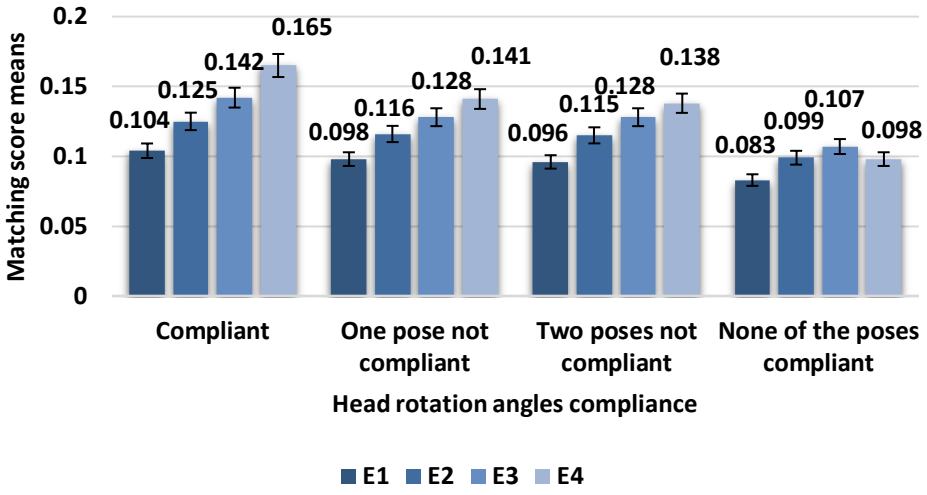


Figure 7.18: Matching score means by VeriLook 10.0 assessing the compliance of user's poses in the image.

Interestingly, the Face_recognition system reported the highest values for the E1 when none of the angular poses were compliant, as shown in Figure 7.19. it could be possible that since the E1 enrolment scenario considers SLR images, the angular rotations that were not compliant presented a degree of difference from the requirements that was not too high and this could explain the good performance compared to the other enrolment scenarios. The other enrolment scenarios reported the usual trend: the images where all the poses were compliant receive higher scores, while those with none are the lowest.

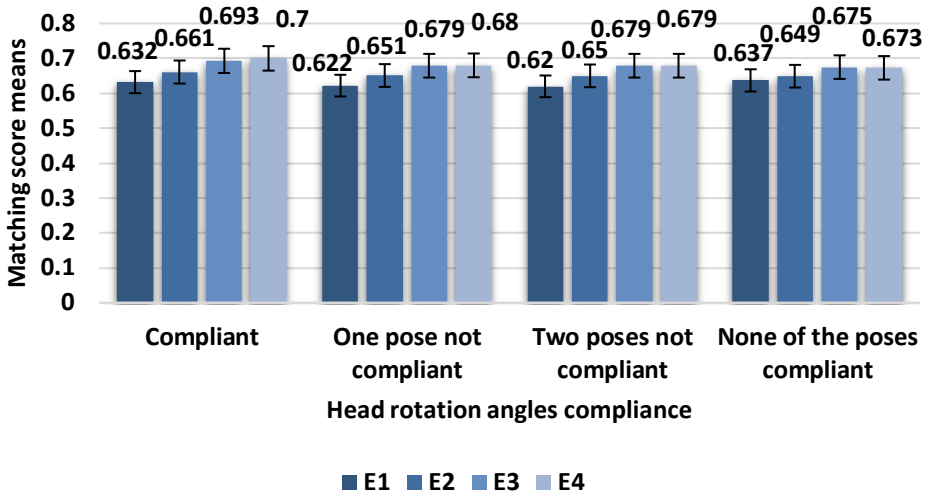


Figure 7.19: Matching score means by Face_recognition assessing the compliance of user's poses in the image.

There are important observations that can be determined from this analysis. First of all, the importance of having constrained angular poses, as it is verified that they affect the verification performance, but they should be adjusted for a mobile scenario and being less strict. In particular yaw angles since the majority of the cases (85%) is not compliant.

It could be possible, for instance, modify the requirements for yaw angles from being within a range of ±5 to a range of ±10. When using the ISO/IEC 19794-5 Standard requirement, the percentage of "Successful" verification that could be discarded since the pose angle do not conform with the requirements is extremely high considering the application in mobile authentication. When changing the degrees of requirements to a more permissive head pose rotation angles, the percentage of discarded "Successful" verification is more acceptable within the context of facial verification for smartphone devices.

Table 7.18: "Successful" verification percentages that presented yaw angles not compliant within two different requirement ranges.

Verification systems	Enrolment scenarios	Yaw angles ±5 degrees	Yaw angles ±10 degrees
VeriLook 10.0	E1	84.8%	1.12%
	E2	84.7%	1.3%
	E3	83.8%	1.28%
	E4	84.8%	1.26%
Face_recognition	E1	84.9%	0.18%
	E2	85%	0.13%
	E3	84.9%	0.1%
	E4	85%	1.67%

7.5 Quality assessment in relation to the performance

This last Section of this Chapter presents the analysis of biometric performance considering the relationship with each of the quality metrics considered:

- Brightness
- Contrast
- GCF
- Blurriness
- Exposure

Each of them was normalised to be in the same range as explained in Section 4.4.3.1. For the purpose of this analysis, each metric was divided in 5 different levels to classify the values of FIQ metrics that an image should present to have better biometric performance and each range level is show in Table 7.19.

Table 7.19: Frequency and percentage of images divided into groups according to the FIQ metric level.

FIQ metrics	Level	Range	Frequency	Percentage
Brightness	**1**	0-1	72	0.9%
	2	1-2	1325	16.7%
	3	2-3	4756	60.1%
	4	3-4	1697	21.4%
	5	4-5	64	0.8%
Contrast	**1**	0-1	5	0.1%
	2	1-2	60	0.8%
	3	2-3	851	10.8%
	4	3-4	5782	73.1%
	5	4-5	1216	15.4%
GCF	**1**	0-1	132	1.7%
	2	1-2	1058	13.4%
	3	2-3	3656	46.2%
	4	3-4	2700	34.1%
	5	4-5	368	4.6%
Blurriness	**1**	0-1	120	1.5%
	2	1-2	2498	31.6%
	3	2-3	4177	52.8%
	4	3-4	1046	13.2%
	5	4-5	73	0.9%
Exposure	**1**	0-1	0	0%
	2	1-2	6	0.1%
	3	2-3	35	0.4%
	4	3-4	695	8.8%
	5	4-5	7178	90.7%

For example, extremely bright images will be indicated with Brightness level 5, while an extreme level of darkness in the image will be indicated with level 1. In the collected database, only a few images presented extreme values in Brightness, as the majority

presented an intermediate level that can be indicated as level 3. Similar considerations can be seen for the level of Contrast presented in the image, where level 1 indicates lower values of Contrast in the image, and 5 higher values.

As previously seen in Section 6.2, Contrast presented a distribution slightly shifted to higher values compared to Brightness. A small percentage of images is included in the first two Contrast levels of this FIQ metric. Image GCF and Blurriness similarly also followed the same trend, with lower percentages of images included in the highest (level 5) and lowest (level 1) values of the metric. Exposure presented a distribution that was more skewed on the higher values, and did not present any image with a level 1 of Exposure, but the percentage of images included in each level increases with the values of Exposure presented.

The analysis assessed each of the FIQ metrics to understand their relationship with the matching scores and generally if it is possible to regulate the FIQ levels to obtain a higher performance of the system.

7.5.1 Brightness

The means of the matching scores calculated with VeriLook 10.0 are shown in Figure 7.20. The enrolment scenarios reported similar trends, apart from E4, the scenario that considered images taken outdoors. E4 in fact reported the highest matching scores, but the highest and lowest level of Brightness resulted in better results. The other three enrolment scenarios reported a similar trend in which the lowest level of Brightness (level 1) reported the worse performance, and the matching score values increase with the level of Brightness until the maximum values in level 4 and 5. E1 presented a small difference between the two highest values, but for E2 and E3 the matching scores decreased a little from level 4 to 5.

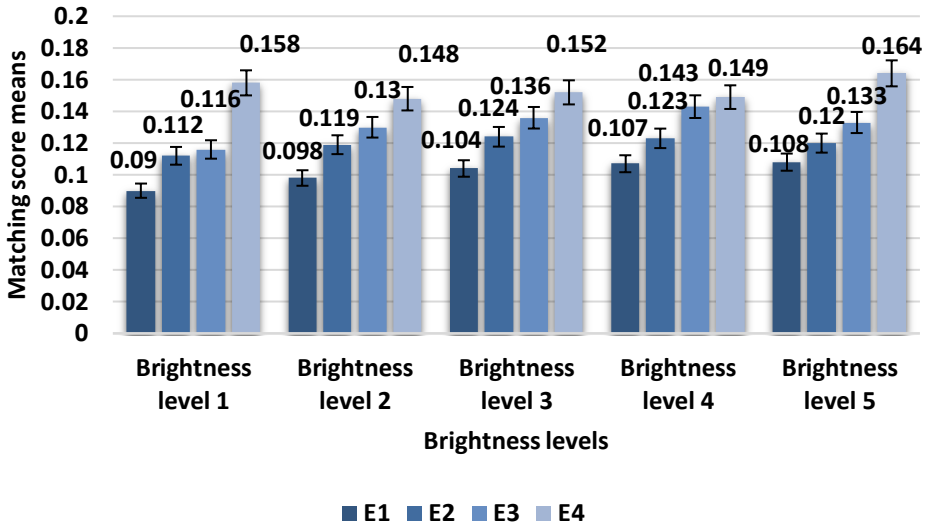


Figure 7.20: Matching score means for VeriLook 10.0 according to the level of Brightness.

When the scores are calculated with Face_recognition, the results are similar for E1 and E3, but the other two enrolment scenarios reported values different from VeriLook 10.0, as can be seen in Figure 7.21.

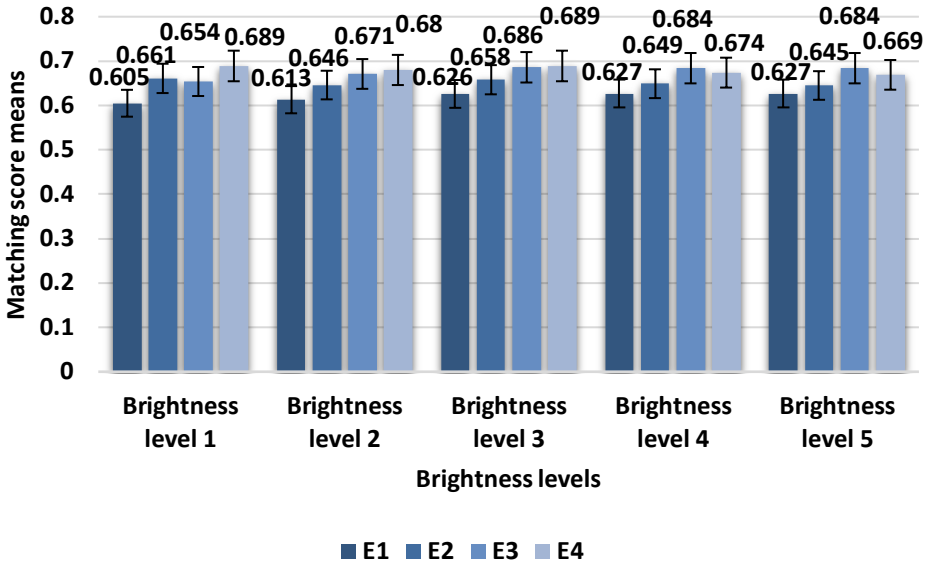


Figure 7.21: Matching score means for Face_recognition according to the level of Brightness.

The highest values in the cases of E2 and E4 are higher for Level 1 of Brightness and for Level 3, and then decrease for the last two levels, where the brightness presented the highest values. A One-way ANOVA test was performed to check these differences reporting the significant results shown in Table 7.20.

Table 7.20: One-way ANOVA results for group comparison across Brightness levels.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	E1	F(4,176) = 15.15 at p < 0.001
	E2	F(4, 178) = 3.8 at p = 0.005
	E3	F(4,179) = 13.70 at p < 0.001
	E4	F(4,175) = 1.06 at p =0.379
Face_recognition	E1	F(4,177) = 8.53 at p < 0.001
	E2	F(4, 177) = 6.29 at p < 0.001
	E3	F(4, 178) = 13.45 at p < 0.001
	E4	F(4, 176) = 19.15 at p < 0.001

The results were significant for each enrolment scenarios apart from the E4 in VeriLook 10.0. The Post-hoc comparisons with Tukey test was performed to understand the relationship between the groups that confirmed the trends that were seen on the charts.

Looking at the binary outcome for each verification system (Table 7.21) it can be seen that there are mainly two situations: when considering an enrolment with the SLR, the

comparisons with smartphone images resulted in being negatively influenced by the extreme levels of Brightness in the image: when the image is too dark (level 1) or too bright (level 5) the performance for both system were lower. In the other scenarios, the performance improves with a higher value of Brightness, with the exception for E4 in VeriLook 10.0 that presents a lower performance for the extreme cases but that did not present any significance in the distance between the means amongst the groups.

Table 7.21: FRR for Brightness levels.

Brightness level	VeriLook 10.0				Face_recognition			
	E1	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1	27%	8.1%	5.4%	5.4%	0 %	0.2%	0.2%	0.2%
2	11.3%	4.2%	2.6%	3.7%	0.1%	0.1%	0.1%	0.1%
3	8.8%	3.1%	1.7%	2.5%	0.2%	0%	0%	0%
4	7.6%	3%	1.2%	3.2%	0%	0%	0%	0%
5	10.2%	0%	0%	3.4%	0%	0%	0%	0%

Generally, we can say that to have a better biometric performance, smartphone images are more affected in Brightness. For the enrolment scenario that included SLR images the extreme lower and higher levels of Brightness should be avoided. While in the mobile scenarios, the higher the level of Brightness in the images, the better the performance. These observations were generally valid for both verification algorithms, although small differences in matching scores were observed between VeriLook 10.0 and Face_recognition.

7.5.2 Contrast

When considering Contrast, there were only a few images that presented a Level 1. A One-way ANOVA performed to check the differences between the Contrast groups reported significant results in any of the enrolment scenarios (Table 7.22).

Table 7.22: One-way ANOVA results for group comparison across Contrast levels.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	**E1**	$F(4,7909) = 151.25$ at $p < 0.001$
	E2	$F(4, 30) = 120.76$ at $p < 0.001$
	E3	$F(4,30) = 134.69$ at $p < 0.001$
	E4	$F(4,30) = 109.37$ at $p < 0.001$
Face_recognition	**E1**	$F(4,30) = 50.33$ at $p < 0.001$
	E2	$F(4, 30) = 34.86$ at $p < 0.001$
	E3	$F(4, 30) = 43.67$ at $p < 0.001$
	E4	$F(4, 30) = 18.34$ at $p < 0.001$

The mean values for matching scores calculated with VeriLook 10.0 can be seen in Figure 7.22.

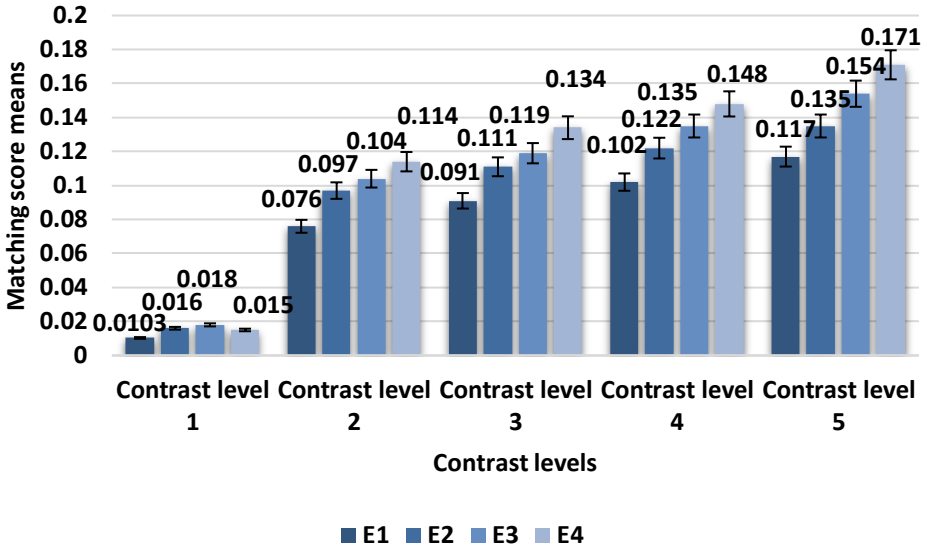


Figure 7.22: Matching score means for VeriLook 10.0 according to the level of Contrast.

Apart from huge difference observed in Contrast level 1, the main trend is that the higher the level of contrast in the image, the better the performance. In the case of Contrast, these trends were similar in all the enrolment scenarios. These observations were also considered for Face_recognition, as can be seen in Figure 7.23.

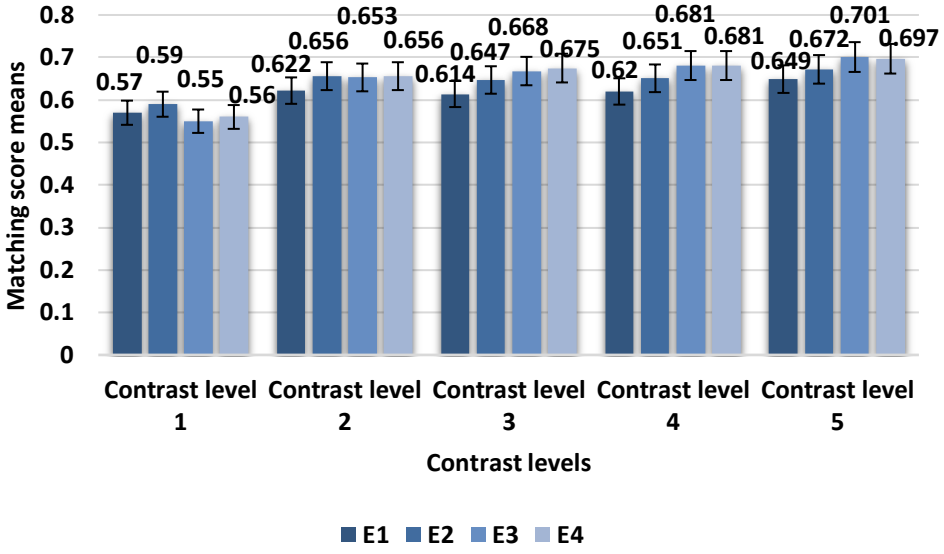


Figure 7.23: Matching score means for Face_recognition according to the level of Contrast.

The trend for FRR showed an enhanced performance for the higher level of Contrast (Table 7.23), although the frequency of images in Level 1 and Level 2 was low, as

previously observed in Table 7.19, which is probably why the scores for each of the enrolment scenarios is so high when compared to images with such low contrast.

Table 7.23: FRR for Contrast levels.

Contrast level	VeriLook 10.0				Face_recognition			
	E1	E2	E3	E4	E1	E2	E3	E4
1	100%	0%	0%	0%	0 %	0%	0%	0%
2	26.7%	0%	0%	0%	0%	0%	0%	0%
3	15.6%	6.4%	4.1%	5.3%	0.1%	0.1%	0.1%	0.1%
4	9.1%	3.3%	1.6%	2.7%	0.1%	0.1%	0.1%	0.1%
5	3.9%	1.2%	0.8%	1.9%	0%	0%	0%	0%

It is possible to conclude that the higher the values for Contrast, the better the performance of the system, as confirmed by both algorithms.

7.5.3 GCF

The trend presented for GCF lowers down the verification performance when the level is GCF is either Level 1 or Level 5. As shown in Figure 7.24, the matching scores calculated by VeriLook 10.0 presented the best results for GCF that were level 3 (medium values) and decrease when the GCF tends to lower or higher values.

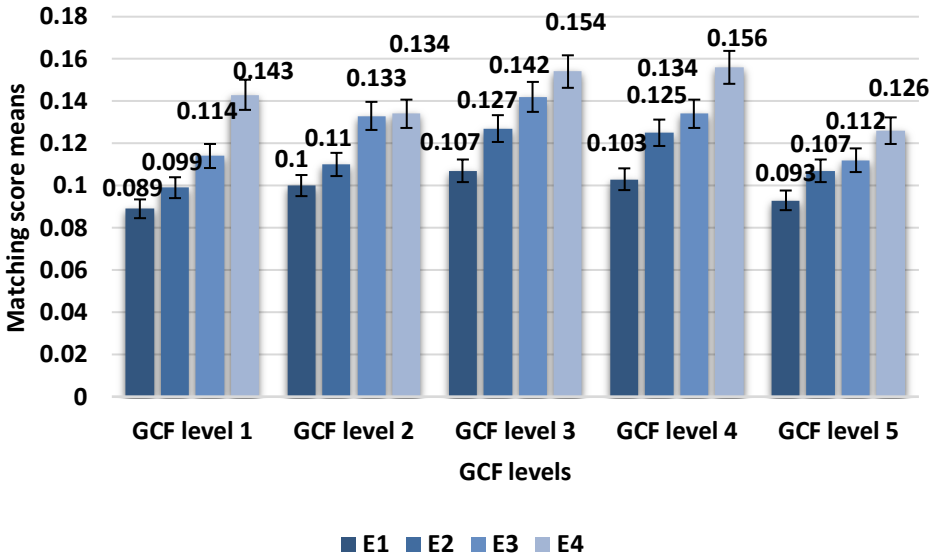


Figure 7.24: Matching score means for VeriLook 10.0 according to the level of GCF.

The values calculated from Face_recognition reported similar results and can be seen in Figure 7.25. In both algorithms the only enrolment scenario that behaves slightly different is E4. In this scenario the lower matching scores were also recorded when the level of GCF is 2.

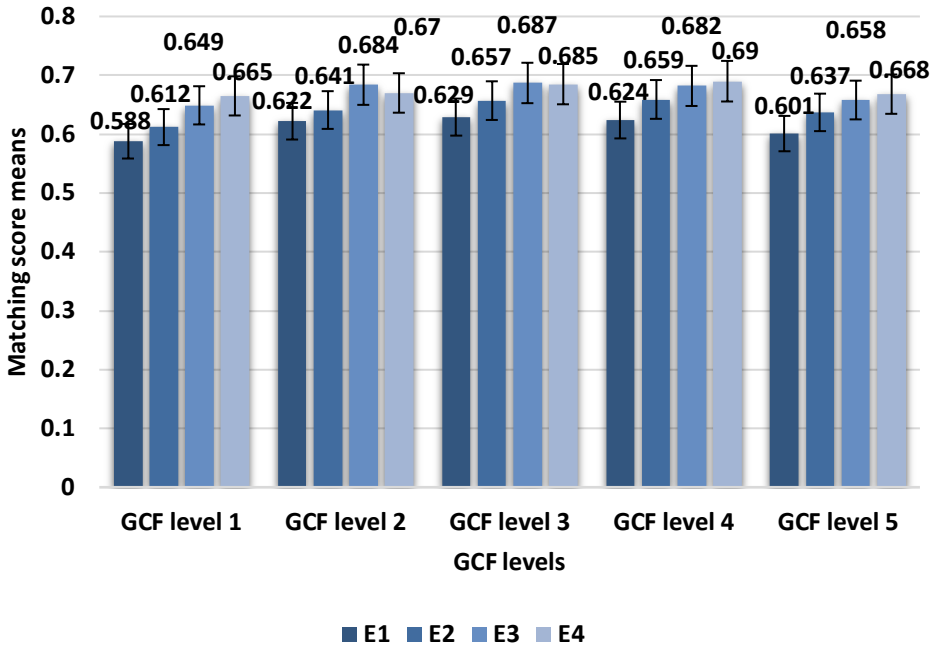


Figure 7.25: Matching score means for Face_recognition according to the level of GCF.

It can be seen that to have better performance with GCF, the values should be between 2 and 4, but for E4 when the enrolment was outdoors, level 2 recorded a lower verification performance.

The trends observed were significant as shown in the results from a One-way ANOVA test that was performed considering post-hoc multiple comparisons. The significant values were seen as shown in Table 7.24.

Table 7.24: One-way ANOVA results for group comparison across GCF levels.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	E1	F(4,659) = 24.72 at p < 0.001
	E2	F(4, 679) = 60.08 at p < 0.001
	E3	F(4,691) = 53.94 at p < 0.001
	E4	F(4,668) = 26.57 at p < 0.001
Face_recognition	E1	F(4,662) = 25.39 at p < 0.001
	E2	F(4, 660) = 37.50 at p < 0.001
	E3	F(4, 661) = 25.06 at p < 0.001
	E4	F(4, 660) = 19.08 at p < 0.001

The binary outcomes from both systems can be seen in Table 7.25. Observing the VeriLook 10.0 FRR trends, it can be seen that the best performance was presented for the three levels of GCF that range between 2 and 4.

Table 7.25: FRR for GCF levels.

GCF level	VeriLook 10.0				Face_recognition			
	E1	E2	E3	E4	E1	E2	E3	E4
1	27%	9.6%	1.7%	4.3%	0%	0%	0%	0%
2	10.1%	5.5%	1.5%	3.4%	0%	0%	0%	0.1%
3	6.8%	2.8%	1.7%	3.3%	0.1%	0.1%	0.1%	0.1%
4	10%	2.7%	1.8%	2.2%	0.2%	0.2%	0.1%	0.1%
5	15.1%	4.6%	2.9%	1.4%	0.6%	0%	0%	0%

7.5.4 Blurriness

The level of Blurriness in the image appeared to have similar trend as GCF: the extreme low and high values obtained the lowest matching scores, as can be seen in Figure 7.26.

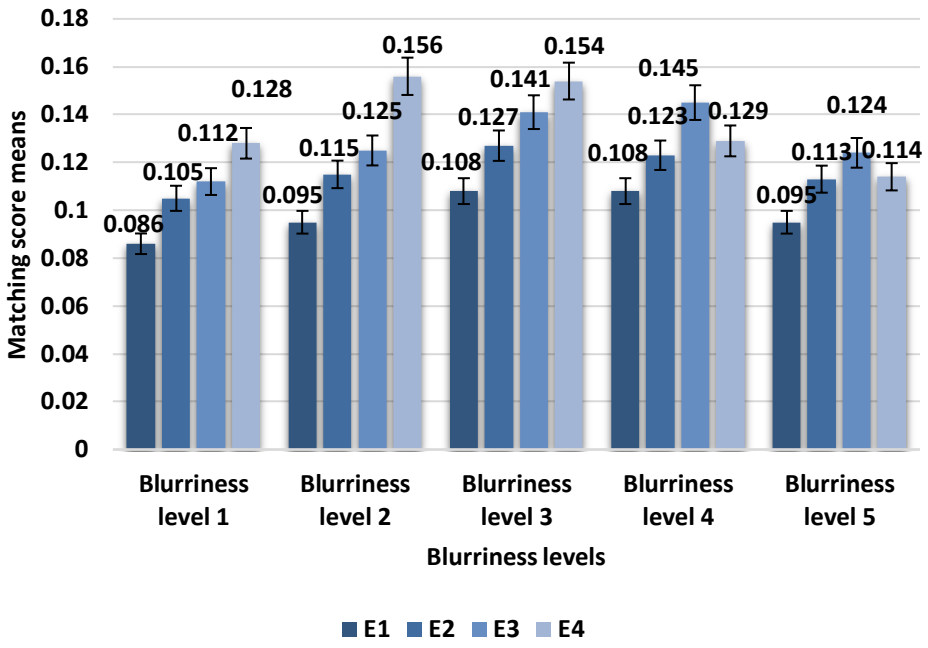


Figure 7.26: Matching score means for VeriLook 10.0 according to the level of Blurriness.

When the image is too sharp, as occurring in Level 1, or too blurred, as for Level 5, the matching scores recorded are lower. VeriLook 10.0 reported matching scores that were higher for levels 3 and 4 of Blurriness. Again, there is a different result for those images included in E4. The level of blur in this case should be excluded for categories 2 and 3. In particular, this enrolment scenario seems more affected by extreme level of Blurriness than the others.

Similar observations were valid also for Face_recognition, as shown in Figure 7.27.

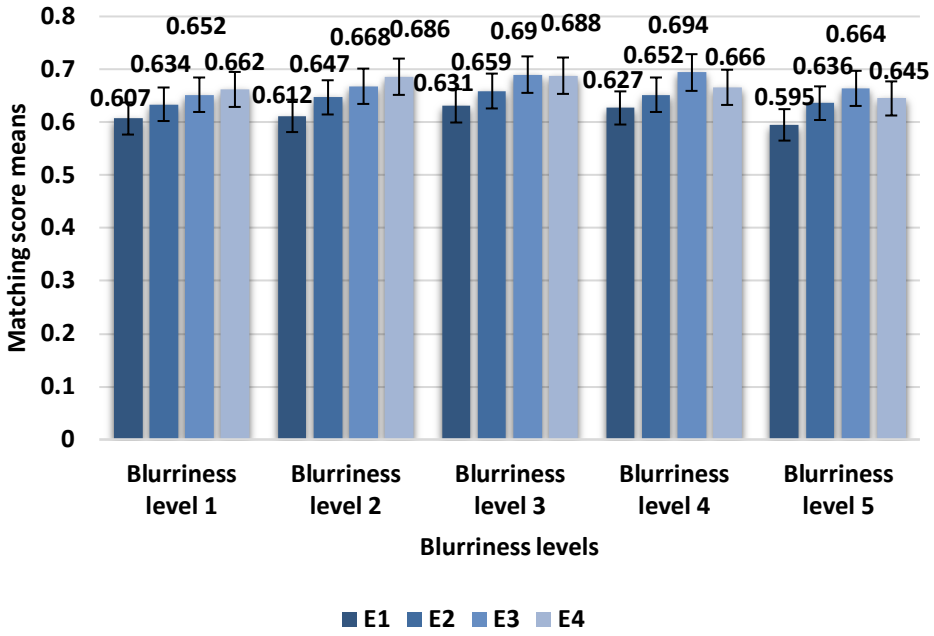


Figure 7.27: Matching score means for Face_recognition according to the level of Blurriness.

Level 3 and 4 are the blurriness values that most obtained higher scores in each enrolment scenarios, with the exception of E4. Furthermore, in this algorithm, it seems that the enrolment scenario is more affected by the highest level of blurriness, recording higher scores for levels 2 and 3 instead.

A One-way ANOVA and corresponding multiple comparisons with Tukey test confirmed the significance of this trends (Table 7.26).

Table 7.26: One-way ANOVA results for group comparison across Blurriness levels.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	E1	F(4,335) = 73.86 at p < 0.001
	E2	F(4,337) = 32.74 at p < 0.001
	E3	F(4,7534) = 49.46 at p < 0.001
	E4	F(4,335) = 43.02 at p < 0.001
Face_recognition	E1	F(4,330) = 38.5 at p < 0.001
	E2	F(4,7534) = 17.13 at p < 0.001
	E3	F(4,332) = 64.75 at p < 0.001
	E4	F(4,331) = 27.28 at p < 0.001

The trends were also reflected on the FRR. E1 as always is the enrolment scenarios that recorded the highest number of FRRs, but the trends are similar as for the others, with lower FRRs recorded for Level 3 or 4 of Blurriness. E4 instead recorded lower FRRs when the level of Blurriness was lower.

Table 7.27: FRR for Blurriness levels.

Blurriness level	VeriLook 10.0				Face_recognition			
	E1	E2	E3	E4	E1	E2	E3	E4
1	14.3%	5.4%	7.1%	0.9%	0%	0%	0%	0.9%
2	13.2%	4.8%	3.3%	2.1%	0.01%	0.1%	0.1%	0.1%
3	6.4%	2.5%	1%	2.8%	0.2%	0.1%	0.01%	0.1%
4	8.5%	2.8%	0.5%	4.5%	0.1%	0.1%	0.1%	0.1%
5	20.3%	2.9%	2.9%	11.6%	0%	0%	0%	0%

We can summarise that an extreme case in this FIQ metric brings consequences. Even when there is not Blur in the image, a too-sharp image resulted with have a negative effect on verification performance. The best compromise is to find a level of blurriness that is considered between 2 and 3.

7.5.5 Exposure

Exposure was more skewed to higher values, so since the metrics were reported on the same scale, the level of this variables were considered as for the other metrics, with the consequence of no images presenting a Level 1 of Exposure. The chart in Figure 7.28 shows the mean values for matching scores calculated with VeriLook 10.0.

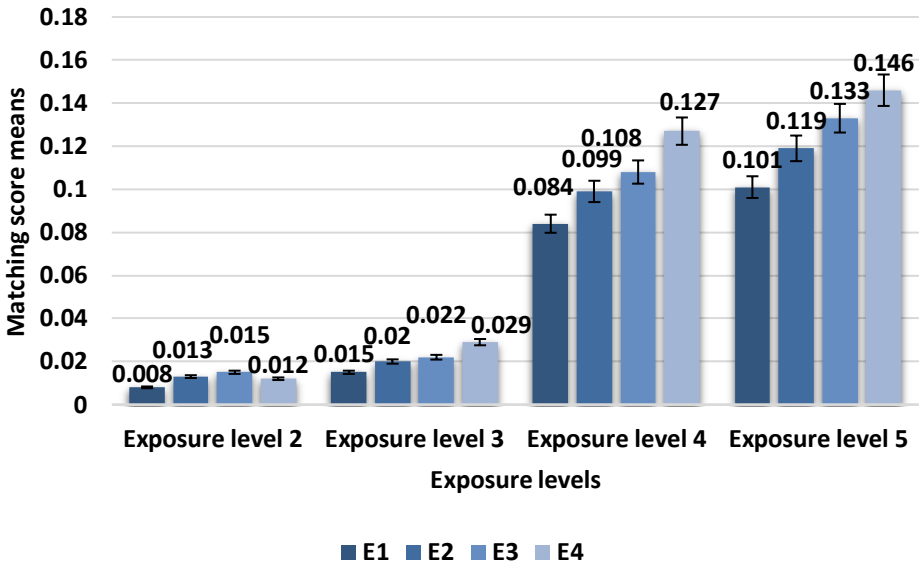


Figure 7.28: Matching score means for VeriLook 10.0 according to the level of Exposure.

The algorithms seem to receive a huge effect from this metric when the level of Exposure is low (level 2 and 3), more than for Face_recognition algorithm as shown in Figure 7.29.

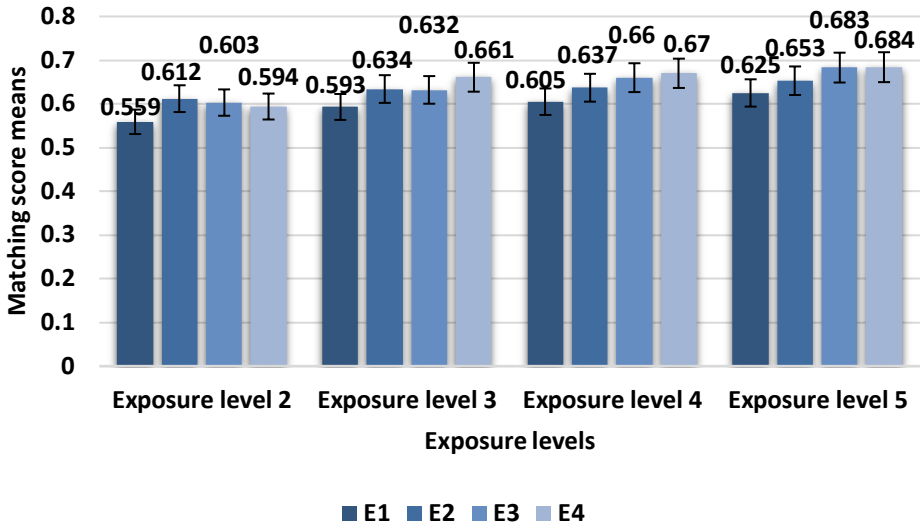


Figure 7.29: Matching score means for Face_recognition according to the level of Exposure.

The trend is the same for both algorithms and for each enrolment scenarios: the higher the Exposure level, the better the performance. The differences between each group considering the Exposure level reported a significant difference using a One-way ANOVA test, and the post-hoc multiple comparisons underline the same trend for all the scenarios (Table 7.28), with a gradual increase for VeriLook 10.0 and a sharper increase for the Face_recognition algorithm.

Table 7.28: One-way ANOVA results for group comparison across Exposure levels.

Verification systems	Enrolment scenarios	One-way between-groups ANOVA
VeriLook 10.0	**E1**	$F(3,7910) = 101.07$ at $p < 0.001$
	E2	$F(3, 7910) = 81.42$ at $p < 0.001$
	E3	$F(3, 7910) = 83.17$ at $p < 0.001$
	E4	$F(3, 7910) = 34.50$ at $p < 0.001$
Face_recognition	**E1**	$F(3, 7910) = 23.59$ at $p < 0.001$
	E2	$F(3,23) = 11.38$ at $p < 0.001$
	E3	$F(3, 7910) = 39.92$ at $p < 0.001$
	E4	$F(3,22) = 26.04$ at $p < 0.001$

According to the results, to ensure higher verification performance there should not be extreme lower values of Exposure in the image. When comparing the FRR, shown in Table 7.29, higher FRR can be seen for higher level of Exposure. This could be explained by the higher number of images that presented higher Exposure, increasing the probability of having an FRR compared to the few images that presented lower levels of Exposure. The trend of an improvement between lower levels and higher ones is still evident for level 4 and 5.

Table 7.29: FRR for Exposure levels.

Exposure level	VeriLook 10.0				Face_recognition			
	E1	E2	E3	E4	E1	E2	E3	E4
1	-	-	-	-	-	-	-	-
2	100%	0%	0%	0%	0%	0%	0%	0%
3	42.9%	0%	0%	0%	0%	0%	0%	0%
4	16.1%	6.2%	2.6%	4.3%	0.2%	0.2%	0.2%	0.3%
5	8.3%	3%	1.7%	2.7%	0.1%	0.1%	0.1%	0.1%

It can be concluded that to enhance the verification performance, the level of Exposure should be between 4 and 5, as an increase in the matching scores was observed for a higher level of Exposure.

7.6 Mobile facial verification: overall observations

This Chapter assessed biometric performance across a different range of variables. The matching scores obtained from the two considered verification algorithms were different even if normalised to the same scale, probably due to the different methods used for image comparisons. Nevertheless, it was possible to observe similar behaviours within our analysis, enabling to formulate general observations regarding the variations added from the considered variables to the verification performance.

The analysis was carried out considering four enrolment scenarios. It was observed from the results that the enrolment scenario that performed better across the different conditions was E3, which considers images selected in indoors locations.

Verification matching scores were compared across the three sessions considered in the experimental data collection. It was seen that the verification performance decreased within the three session. It could be possible that the repetitive task assessed by the participants for taking images during the data collection had a negative impact in the system performance. Moreover, participants did not receive a feedback from the system, and this probably affected the performance since there was not an interested not an indication on how improve the facial image presentation during the data collection.

Interesting observations were made when considering differences between the environmental locations. The matching scores did not significantly differ when considering verification images taken indoors or outdoors. The location types affected more the performance when considering the differences within the enrolment scenarios more than the verification conditions. When verification images were compared to E1, the results were lower than for the other scenarios. E3 and E4 were instead the enrolment scenarios that reported the best performance.

Significant differences were observed when comparing the matching scores across demographics groups and when considering the user's static and dynamic characteristics. The magnitude of the differences reported mainly small values, indicating a low effect that these variables have on verification performance. However, when a larger population of users is involved, as considered for the application of facial verification on mobile devices, the size effect of the observed differences can have a bigger impact. The observations shown in this analysis can be used to understand the relationship within the variables and could be used to adjust the thresholds for the verification system accordingly.

Facial expressions were detected and assessed to understand how they affect performance of the verification system. Results reported that the most detected facial expression was Neutral as it resulted in the highest matching scores. However, there were interesting results observed from Surprise and Happiness, while lower performance was recorded for facial expressions that are usually associated with negative emotions, like Sadness and Anger.

The analysis also considered the users' angular poses, according to the yaw, pitch and roll rotations, that were studied in regards to the verification performance. When applying the ISO/IEC 19794-5 Standard for compliance with facial pose angles, a good percentage of images that would result in a "Successful" verification would be erroneously discarded, especially due to the wide variations reported for yaw angles in the context of mobile devices. The requirements specified for head pose should be adapted to the variations that the smartphone scenarios bring over the user pose.

Finally, the quality assessment with biometric performance reported fundamental results, indicating the level of the quality metric that should be ranged to obtain the higher performance. Brightness, Contrast and Exposure presented the highest performance when higher levels of the quality metrics were observed in the images. GCF and Blurriness presented instead the best performance when values were not extremely low (level 1) or extremely high (level 5).

Conclusions and future work

8.1 Introduction

This work presented a complete assessment of facial image quality and user interaction of facial verification system applied to a mobile scenario. The aim was to understand the relationship between the variables that can affect a facial verification system concerning quality and biometric performance. The results and the observations that were considered in the analysis are a contribution to the state-of-the-art guidelines and best practice to enhance the system performance.

The novelty of this work is the assessment of user interaction and image quality over a dataset that comprises of images taken with a smartphone camera in the unconstrained environment by unsupervised users. The locations considered were representative of real-life environmental scenarios. The study also proposed innovative approaches to assess the user interaction and the acquisition process by analysing the data collected through the sensor available and implemented on the mobile device.

The results obtained were compared to conditions similar to those for passport scenarios, as the majority of the documentation and best practice of facial recognition is based on passport images. The user and the camera were considered in this study as they are an integral part of the system. Results underline the importance of considering these two aspects in the context of mobile devices, by proposing observations and requirements for designing and implementing facial verification systems on smartphones.

This Chapter presents a comprehensive summary of the observations and findings of this work, including lessons learned and considerations that should be addressed in future research.

8.2 Thesis contributions

Initially, the study presented an assessment of the users' perspectives of biometric technologies in the context of mobile devices. From the online survey, it was observed that there is still a high number of participants that are not aware or not sure of storing sensitive data on their mobile devices. Educating the users about the nature of personal data stored on the device and the risks they might occur in terms of security is of critical importance as it was shown from the results that the awareness of this information does affect the attitude that users have towards the security adoption.

An encouraging shift in the acceptance of biometric authentication on mobile devices was noted from the results, especially fingerprint verification, as it was considered more trustworthy than PINs and passwords in specific real-life application scenarios presented

in the survey. However, other biometric modalities, such as voice and face verification, resulted in being more unlikely to be used on mobile devices. Face recognition, in particular, despite being already implemented and used in the mobile context, presented a low percentage of subjects that experienced it or that are willing to adopt it in real-life applications.

Face verification systems present a series of challenges when implemented on mobile devices that can affect the user's perception and acceptance. Based on the outcome of the online survey we focused our attention on the considerations to obtain the higher performance of facial verification systems, in particular considering the user interaction and the quality assessment. The analysis was performed to investigate what variables affect verification performance and how to adapt them to ensure high system performance. With this aim, an experimental database was collected to enable an analysis of the environmental variables affecting the system, as well as providing an innovative approach to investigate user interaction and camera movements.

As seen in previous work, there is not a unique way to assess quality. The ISO/IEC 24979-5 TR describes quality metrics that can be used for quality assessment; we selected five metrics that were commonly used across the state-of-the-art. One of the main contributions of this work was providing a general observation on how each FIQ metric variates in realistic mobile scenarios. The quality metrics were normalised over the same scale from 0 to 5, to provide a comparison between the different range observed for each metric and to identify the levels for which the system presented higher performance.

Two verification algorithms were considered to assess the system performance: the aim was not to compare the accuracy and performance of the selected algorithms, but rather to provide a general perspective by observing common variations from the analysis. The matching scores were obtained from genuine users comparisons between four different enrolment scenarios and each verification image collected from the participants.

Higher matching scores were obtained for a higher level of Brightness, Contrast and Exposure, while GCF and Blurriness reported better performance when the images presented a medium-range between 2 and 4. These observations were valid for both the verification algorithms. When considering the detection of the facial area in an image, Contrast and Blurriness were the two metrics that most contributed in estimating the detection outcome, as observed across all the algorithm considered in the analysis. Table 8.1 indicates the values for which each quality metric should range in order to obtain better biometric verification performance. Future biometric system developers can base their model considering these values for the quality of facial area images. The quality thresholds can be adjusted depending on the facial verification algorithm used: within these ranges the system will adapt to the requirements for the specific scenario but without losing the quality necessary to ensure high performance.

Table 8.1: Quality metrics values to ensure high verification performance.

Quality metrics	Level on normalised scale	Values
B	Higher than 2 (3:4)	Higher than 80.74 (118.10:155.46)
C	Higher than 2, (4:5)	Higher than 7.59 (10.90:12.55)
GCF	From 2 to 4	From 3.23 to 7.92
Blur	From 2 to 4	From 0.35 to 0.49
E	Higher than 3 (4:5)	Higher than 6.42 (7.20:7.97)

When implementing facial biometric systems on mobile devices, checking the quality metrics is not enough, as there are other factors, as the user and camera's static and dynamic characteristics, that need to be taken into account. This research has allowed to understand which element influenced the quality metrics and in which way so that future application developers can consider these effects when implementing a biometric system. Table 8.2 summarise the list of characteristics that were described and analysed in this study.

Table 8.2: Effects that user and camera's characteristics present over quality.

User and camera's characteristics	Affecting quality
Presence of glasses	Higher metrics values overall, In particular high GCF level
Presence of heavy make-up	Higher metrics values overall, In particular high Contrast level
Presence of facial hair	Lower metrics values overall, In particular, low Contrast level
Blink	Did not affect
Mouth opened	Did not affect
ISO	Should be between ISO 100-800 Affecting particularly Blurriness
Light Values	Should be between 11.10-16.90 Affecting particularly Blurriness
Camera movements	Should not register peaks over 2 m/s^2

When designing a biometric system, the requirements set for the quality metrics to ensure high performance need to be flexible in order to include the variations introduced by the user and camera's characteristics. For example, when setting a threshold for the maximum value required for GCF, the developer needs to consider that when the user present glasses in the image, the quality value will increase compared to when the characteristic is not present. Therefore, the threshold should be set so that the genuine user is not excluded based only on the quality score.

Main considerations from this work can be summarised as follows:

- **Environmental factors:** when considering the environmental locations selected for the verification images, indoors and outdoors did not report significant results. However, differences were observed within the enrolment scenarios. The verification algorithms reported the highest matching scores when the enrolment scenarios were similar to the environmental conditions

in which the verification images were presented. These results could lead to a new idea of having an enhanced enrolment image for better results to give space to an enrolment that comprises a higher variation of images that could occur in realistic scenarios.

- **Background analysis:** the image background can provide useful information for estimating the detection of an image. The texture assessment that was proposed in this work demonstrates that it is possible to estimate whether the biometric presentation is occurring in an indoor or outdoor location and estimate the FTDs considering the complexity level of the background.
- **Users and camera characteristics:** the study presented an assessment of the variations introduced by the user and the camera over face detection and recognition. The observations indicated in this work will help the understanding of the variations that image quality and biometric performance receive from each variable and adapt the thresholds accordingly considering specific applications.

The analysis underlines issues with the current Standards and best practice, that needs to be adapted to the specific context of mobile devices. The quality assessment presented in this work demonstrates that the FIQ metrics selected presented different values between the camera types even when the images were taken under the same conditions that were similar to the passport image enrolment scenarios. This consideration is of critical importance, especially considering the future application might consider the smartphone as a passport for cross boarding at the airport.

Other differences assessed in this study involved the requirements for user's facial expressions and head pose. It was demonstrated that not only the Neutral expression can achieve high matching scores, but that also facial expressions that present "positive" emotions, like Happiness, can record high performance. The lower performance was instead associated with the more "negative" emotions like Anger and Sadness. It was also shown that head angular rotations requirements formulated for passport scenarios could not be applied in a mobile environment since the variation in the pose is higher in this scenario. The exclusions over the images that do not conform with the requirements also imply the rejection of "Successful" detected or verified facial images.

Furthermore, innovative methods were proposed to assess user interaction and camera movements:

- The accelerometer data was used to extract features that enabled an assessment of camera movements during image acquisition. The quality metrics were studied to see how the level of estimated movements affected the images. A higher level of movements resulted in an increase of Blurriness and a decrease of Brightness and Exposure.
- Assessing the user's opinions and experience is fundamental to understand the acceptability and usability of biometric systems. The quality of the images

and the detection of facial areas were both aspects that were influenced by the user's opinions during the data collection. Participants expressed concerns in taking images with adverse weather conditions, in particular, rain and wind. The subjects that reported feeling uncomfortable in taking the images when other people were present resulted in a higher number of FTDs. Understanding how users feel towards technology is a crucial point for solving the acceptability issue that this biometric modality is still currently presenting.

Moreover, issues and new challenges were identified in this study that should be addressed in future research.

8.3 Lessons learned and future work

The main idea when collecting the data was to assess the user interaction using the embedded sensors available on the device. The focus was on the sensors that could be available in the most common devices. The accelerometer and the gyroscope are the two main sensors that can be found in the majority of the devices. However, the information provided from these two sensors is similar and when used singularly does not allow a proper distinction between the user's movements to adjust the camera and whether the subject is walking while presenting the images.

The ActivityRecognition API from Android was considered to collect information regarding the user's movements to compared to the accelerometer data, but the accuracy of the Activity estimation was not enough to allow analysis over this aspect. Therefore, future research will focus on assessing sensing data collected from more embedded sensors to enable a more accurate analysis of the user interaction. Moreover, the magnetometer information combined with the gyroscope and accelerometer could be used to extract estimations about the pitch, yaw and roll angles of the smartphone device, as this was not possible to be adequately estimated over our database using only gyroscope and accelerometer data.

Other aspects that need to be included in the future analysis is the effect of providing feedback to the users, either by indicating a quality score or by real-time verification using a biometric system. The feedback provided could result in an improvement or a decrease over the performance across multiple sessions. Finally, another important aspect for future research would be to extend the observations made in this work to different smartphone cameras, to ensure that there is consistency within the results.

References

[1] C.-Y. Chen, B.-Y. Lin, J. Wang, and K. G. Shin, "Keep Others from Peeking at Your Mobile Device Screen!," *25th Annu. Int. Conf. Mob. Comput. Netw. (MobiCom '19)*, no. October 21–25, 2019.

[2] R. Blanco-Gonzalo and R. Sanchez-Reillo, "Biometrics on Mobile Devices," in *Encyclopedia of Biometrics*, Boston, MA: Springer US, 2015.

[3] M. F. Theofanos, B. C. Stanton, and C. Wolfson, "Usability and Biometrics: Ensuring Successful Biometrics Systems," *Int. Work. Usability Biometrics*, pp. 37–105, 2008.

[4] E. Vazquez-Fernandez and D. Gonzalez-Jimenez, "Face recognition for authentication on mobile devices," *Image Vis. Comput.*, vol. 55, pp. 31–33, 2016.

[5] L. J. Karam and T. Zhu, "Quality labeled faces in the wild (QLFW): a database for studying face recognition in real-world environments," in *Human Vision and Electronic Imaging XX*, 2015, vol. 9394, p. 93940B.

[6] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in Face Detection and Facial Image Analysis*, 2016, pp. 189–248.

[7] P. Editor, "ISO/IEC 29794 Biometric Sample Quality — Part 5: Face image Data Sample Quality - Technical Report -," *Image (Rochester, N.Y.)*, 2009.

[8] ISO/IEC, "ISO/IEC 19794-5 Biometric Data Interchange Formats — Part 5: Face Image Data," no. 30, pp. 1–522, 2004.

[9] J. Sang, Z. Lei, and S. Z. Li, "Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5558 LNCS, pp. 229–238.

[10] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing face recognition vendor test (FRVT) part 2:," Gaithersburg, MD, Nov. 2018.

[11] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test Ongoing Face Recognition Quality Assessment Concept and Goals VERSION 1.0," 2019.

[12] M. Abdel-Mottaleb and M. H. Mahoor, "Algorithms for assessing the quality of facial images," *IEEE Comput. Intell. Mag.*, vol. 2, no. 2, pp. 10–17, 2007.

[13] M. Turk and A. Pentland, "Eigenfaces for recognition.," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[14] National Institute of Standards and Technology, "Face Recognition Technology (FERET)," *National Institute of Standards and Technology*, 2017.

[15] X. Gao, S. Z. Li, R. Liu, and P. Zhang, “Standardization of Face Image Sample Quality,” in *Advances in Biometrics*, 2007, pp. 242–251.

[16] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, 2001.

[17] P. Kocjan and K. Saeed, “Face recognition in unconstrained environment,” in *Biometrics and Kansei Engineering*, vol. 9781461456, 2012, pp. 21–42.

[18] T. Bourlai, A. Abaza, A. Ross, and M. A. Harrison, “Design and evaluation of photometric image quality measures for effective face recognition,” *IET Biometrics*, vol. 3, no. 4, pp. 314–324, 2014.

[19] P. Wasnik, K. B. Raja, R. Ramachandra, and C. Busch, “Assessing face image quality for smartphone based face recognition system,” *Proc. - 2017 5th Int. Work. Biometrics Forensics, IWBF 2017*, pp. 1–6, 2017.

[20] Q. C. Truong, T. K. Dang, and T. Ha, “Face Quality Measure for Face Authentication,” in *Future Data and Security Engineering*, 2017, vol. 10646.

[21] T. Allen, “Visualization and Usability Group,” 2008.

[22] ISO, “ISO 9241-11, Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts,” p. 1, 1998.

[23] R. Blanco-Gonzalo, L. Diaz-Fernandez, O. Miguel-Hurtado, and R. Sanchez-Reillo, “Usability evaluation of biometrics in mobile environments,” *2013 6th Int. Conf. Hum. Syst. Interact.*, pp. 123–128, 2013.

[24] R. Blanco-Gonzalo, R. Sanchez-Reillo, O. Miguel-Hurtado, and J. Liu-Jimenez, “Usability analysis of dynamic signature verification in mobile environments,” *Biometrics Spec. Interes. Gr. (BIOSIG), 2013 Int. Conf.*, pp. 1–9, 2013.

[25] V. Conti, M. Collotta, G. Pau, and S. Vitabile, “Usability analysis of a novel biometric authentication approach for android-based mobile devices,” *J. Telecommun. Inf. Technol.*, vol. 2014, no. 4, pp. 34–43, 2014.

[26] O. Miguel-Hurtado, R. Guest, and C. Lunerti, “Voice and face interaction evaluation of a mobile authentication platform,” *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2017-Octob, pp. 1–6, 2017.

[27] Z. Sitova *et al.*, “HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users,” *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 5, pp. 877–892, 2016.

[28] J. Bonneau, “The science of guessing: Analyzing an anonymized corpus of 70 million passwords,” *Proc. - IEEE Symp. Secur. Priv.*, no. Section VII, pp. 538–552, 2012.

[29] S. Uellenbeck, M. Dürmuth, C. Wolf, T. Holz, H. Görtz, and R. Bochum, “Quantifying the Security of Graphical Passwords : The Case of Android Unlock Patterns Categories and Subject Descriptors,” *CCS ’13 Proc. 2013 ACM SIGSAC Conf.*

Comput. Commun. Secur., pp. 161–172, 2013.

[30] C. Sun, Y. Wang, and J. Zheng, "Dissecting pattern unlock: The effect of pattern strength meter on pattern selection," *J. Inf. Secur. Appl.*, vol. 19, no. 4–5, pp. 308–320, 2014.

[31] iMore, "How Touch ID works: Making sense of Apple's fingerprint identity sensor | iMore," 2013. [Online]. Available: https://www.imore.com/how-touch-id-works. [Accessed: 12-Dec-2018].

[32] Molly McLaughlin, "Using Google Smart Lock on Your Android Device," *18 September*, 2018. [Online]. Available: https://www.lifewire.com/using-google-smart-lock-on-android-121682. [Accessed: 12-Dec-2018].

[33] Android help, "Set your Android device to automatically unlock - Android Help." [Online]. Available: https://support.google.com/android/answer/9075927?visit_id=1-636154992314693019-839047886&hl=en&rd=2#facial_recognition. [Accessed: 12-Dec-2018].

[34] Samsung, "Camera | Samsung Galaxy S8." [Online]. Available: https://www.samsung.com/uk/smartphones/galaxy-s8/security/. [Accessed: 12-Dec-2018].

[35] Samsung, "Specifications | Samsung Galaxy S9 and S9+ – The Official Samsung Galaxy Site." [Online]. Available: https://www.samsung.com/global/galaxy/galaxy-s9/performance/. [Accessed: 06-Apr-2019].

[36] Apple Inc., "Use Touch ID on iPhone and iPad - Apple Support," 2015. [Online]. Available: https://support.apple.com/en-gb/HT201371. [Accessed: 12-Dec-2018].

[37] I. Apple, "iPhone XS - Face ID - Apple," 2019. [Online]. Available: https://www.apple.com/uk/iphone-xs/face-id/. [Accessed: 06-Apr-2019].

[38] "How the Iphone X has inspired the notch age - TECH NAV." [Online]. Available: http://technav.co.uk/iphone-x-inspired-notch-age/. [Accessed: 06-Apr-2019].

[39] Mauro Huculak, "How the iris scanner on the Lumia 950 and 950 XL works | Windows Central," 2015. [Online]. Available: https://www.windowscentral.com/how-iris-scanner-lumia-950-and-950-xl-works. [Accessed: 12-Dec-2018].

[40] N. Micallef, M. Just, L. Baillie, M. Halvey, and H. G. Kayacik, "Why aren't Users Using Protection? Investigating the Usability of Smartphone Locking," *Proc. 17th Int. Conf. Human-Computer Interact. with Mob. Devices Serv. - MobileHCI '15*, pp. 284–294, 2015.

[41] NIST, "Usability and Biometrics: Ensuring Successful Biometric Systems," *Des. Perform. Biometric Syst.*, pp. 37–105, 2008.

[42] SurveyMonkey, "SurveyMonkey: The World's Most Popular Free Online Survey Tool," *SurveyMonkey*, 2018. [Online]. Available:

https://www.surveymonkey.com/. [Accessed: 12-Dec-2018].

[43] S. M. Furnell, P. S. Dowland, and H. M. Illingworth, “Authentication and Supervision: A Survey of User Attitudes,” *Comput. Secur.*, vol. 19, pp. 529–539, 2000.

[44] F. Schaub, R. Deyhle, and M. Weber, “Password entry usability and shoulder surfing susceptibility on different smartphone platforms,” *Proc. 11th Int. Conf. Mob. Ubiquitous Multimed. - MUM ’12*, p. 1, 2012.

[45] J. Moody, “Public perceptions of biometric devices: The effect of misinformation on acceptance and use,” *J. Issues Informing Sci. Inf. Technol.*, pp. 753–761, 2004.

[46] L. A. Jones, A. I. Antón, and J. B. Earp, “Towards understanding user perceptions of authentication technologies,” in *Proceedings of the 2007 ACM workshop on Privacy in electronic society*, 2007, pp. 91–98.

[47] A. Morales, M. A. Ferrer, C. M. Travieso, and J. B. Alonso, “About user acceptance in hand, face and signature biometric systems,” *Proc. Work. Tecnol. multibiométricas para la identificación Pers. (1er WTM-IP)*, pp. 14–17, 2010.

[48] M. El-Abed, R. Giot, B. Hemery, and C. Rosenberger, “A study of users’ acceptance and satisfaction of biometric systems,” *Proc. - Int. Carnahan Conf. Secur. Technol.*, pp. 170–178, 2010.

[49] M. El-abed, C. Charrier, M. El-abed, C. Charrier, B. Systems, and M. El-abed, “Evaluation of Biometric Systems,” *IEEE Int. Conf. Hand-Based Biometrics*, pp. 149–169, 2014.

[50] R. Blanco-Gonzalo, R. Sanchez-Reillo, R. Ros-Gomez, and B. Fernandez-Saavedra, “User acceptance of planar semiconductor fingerprint sensors,” *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2015-Janua, pp. 31–36, 2015.

[51] N. L. Clarke and S. M. Furnell, “Authentication of users on mobile telephones - A survey of attitudes and practices,” *Comput. Secur.*, vol. 24, no. 7, pp. 519–527, 2005.

[52] F. Breitinger and C. Nickel, “User Survey on Phone Security and Usage,” *BIOSIG 2010 Biometrics Electron. Signatures. Proc. Spec. Interes. Gr. Biometrics Electron. Signatures*, no. May 2010, pp. 139–144, 2010.

[53] C. Bhagavatula, B. Ur, K. Iacovino, S. M. Kywe, L. F. Cranor, and M. Savvides, “Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption,” 2015.

[54] M. Harbach, A. De Luca, N. Malkin, and S. Egelman, “Keep on Lockin’ in the Free World,” 2016, pp. 4823–4827.

[55] S. McLeod, “Likert Scale | Simply Psychology,” 2008. [Online]. Available: https://www.simplypsychology.org/likert-scale.html. [Accessed: 12-Dec-2018].

[56] “SailfishOS - Sailfish OS.” [Online]. Available: https://sailfishos.org/. [Accessed: 12-Dec-2018].

[57] “Symbian Open Source : Symbian Foundation : Free Download, Borrow, and Streaming : Internet Archive.” [Online]. Available: https://archive.org/details/SymbianOpenSource. [Accessed: 12-Dec-2018].

[58] “Data Protection Act 1998.”

[59] “Data Protection Act 2018.” [Online]. Available: https://www.gov.uk/data-protection. [Accessed: 13-Dec-2018].

[60] Iritech, “ForYourIrisOnly | Iris Scanner | Iris Biometrics Technology | Iris Recognition.” [Online]. Available: http://www.iritech.com/products/software/foryouririsonly-iris-recognition-software. [Accessed: 12-Dec-2018].

[61] H. Lu, J. Huang, T. Saha, and L. Nachman, “Unobtrusive gait verification for mobile phones,” *Proc. 2014 ACM Int. Symp. Wearable Comput. - ISWC ’14*, pp. 91–98, 2014.

[62] D. Kim, Y. Jung, K. A. Toh, B. Son, and J. Kim, “An empirical study on iris recognition in a mobile phone,” *Expert Syst. Appl.*, vol. 54, pp. 328–339, Jul. 2016.

[63] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, “Continuous User Authentication on Mobile Devices: Recent progress and remaining challenges,” *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 49–61, Jul. 2016.

[64] P. S. Teh, N. Zhang, A. B. J. Teoh, and K. Chen, “A survey on touch dynamics authentication in mobile devices,” *Comput. Secur.*, vol. 59, pp. 210–235, Jun. 2016.

[65] R. Ferrero, F. Gandino, B. Montrucchio, M. Rebaudengo, A. Velasco, and I. Benkhelifa, “On gait recognition with smartphone accelerometer,” *Proc. - 2015 4th Mediterr. Conf. Embed. Comput. MECO 2015 - Incl. ECyPS 2015, BioEMIS 2015, BioICT 2015, MECO-Student Chall. 2015*, pp. 368–373, 2015.

[66] Adam C. Uzialko, “Google Pay vs. Apple Pay vs. Samsung Pay,” *Business.com*. [Online]. Available: https://www.business.com/articles/google-pay-vs-apple-pay-vs-samsung-pay/. [Accessed: 10-Apr-2019].

[67] “Canon EOS 30D.” [Online]. Available: http://web.canon.jp/imaging/eos30d/spec/index.html. [Accessed: 23-Feb-2019].

[68] “Google Nexus 5 Specifications.” [Online]. Available: https://www.ubergizmo.com/products/lang/en_us/devices/nexus-5/. [Accessed: 23-Feb-2019].

[69] “Canon EOS 30D Review: Digital Photography Review.” [Online]. Available: https://www.dpreview.com/reviews/canoneos30d. [Accessed: 01-Apr-2019].

[70] “Google Nexus 5 review | TechRadar.” [Online]. Available: https://www.techradar.com/reviews/phones/mobile-phones/google-nexus-5-1198568/review. [Accessed: 01-Apr-2019].

[71] “Android Studio features.” [Online]. Available: https://developer.android.com/studio/features. [Accessed: 23-Feb-2019].

[72] “Research Ethics - Faculty of Social Sciences - University of Kent.” [Online]. Available: https://www.kent.ac.uk/socsci/faculty/research-ethics/index.html. [Accessed: 23-Feb-2019].

[73] National Institutes of Health (NIH), “Racial and ethnic categories and definitions for NIH diversity programs and for other reporting purposes,” *Notice Number: NOT-OD-15-089*, 2015. [Online]. Available: https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html. [Accessed: 22-Feb-2019].

[74] “DetectedActivity | Google APIs for Android | Google Developers.” [Online]. Available: https://developers.google.com/android/reference/com/google/android/gms/location/DetectedActivity. [Accessed: 24-Feb-2019].

[75] “How To Disable Proximity Sensor In Any Android.” [Online]. Available: https://techviral.net/how-to-disable-proximity-sensor-android/. [Accessed: 01-Apr-2019].

[76] “Proximity Actions 3.09 apk | androidappsapk.co.” [Online]. Available: https://androidappsapk.co/detail-novum-proximity-actions/. [Accessed: 01-Apr-2019].

[77] P. Viola and M. J. Jones, “Robust Real-Time Face Detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[78] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, “Representation and Description,” in *Digital Image Processing Using MATLAB*, Prentice Hall: Upper Saddle River, 2004, pp. 644–656.

[79] Geolocation API - Google, “Developer Guide | Geolocation API | Google Developers.” [Online]. Available: https://developers.google.com/maps/documentation/geolocation/intro?fbclid=IwAR3fbh8kyWJEh_v8bAEig5DTGm3PNM2UC-SdnBHlWg057X8gF8NIAjYfqP4. [Accessed: 04-Apr-2019].

[80] Neurotechnology, “VeriLook Face Verification SDK,” 2017. [Online]. Available: https://www.neurotechnology.com/verilook-technical-specifications.html. [Accessed: 21-Mar-2019].

[81] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, “Robust discriminative response map fitting with constrained local models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3444–3451.

[82] Marcus Hawkins, “The Exposure Triangle: aperture, shutter speed and ISO explained | TechRadar,” 2017. [Online]. Available: http://www.techradar.com/how-to/photography-video-capture/cameras/the-exposure-triangle-aperture-shutter-speed-and-iso-explained-1320830. [Accessed: 18-Mar-2019].

[83] P. Fernandez-Lopez, J. Sanchez-Casanova, P. Tirado-Martin, and J. Liu-Jimenez,

"Optimizing resources on smartphone gait recognition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 31–36.

[84] "Biometric Image Autocapture for Facial Recognition - PreFace." [Online]. Available: https://www.aware.com/biometrics/preface/. [Accessed: 09-Apr-2019].

[85] A. Geitgey, "Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning." [Online]. Available: https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78. [Accessed: 09-Apr-2019].

[86] K. Matković, L. Neumann, A. Neumann, T. Psik, and W. Purgathofer, "Global contrast factor - a new approach to image contrast," *Proc. First Eurographics Conf. Comput. Aesthet. Graph. Vis. Imaging*, pp. 159–167, 2005.

[87] F. Crété-Roffet, T. Dolmiere, P. Ladret, M. Nicolas, and F. Crete, "The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric," *Hum. Vis. Electron. imaging XII*, vol. 6492, no. International Society for Optics and Photonics, p. 64920I, 2007.

[88] "dlib C++ Library." [Online]. Available: http://dlib.net/. [Accessed: 10-Apr-2019].

[89] "OpenFace." [Online]. Available: https://cmusatyalab.github.io/openface/. [Accessed: 10-Apr-2019].

[90] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection."

[91] K. Vahid and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 1867–1874, 2014.

[92] C. Thongleng and W. Kaewapichai, "Case Studies to Improve Viola-Jones for Eye Detection," 2018.

[93] H. Lahiani, M. Kherallah, and M. Neji, "Hand pose estimation system based on Viola-Jones algorithm for Android devices," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, pp. 1–6, 2017.

[94] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2879–2886, 2012.

[95] M. Saquib Sarfraz and O. Hellwich, "Head pose estimation in face recognition across pose scenarios," 2011, pp. 235–242.

[96] A. L. Whitehead, S. A. Julious, C. L. Cooper, and M. J. Campbell, "Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable," *Stat. Methods Med. Res.*, vol. 25, no. 3, pp. 1057–1073, 2016.

[97] "Front, not back cameras, will be the new big smartphone trend | TechRadar." [Online]. Available: https://www.techradar.com/uk/news/front-not-back-

cameras-will-be-the-new-big-smartphone-trend. [Accessed: 28-Apr-2019].

[98] “Exposure program chart - File:Exposure program chart.gif - Wikimedia Commons.” [Online]. Available: https://commons.wikimedia.org/wiki/File:Exposure_program_chart.gif#/media/File:Exposure_program_chart.gif. [Accessed: 12-May-2019].

[99] “Film Photography - Fortismere Art Department.” [Online]. Available: https://fortismereartdepartment.weebly.com/film-photography.html. [Accessed: 12-May-2019].