UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Usability in Biometric Recognition Systems

Autor:
Ramón Blanco Gonzalo

Director:
Raúl Sánchez Reíllo
Co-Directora:
Judith Liu Jiménez

DEPARTAMENTO DE TECNOLOGÍA ELECTRÓNICA

L , F 2016

PhD Thesis

Usability in Biometric Recognition Systems

Author: Ramón Blanco Gonzalo

Advisors: Raúl Sánchez Reíllo and Judith Liu Jiménez

Signature from the Thesis Committee:

	Name and Surname	Signature
President:	Dr. PATRIZIO CAMPISI	
Vocal:	Dr. MARCOS FAUNDEZ ZANUY	
Secretary:	Dr. ENRIQUE CABELLO PARDOS	

Mark:

Leganés, February, 2016

"Si ese otro puede, ¿por qué tú no vas a poder?

Está claro que nunca correrás los 100 metros en 9 segundos…

Pero si ese otro puede, tú también puedes"

Miguel, profesor de matemáticas

"I have always wished for my computer to be as easy to use as my telephone;

My wish has come true because I can no longer figure out how to use my telephone"

Bjarne Stroustroup, IAAP

Acknowledgements / Agradecimientos

This little section is written mostly in my mother tongue as several people who may read it do not understand English. My apologies.

Este trabajo tiene muchas horas de dedicación, pero no han sido sólo mías…Sería muy injusto no escribir unas palabras de agradecimiento para aquellas personas que me han traído hasta aquí.

En primer lugar quiero agradeceyos muchu a Mari Carmen y Juanra. Ejemplu de muches coses: superación personal, esfuerzu diariu, constancia…Siempre tuve vuestru apoyu incondicional y tou lo que conseguí fue en gran parte culpa vuestra. Gracies papá y mamá. También yos debo muchu a Julio y Ana que me apoyaron cuando más lo necesité y a César que siempre me quiso enseñar a programar. Tengo mucha suerte.

Mi familia también son mis amigos Ana, Fran, Julia, Moya, Pablo, Nuria, Víctor, Sofía, Jesús, Alberto, Nico, Sergio, David, Guti y Antonio, que siempre están ahí y me llevan apoyando muchos años. Sois grandes chavales. Patri también me ha ayudado mucho en esta última parte de la tesis y espero que sea por mucho tiempo más ;)

Aunque no es una práctica muy común, tengo que dar mucho las gracias al GUTI y alabar las bonanzas de los gutitos y exgutitos con los que coincidí: Jaume, Inma, Aitor, Ram, Iván, Sandra, Pepe, Eugenio, Jorge, Eric, Dani, Rubén, Marcos, Amores, Roberto, Carlos, Inés, Helga, Pablo, Rodrigo, Wendy… ¡y los que vienen! He aprendido de todos vosotros más de lo que pensáis. Muchas gracias ☺

Especial agradecimiento para los androids Dani, Eric y Carlos, que me han ayudado en momentos clave: ¡cracks! También debo mucho a mis proyectandos Alfredo, Luis y Jose. Me habéis enseñado mucho más que yo a vosotros, ¡gracias mil!

También debo agradecer a todos aquellos biométricos con los que me he cruzado y me han enseñado tantas cosas o me han servido mucho de ejemplo…me acuerdo de Daniel Hartung, Alberto de Santos, Carmen Sánchez, Marcos Faundez, Virginia Espinosa, Enrique Cabello, Marek Rejman, Richard Guest, Andreas Nautsch, Michael Fairhurst, Patrick Grother y un largo etc…

Thanks Norman for hosting me, for your time and advices. Your dedication to research has encouraged me to improve my scientific skills. Terima kasih. Thanks also to Stathis, Veronika, Franziskus, Ali, Andrew, Nurulhuda, Aamo, Santosh and many others. All of them great researchers of the Surrey University. I had a great time with all of you guys ☺

Loïc me enseñó lo que era la accesibilidad. El capítulo 4 de esta tesis lleva mucho trabajo suyo. Gracias por todo, espero que sigamos muchos años en contacto.

GRACIAS a todos los usuarios que han participado en las evaluaciones, empleando su tiempo y esforzándose para que todo saliese bien. Gracias especiales al CRMF que han sido un gran apoyo en todo momento. Esperamos haberos ayudado aunque sea mínimamente.

Óscar es uno de los que más ha intervenido en mi tesis a nivel de trabajo y de consejos. Un buen investigador ha de saber adaptarse a los cambios y tener buenas ideas. De Óscar he aprendido a mantener los pies en el suelo y a ver las cosas con perspectiva. Gracias de verdad.

Judith fue de las primeras personas que intervino en mi tesis y una de las últimas ☺ siempre me ha dado buenos consejos y me ha ayudado en lo posible. Gran investigadora, madre, currante y amiga. Muchas buenas ideas de este trabajo vienen directa o indirectamente de Judith ¡Gracias!

Belén desarrolló la metodología sobre la que va esta tesis, me enseñó a evaluar, me enseñó de biometría, de investigación, de coser y de miles de cosas más. Siempre dispuesta a ayudar y siempre con criterio. Trabajadora incansable y gran compañera de trabajo. Habla con las ovejas. Gracias por todo ☺

De **Raúl** ha salido todo esto: el GUTI, la investigación, la biometría, las normas…Estoy muy orgulloso de poder decir que he hecho la tesis con uno de los mejores biométricos del mundo. Hemos tenido muchas conversaciones (unas de buenas formas y otras no tan buenas) y de todas he aprendido lecciones, no solo científicas, sino de vida. Siempre estaré en deuda con Raúl. GRACIAS.

Abstract

Biometric recognition, which is a technology already mature, grows nowadays in several contexts, including forensics, access controls, home automation systems, internet, etc. Now that technology is moving to mobile scenarios, biometric recognition is being also integrated in smartphones, tablets and other mobile devices as a convenient solution for guaranteeing security, complementing other methods such as PIN or passwords.

Nevertheless, the use of biometric recognition is not as spread as desired and it is still unknown for a wide percentage of the population. It has been demonstrated [1] that some of the possible reasons for the slow penetration of biometrics could be related to usability concerns. This could lead to various drawbacks like worst error rates due to systems misuses and it could end with users rejecting the technology and preferring other approaches.

This Thesis is intended to cover this topic including a study of the current state of the art, several experiments analysing the most relevant usability factors and modifications to a usability evaluation methodology. The chosen methodology is the H-B interaction, carried out by Fernandez-Saavedra [2], based on the ISO/IEC 19795 [3], the HBSI [4], the ISO 9241-210 [5] and on Common Criteria [6].

Furthermore, this work is focused on dealing with accessibility concerns in biometric recognition systems. This topic, usually included into the usability field, has been addressed here separately, though the study of the accessibility has followed the same steps as the usability study: reviewing the state of the art, pointing and analysing the main influential factors and making improvements to the state of the art. The recently published standard EN 301 549 – "*Accessibility requirements suitable for public procurement of ICT products and services in Europe*" [7] has been also analysed.

These two topics have been overcome through the well-known user-centric-design approach. In this way, first the influential factors have been detected. Then, they have been isolated (when possible) and measured. The results obtained have been then interpreted to suggest new updates to the H-B interaction. This 3-steps approach has been applied cyclically and the factors and methodology updated after each iteration.

Due to technology and usability trends, during this work, all the systems/applications developed in the experiments have been thought to be mobile directly or indirectly. The biometric modalities used during the experiments performed in this Thesis are those pointed as suitable for biometric recognition in mobile devices: handwritten recognition signature, face and fingerprint recognition. Also, the scenarios and the applications used are in line with the main uses of biometrics in mobile environments, such as sign documents, locking/unlocking devices, or make payments.

The outcomes of this Thesis are intended to guide future developers in the way of designing and testing proper usable and accessible biometrics. Finally, the results of this Thesis are being suggested as a new International Standard within ISO/IEC/JTC1/SC37 – Biometric Recognition,

as standardization is the proper way of guaranteeing usability and accessibility in future biometric systems.

The contributions of this Thesis include:

- Improvements to the H-B interaction methodology, including several usability evaluations.
- Improvements on the accessibility of the ICT (Information and Communications Technology) products by means of the integration of biometric recognition systems
- Adaptation and application of the EN 301 549 to biometric recognition systems.

Resumen

El reconocimiento biométrico, que es una tecnología ya madura, crece hoy en día en varios contextos, incluyendo la medicina forense, controles de acceso, sistemas de automatización del hogar, internet, etc. Ahora que la tecnología se está moviendo a los escenarios móviles, el reconocimiento biométrico está siendo también integrado en los teléfonos inteligentes, tabletas y otros dispositivos móviles como una solución conveniente para garantizar la seguridad, como complemento de otros métodos de seguridad como el PIN o las contraseñas.

Sin embargo, el uso del reconocimiento biométrico es todavía desconocido para un amplio porcentaje de la población. Se ha demostrado [1] que algunas de las posibles razones de la lenta penetración de la biometría podrían estar relacionadas con problemas de usabilidad. Esto podría dar lugar a diversos inconvenientes, ofreciendo un rendimiento por debajo de lo esperado debido al mal uso de los sistemas y podría terminar con los usuarios rechazando la tecnología y prefiriendo otros enfoques.

Esta tesis doctoral trata este tema incluyendo un estudio del estado actual de la técnica, varios experimentos que analizan los factores de usabilidad más relevantes y modificaciones a una metodología de evaluación de la usabilidad, la "*H-B interaction*" [2] basada en la ISO / IEC 19795 [3], el HBSI [4], la ISO 9241 [5] y *Common Criteria* [6].

Además, este trabajo se centra también en los problemas de accesibilidad de los sistemas de reconocimiento biométrico. Este tema, que por lo general se incluye en el campo de la usabilidad, se ha tratado aquí por separado, aunque el estudio de la accesibilidad ha seguido los mismos pasos que el estudio de usabilidad: revisión del estado del arte, análisis de los principales factores influyentes y propuesta de cambios en la metodología *H-B interaction*. Han sido también analizados los requisitos de accesibilidad para las Tecnologías de la Información y la Comunicación (TIC) en Europa, bajo la norma EN 301 549 [7].

Estos dos temas han sido estudiados a través de un enfoque centrado en el usuario (*User Centric Design* - UCD). De esta manera, se han detectado los factores influyentes. A continuación, dichos factores han sido aislados (cuando ha sido posible) y medidos. Los resultados obtenidos han sido interpretados para sugerir nuevos cambios a la metodología *H-B interaction*. Este enfoque de 3 pasos se ha aplicado de forma cíclica a los factores y a la metodología después de cada iteración.

Debido a las tendencias tecnológicas y de usabilidad, durante este trabajo, todos los sistemas / aplicaciones desarrolladas en los experimentos se han pensado para ser móviles, directa o indirectamente. Las modalidades utilizadas durante los experimentos realizados en esta tesis doctoral son las que se señalaron como adecuados para el reconocimiento biométrico en dispositivos móviles: la firma manuscrita, la cara y el reconocimiento de huellas dactilares. Además, los escenarios y las aplicaciones utilizadas están en línea con los principales usos de la biometría en entornos móviles, como la firma de documentos, el bloqueo / desbloqueo de dispositivos, o hacer pagos.

Los resultados de esta tesis tienen como objetivo orientar a los futuros desarrolladores en el diseño y evaluación de la usabilidad y la accesibilidad en los sistemas de reconocimiento biométrico. Por último, los resultados de esta tesis doctoral se sugerirán como un nuevo estándar de *ISO / IEC / JTC1 / SC37 - Biometric Recognition*, ya que la normalización es la manera adecuada de garantizar la usabilidad y la accesibilidad en los futuros sistemas biométricos.

Las contribuciones de esta tesis incluyen:

- Mejora de la metodología de evaluación *H-B interaction*, incluyendo varias evaluaciones de usabilidad.
- Mejora de la accesibilidad de los sistemas de información / electrónicos mediante la integración de sistemas biométricos y varias evaluaciones.
- Adaptación y aplicación de la norma de accesibilidad EN 301 549 al campo de los sistemas biométricos.

Contents

List of Figures

List of Tables

Acronyms

ABC	Automated Border Control
CC	Common Criteria for Information Technology Security Evaluation
CEN	European Committee for Standardization
CI	Concealed Interactions
DET	Detection Error Trade-off
DSV	Dynamic Signature Verification
DI	Defective Interactions
FAR	False Accept Rate
FI	False Interactions
FMR	False Match Rate
FNMR	False Non-Match Rate
FRR	False Reject Rate
FTA	Failure To Acquire rate
FTD	Failure To Detect
FTE	Failure To Enrol rate
FTP	Failure To Process
H-B	Human-Biometric system interaction
HCI	Human-Computer Interaction
HBSI	Human-Biometric Sensor Interaction
ID	Identifier
ICT	Information and Communications Technology
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
IT	Information Technology
LDA	Linear Discriminant Analysis
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
REC	Reference Evaluation Conditions
REE	Reference Evaluation Environment
ROC	Receiver Operating Characteristic
SAS	Successful Acquisition Sample
SIFT	Scale-Invariant Feature Transform
SPS	Successful Processes Sample
SDK	Software Development Kit
TEC	Target Evaluation Conditions
TEE	Target Evaluation Environment
UCD	User Centered Design
UX	User eXperience

Part I: Introduction

Chapter 1

Introduction

The amount of sensitive data and resources that need to be protected, not only at institution or company levels but also for regular citizens, is increasing exponentially. In this context the use of biometric recognition is extended worldwide as a trustable way to identify individuals and guarantee security. Biometric systems are used in many contexts such as airports, points of sale, internet or companies and its use is being increased. Moreover, biometric recognition is moving to mobile environments [8] and the range of possibilities for integrating biometrics is promising, with potential applications such as signing documents unequivocally, accessing to websites securely, executing administrative procedures and other electronic transactions. Indeed, according to market reports [9], young users do prefer biometrics embedded in mobile devices, e.g. fingerprint or face recognition to unlock the smartphone.

Unfortunately, in the intent to develop systems with high performance the users' satisfaction is most of the times placed aside. Almost all the work done in biometrics is directed to improve algorithms performance and bringing the Equal Error Rate (EER) close to zero. But while this kind of research is necessary, working on improving user interaction with systems is also extremely important, as a lack of usability could mean not only the rejection of the system by the users, but also a reduction in the expected performance of the biometric system. In order to increase the easiness and encourage the use of biometrics it is necessary to improve its usability, making biometrics reachable for a wider percent of population. Therefore, it is necessary to involve users from the first stages of the development for designing systems from the user-centric systems point of view and improve the whole throughput as a consequence [10].

One of the collectives usually excluded at the time of design security systems is the people with disabilities, who are around 15% of the world population [11]. Furthermore, it is important to highlight that every individual is potentially dependent (illnesses, age, pregnancy, etc.). Improving biometric designs would be beneficial, not only for people with disabilities, but for many others who find the technology complicated to use. It could be thought that biometric recognition is challenging for disabled people but a correct design can make the process easy for everyone.

Regarding the technology trends, it is common to use the smartphone to access bank accounts [12], make payments or handle important information in general, which leads to the necessity of increase the security in those devices [13] [14]. Usually, the applied methods to assure security in mobile devices are based on PINs or passwords, which can be easy to forget and forge, so that, other approaches are arising. In particular, biometric recognition is suggested to be embedded in mobile devices for many reasons. The first one is the large amount of devices already deployed, which has reached the situation that it is difficult to find someone that does not possess and use mobile devices daily such as smartphones or tablets. The second one is that for some biometric modalities, the capture device is already included in mobile devices (e.g., camera for face recognition, touch screen for handwritten signature recognition, microphone for speaker recognition, or the inclusion of some sensors for fingerprint verification). Handwritten signature, voice and face recognition has been suggested as the most suitable modalities [15]. Thus, in accordance with the transition to mobile environments, most of the experiments were carried out in mobile scenarios during this Thesis.

Specifically, this work analyses the H-B interaction usability methodology published by Fernandez-Saavedra and points out its weaknesses and the non-covered points. Afterwards, several experiments were done analysing specifically all those highlighted concerns as validations. These validations allowed us to make improvements to the H-B interaction.

Improving the methodology for measuring usability as well as guidelines for convenient designs, future biometric systems will be not only interoperable but more reliable and usable by a wider percentage of users.

The accessibility (usually included within the usability concept) in biometric recognition systems has been also analysed. This topic, included into the improvements made to the H-B interaction was considered relevant enough as to be presented in an additional section. The methodologies applied to measure the accessibility are the EN 301 549 and the well-known usability measurements (ISO 9241).

This document is divided in five main parts: Introduction, State of the art, Usability evaluation methodology, Accessibility in biometrics and Conclusions and Future work.

- **State of the Art in Biometrics, Usability and Accessibility**. This part of the document includes the main researches in usability and accessibility in the context of mobile biometrics. The main results of the state of the art are analysed and the points to be improved are highlighted.

- **Usability evaluation methodology**. This is the main part of the Thesis. First, the evaluation of biometrics is explained (ISO/IEC 19795) and then the H-B interaction methodology is analysed. Once we have reviewed the starting points and the uncovered weaknesses, modifications to do in the methodology are highlighted. Afterwards, all the experiments where the points to be covered were studied are shown and the results analysed. When all the factors had been analysed, the modifications to the H-B interaction methodology are applied. In this part, the user-centred-design was used and the abovementioned process is repeated after each experiment. As the user – computer interaction is higher in dynamic modalities (users are required to perform more actions), most of the experimental studies relate to dynamic handwritten signature recognition. This modality was chosen due to sign is a well-known procedure and users feel generally comfortable, as also shown during the experiments. Moreover, handwritten signature recognition in mobile devices is a trending topic nowadays.

  Accessibility in Biometrics. Once the necessity of assess accessibility in biometric systems has been highlighted in the usability part, this section replies to that necessity. As accessibility was included within the H-B interaction improvements, we have carried out experiments to validate the new methodology update. Then, it is measured through the well-known usability measurements (ISO 9241) and the EN 301 549 both added to the H-B interaction.
  Finally a use case where blind users interact with a face recognition system is studied. This use case illustrates the difficulties to face when including people with disabilities in biometrics and how a proper design allows them to interact with the technology properly. The accessibility measurements in the blind users case has been done through efficiency (in terms of time) and performance comparatives.

- **Conclusions and Future Work**. This section includes discussions of the main topics addressed during the Thesis and best practices on usability of biometrics. As several outcomes are detailed after each experiment, this part contains only the main conclusions reached during the Thesis. Finally, the possible future lines are detailed and discussed.

Part II: State of the Art

Chapter 2

Biometric Recognition

2.1 Biometric Recognition

According to the Handbook of Biometrics [16], Biometrics is "*the science of establishing the identity of an individual based on the physical, chemical or behavioural attributes of the person*". The possibility of univocally identify people has encouraged the use of biometrics in several contexts, overall traditionally in security systems. Therefore, biometric recognition is the use of biometrics as the univocal differentiation of individuals according to special physical or behavioural attributes such as face, iris, voice or handwritten signature. Biometric recognition systems are common in places where high security is needed and are utilised in many cases as a reinforcement of other security techniques (smart cards, PIN codes, passwords, etc.). Accessing to a secure area (e.g. using iris or finger recognition) or recognizing passengers in an airport (by using face recognition [17] or fingerprint) are two examples of biometric recognition usage.

2.1.1 Highlights of biometric recognition history

Biometric recognition is the differentiation of individuals by special characteristics. Thus, humans have always identified among them by faces, voices or odour for example. There are several evidences in the history of the use of biometrics [16]:

- Fingerprints were used as a person's mark by Babylonians in businesses in 500 B.C
- Chinese used fingerprints to differentiate children.
- Egyptians trusted traders were identified by their physical descriptors.

Some of the most relevant biometric milestones up to now are the following:

- Handwritten signature is considered as the first biometric modality. People not knowing how to write were asked to draw a cross or any other kind of mark.
- By the early 1800s are the first biometric scientific studies made by E. Purkinje on fingerprint recognition, differentiating among the different kinds of reliefs in the fingerprint and the different patterns in accordance.
  - Bertillon was the first of using fingerprints for identification matters, writing the Bertillon's Identification Anthropometrique in 1893.
- Dr. Henry Faulds studied fingerprints and their persistency in the 1870s, building a classification system. His work ended in Sir Francis Galton hands who was one of the most relevant experts in anthropometry.
- In 1888 Sir Francis Galton started his research on fingerprint recognition, which includes the demonstration of the fingerprint uniqueness and persistence in time and the demonstration that fingerprints are internal features (which grow again if the skin is removed). Furthermore, he designed a method for obtaining inked impressions of fingerprints, classified the fingerprint patterns (Loops, Whorls and Arches) and created a 10 fingerprints classification system.
- Scotland Yard started to use the Galton's system in 1894 as accompaniment of the Bertillon's system. Finally they chose only the Galton's method and improved therefore the human classification system based on fingerprint which is the one currently used by most countries in the world.
- With the growth and improvement of computers, biometric recognition systems improved deeply: they started to be automated and new modalities appeared. In 1936, the identification by iris was proposed and in the 1960s face recognition became to be automated. Also in the 1960s speech and handwritten signature recognition were proposed.
- By the late 1990s biometric recognition algorithms have seriously improved, especially with the support of FBI (Federal Bureau of Investigation) and NIST (National Institute of Standards and Technology), who started to get involved in biometrics for forensics (e.g. FBI launched a DNA forensic database in 1998) and organized public competitions an evaluations (e.g. FERET (Face Recognition Technology) [18] in 1993 or IAFIS (Integrated Automated Fingerprint Identification System) in 1994).
- In the 2000s the ISO/IEC standards subcommittee on biometrics is stablished (2002) and ICAO (International Civil Aviation Organization) [19] adopts biometrics to MRTD (Machine Readable Travel Documents) in 2003. The number and quality of researches increased severally and the algorithms are continuously improving.

- Nowadays biometric recognition is used in several contexts, in addition to security and forensics. This technology is moving to ubiquitous environments and being embedded in mobile devices such as smartphones or artefacts for border control purposes, in many cases oriented to personalization and societal good.

2.1.2 Biometric modalities

The biometric recognition modalities are divided in two main categories: physical and behavioural. Physical modalities are those based on human physical characteristics and the best known are fingerprint, iris, face (3D and 2D), palmprint, hand geometry and ear shape recognition. Behavioural modalities are those based on behavioural traits of individuals and usually require a higher extent of user interaction. Examples of behavioural modalities are handwritten recognition signature, gait recognition, keystroke dynamics recognition [20] or voice recognition (which is considered also a physical modality). There are seven factors [16] considered desirable to determine if a trait is adequate or not:

1. Universality. Everyone has to possess the trait.
2. Uniqueness. The trait should be different enough among the population.
3. Permanence. The trait should not change significantly over time.
4. Measurability. The collectability of the trait should be easy to handle (collect, process and store).
5. Performance. The recognition accuracy should meet the desired requirements.
6. Acceptability. People might want to use the trait or be willing to.
7. Circumvention. The trait should not be easy to forge or imitate.

Then, the use of one modality or another depends on the application's needs and characteristics and there is not a perfect biometric modality which accomplishes all the requirements: all of them have pros and cons. In this section the biometric recognition modalities especially relevant in this work are shown more in depth.

2.1.2.1 Handwritten signature recognition

Traditionally, handwritten signature recognition has been one of the most widespread accepted methods by human being to authenticate themselves and acknowledge the understanding and acceptance of a written text. Therefore the signing process is already familiar for individuals since it is a common procedure in multiple scenarios: administration requirements, delivering services, rental agreements, contracts and so on. Also, with the progress of technology, there are multiple acquisition alternatives for capturing the signature performed, either by using signature recognition in a paper, or by the use of electronic devices. The massive growth of electronic devices allows people to sign even with the fingertip. Figure 1 shows an example of signing with a stylus and signing with the fingertip.

Figure 1. Example of a signature made with the fingertip (left) and with a stylus (right)

As a behavioural modality, handwritten signature recognition has some drawbacks, particularly because of the fact that behaviour depends on a large number of factors such as mood or aging. Therefore it is claimed that even the best algorithms known nowadays do not meet the accuracy percentages of some other modalities such as fingerprint or iris recognition. Several works in signature recognition have been developed in western and in far-east countries, providing information about the diversity of the signing process in different cultures. The increasing number of works in this field has led to significant improvements in reducing error rates, and they have even created new modalities. Recent works like [21] place writing recognition as a field in ongoing research through alphabetical characters and numbers [22].

In the handwritten signature recognition modality, several methods are researched in order to optimise results, but first it is necessary to divide them into two main approaches: static [23] and dynamic [24]. Static methods take on the image of the signature as the source of information for the recognition process, so no other extra data is provided apart from the picture itself. On the other hand, dynamic methods use multiple data channels as the input into to the recognition algorithm. Examples of such data channels are the spatial and temporal variation of the signature, velocity in both axes, time spent when signing, pressure, pen angles and so on. The variability of the channels used and the application of different algorithms involve the big amount of ongoing research works nowadays.

Historically, the most well-known and used is the static approach, typically used for forensic studies such as detecting the signer in a document or bank check. With the improvement of computers, these forensic studies were taken into algorithms as to reach the same level of performance than that of a calligrapher, but this is still quite far to be reached. Main problems with static approaches (also called off-line signature recognition), come from the fact that some personal characteristics of the signing process, such as variations in the pressure and pen inclination are not easily detected by the use of scanning the image of the signature. Also, calligraphers make an exhaustive use of their experience, applying variable heuristic methods whenever they consider that a more appropriate approach is needed. This kind of knowledge has not been translated to computer algorithms yet. Therefore static approaches are less resistive to be forged than dynamic ones, as imitation of velocity or pressure variation needs a really high forgery work [25]. As technology improved, dynamic handwritten signature recognition (also known as on-line signature recognition) gained popularity and capture devices that acquire signals while signing have become common. Data acquisition process is more complex because of the necessity of obtaining various data channels. The number of features that can be obtained is high as it can be seen in [26] and the most common are x and y coordinates, time, azimuth and pressure. There are not a lot of databases composed by handwritten signatures (and even less including the dynamic features). The most well-known are the MCYT (Ministerio de Ciencia y Tecnología, Spanish Ministry of Science and Technology) [27] and the SVC2004 [28]. During this Thesis, dynamic signature verification is also called DSV.

2.1.2.2 Fingerprint recognition

This modality is the most widespread and best known. Fingerprints have been used for personal identification for many decades, mostly because the identification accuracy using fingerprints has been shown to be very high [29]. A fingerprint is the pattern of ridges and valleys on the surface of a fingertip whose formation is determined during the first seven months of fetal development. It has been empirically determined that the fingerprints of identical twins are different [30] and so are the prints on each finger of the same person.

Fingerprint recognition is today used in several national ID documents, automatic border controls, forensics, access controls such as companies' accesses, home automation systems or even for mobile devices protection. Users find the use of fingerprint sensors not too much intrusive [31]. Furthermore, the error rates in fingerprint recognition are much lower than the rest of modalities, making this modality suitable for security environments. One of the main drawbacks in fingerprint recognition is the resistance to forger. There are studies which demonstrate the feasibility of reproducing a fingerprint easily [32]. On the other hand, there are also works in fingerprint recognition anti-spoofing which check for instance the subject's liveness [33].

There are two main types of fingerprint images: offline (where the fingerprint is printed over a paper with ink by subject's fingertip pressure and finally digitized by scanning) and live-scan (which obtains the fingerprint from a sensor capable of directly digitize the fingerprint by contact). There are several kind of live-scan fingerprint sensors such as optical, solid-state or ultrasound sensors. On the other hand, there are also swipe sensors which obtain the fingerprint image through the subject's swipe on a sensor.

As fingerprint recognition is not a new technology, several public and private databases are available. The most relevant are the IAFIS and NIST databases, gathered through the FVC (Fingerprint Verification Competition) competitions [34].

2.1.2.3 Face recognition

Biometric recognition based on face features is one of the most demanding methods for guaranteeing security in many scenarios, especially in crowded or open spaces (e.g. public events or airports). Besides those scenarios, face recognition is also used in automatic border controls [35], in breeder documents, in access control systems, etc. Nowadays it is also used in mobile devices [36]. The face recognition is mainly based on several particular points so called landmarks which can be represented in 2D or in 3D according to the capture sensor. Also face recognition with 3D images could be considered another different modality and may overcome some of the 2D main drawbacks: the adverse environmental conditions (e.g. incorrect lightning) and subject's changes due to the ageing effect (e.g. wrinkles or injuries) or changes in the appearance (e.g. beard, make up or eyeglasses). These drawbacks usually involve low performance and therefore face recognition is not as widely deployed as it is supposed to be. Moreover, the resistance to attacks is not still as satisfactory as desired, involving security problems: good enough masks could easily fake a current face recognition system [37]. On the other hand, face recognition is one of the less intrusive modalities because subjects are not intended to perform specific or complicated movements.

Some of the most representative database examples are the FRVT (Face Recognition Vendor Test) [38], the FRGC (Face Recognition Grand Challenge) [39] and the FERET (Facial Recognition Technology). Furthermore, the number of multimodal databases has increased in recent years, e.g. BIOSECURE [40] or the MOBIO (MObile BIOmetrics) project .

2.1.3 General biometric model

The biometric recognition general model is divided in various phases. According to the application, the system should complete all of them or a subset. Under a very generic perspective, there are two main steps: enrolment and recognition. In the enrolment the system stores the subject's traits and during the recognition the system confirm if the subject is who he claims to be (verification) or checks if the subject is present in a database (identification). From a more specific point of view, the biometric recognition process consists of the following processes (Figure 2):

- *Biometric Presentation*

The subject presents the biometric trait to the system. During this process factors such as the environment and the HCI (Human-Computer-Interaction) are critical and have an important impact on the quality of the sample acquired. This process also involves that even some static modalities (e.g. fingerprint or iris recognition) are in one way or another dependant on the subject's behaviour, because subjects are most of the times required to make an effort to present the trait to the system.

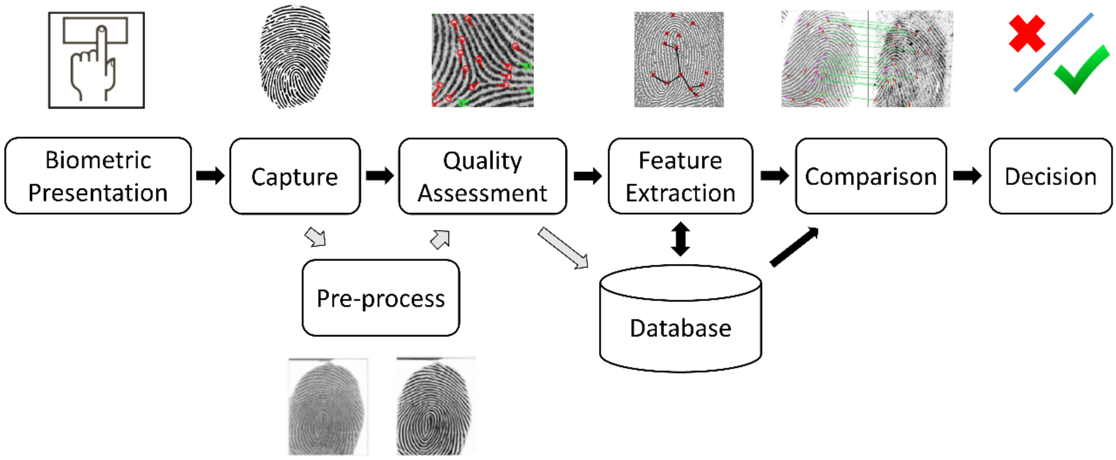

Figure 2. General phases of the biometric recognition process

- *Capture*

The capture is made by the system which is able to detect the biometric trait. Actually, in some contexts the system does not detect the trait and only capture information manually or automatically (e.g. pictures taken by a camera or sound recorded with an audio recorder). There are also specific sensors which detect the trait and capture it, such as in several fingerprint sensors.

- *Pre-process*

The pre-processing improves the image quality and/or adapts the sample to the algorithm inputs requirements. This step in not always mandatory but recommendable for higher performance.

- *Quality assessment*

Once the biometric trait is captured the sample may be good enough for processing or may not. In order to check the sample quality, some systems have quality assessments (this is not always present). Usually, if the sample is below a given threshold the subject is required to present the biometric trait again.

- *Feature Extraction*

In this phase the system extracts the trait characteristics (better known as features vector). This process is necessary for matching samples (actually matching features) and it reduces the required space in the database if the system only stores the features vectors.

- *Comparison*

The comparison between two samples returns a result indicating the likelihood. The higher the matching result the more probable that both samples belong to the same subject.

- *Decision*

According to the comparison result the system decides if the user is genuine or impostor and this is made according to a prefixed threshold. Depending on the application requirements the threshold may be different (e.g. for high security scenarios may be different that for home automation systems).

2.1.4 Biometric system performance

Systems based on biometric recognition are not perfect and do not return unequivocal results. This can happen due to several factors such as ageing, trait erosion or misuse of systems. It means that a subject is correctly identified as genuine or as impostor with a certain probability which is always below 100%. That is why it is necessary to fix a threshold for accepting or rejecting samples. From the degree of similarity between samples obtained through the matching function, the system performance can be extracted. There are several kind of rates to measure a biometric system performance as well as several kind of curves. Moreover, when showing biometric systems performance, it is also necessary to provide data regarding algorithms used, databases, number of genuine and impostor samples used, when the comparisons took place, etc.

2.1.4.1 Error rates

The most common error rates are the FAR (False Accept Rate), the FRR (False Reject Rate), the FMR (False Match Rate), the FNMR (False Non-Match Rate) and the EER (Equal Error Rate). Other common error rates which affect performance are the FTA (Failure To Acquire) and the FTE (Failure To Enrol).

The FNMR is the fraction of genuine scores classified incorrectly as impostors. It is also called Type I error and corresponds to the incorrect rejection of a true null hypothesis. Having a high FNMR involves that a genuine user will be rejected several times. If the FTA (fraction of bad sample acquisitions) is taken into account, then the measurement is called FRR. The relation is:

$$FRR = (1 - FTA)\ FNMR + FTA$$

The FMR is the fraction of impostor scores classified incorrectly as genuine. It is also called Type II error and corresponds to the failure to reject a false null hypothesis. This is the less desired case because having a high FMR involves that a system is easy to forge. If the FTA (fraction of bad sample acquisitions) is taken into account, then the measurement is called FAR. The relation is:

$$FAR = (1 - FTA)\ FMR$$

When there is a fail during the enrolment, it is called failure to enrol (FTE), but during the enrolment FTA may also occur.

As biometric recognition is used in many different contexts, it is interesting to know the system functioning for different thresholds. For instance, the threshold for accessing a bank account should be more restrictive than a threshold for unlock a smartphone. Therefore a low FAR is required in the first case and a low FRR is required in the second. Both error rates are complemented; that is, when the FRR increases in a system, the FAR decreases and vice versa. It is also common to express a system's performance through the EER, which is the threshold in which FAR and FRR have the same value.

2.1.4.2 Performance curves

It is common to represent the system error rates through different threshold values conforming curves which are useful to straightforward understand the system functioning. The most representative curves are the DET (Detection Error Trade-off) and the ROC (Receiver Operating Characteristic).

The DET curve plots directly the FAR against the FRR at various thresholds on a normal deviate scale and interpolates between those points. When a linear, logarithmic or semi-logarithmic scale is used to plot these error rates, then the curve is called ROC [16].

2.1.5 Uses of Biometric recognition: Mobile Devices

Mobile devices are playing a significant role in daily life, not only for communications but also for entertainment, working activities or social relationships. Along with the high increase of the use of smartphones and mobile devices in daily life, the amount of sensitive data that these devices store is also increasing (e.g. bank accounts, personal e-mails, photographs, etc.). This situation leads to the need of protecting the access to such sensitive data, and biometrics is offered as an alternative mechanism for such protection [41]. According to the latest improvements in smartphones, the range of possibilities for integrating biometrics is promising, with potential applications such as signing documents univocally, secured access to websites, execution of administration procedures, etc. Furthermore, the use of other traditional authentication schemas based on passwords, is considered by users as cumbersome due to the necessity to remember a large variety of alphanumeric codes, which usually drives users to re-use the same password for several, if not all, services accessed. The use of biometrics allows the user authentication through "something she/he is" or "something she/he does" avoiding the use of "something she/he knows". Therefore, the possibility to use biometrics to manage and protect sensitive data arouses the interest of users and researchers, furthermore when it is considered a protection mechanism easy to use and secure.

It is important to note some other facts that encourage the inclusion of biometrics into mobile devices. The first one is the large amount of devices already deployed, which has reached the situation that it is difficult to find someone that does not possess and use daily devices such as smartphones or tablets. The second one is that for some biometric modalities, the capture device is already included within the mobile device (e.g. camera for face recognition, touch screen for handwritten signature recognition, microphone for speaker recognition, or the inclusion of some swipe sensors for fingerprint verification). This leads to an important reduction in the cost of the deployment, as users already have those devices and they should only acquire the application. Other important factors are the necessity of having ID portable devices by security forces (e.g. for suspects identification), or for signing documents on the spot.

Also, as users are already familiar with this kind of devices, the usability level achieved could be improved, although, as mentioned below, mobility also creates new usability challenges. Due to marketing needs, mobile devices are improving every day, which will allow powerful biometric algorithms in the near future.

As an important drawback, mobile devices present security concerns related to how the operating system controls the way that installed applications access memory data and communication buffers. A lack of a strict control compromises the integration of biometrics as sensitive data may be endangered.

Usability and accessibility problems can appear also. For example, the use of inadequate interfaces or the adaptation to different user disabilities has to be addressed from scratch in order to offer universality: if the technology is not easy to use or hard to understand, users will reject its use. Moreover, not all the biometric modalities fit perfectly in mobile environments or the migration is far from being easy or cheap. For instance, nowadays the smartphone screen does not allow capturing the fingerprint and then an extra sensor is needed. Therefore many research fields are open.

2.1.5.1 Background

The idea of integrating biometrics into mobile devices started several years ago, while biometric technology started to acquire an acceptable level of maturity and societal acceptance. For instance, in 2005 there is an example of biometric recognition in a rudimentary mobile phone with iris [42]. In other approaches, various biometric modalities are implemented in mobile devices: palmprint, knuckle [43] or fingerprint. Nowadays biometrics is proposed as one of the best solutions to guarantee security within mobile environments. Good examples of it are studies in secure mobile voting [44], and mobile banking.

As biometrics was proposed to be used for multiple purposes (e.g. e-commerce, e-government, etc.), some manufacturers started to create prototypes with fingerprint sensors, as a mean to unlock the device in a comfortable way. This situation also led to the approval of several R&D projects.

The SecurePhone European project (IST-2002-506883 active in 2004-2006) was focused on this topic. Its main target was to develop a biometric recognition system for mobile phones based on face, voice and handwritten signature [45]. This project integrated a biometric recognizer in a 3G-enabled PDA, which allowed users to mutually recognize each other and securely authenticate messages (text or audio). This enabled users to sign legally binding contracts on their PDA/mobile phone.

Two years later, in 2008, the European project Mobio (Mobile Biometry, FP7-214324 http://www.mobioproject.org/) started. In this project the incorporation of biometrics (voice and face) on a mobile device was proposed. Furthermore, it was focused on specific aspects such as the remote biometric identification or the computational limitations of the microprocessor. The intention of the project was to develop new mobile services secured by biometric authentication mechanisms. Its objectives included robust-to-illumination face authentication, robust-to-noise speaker authentication, joint bi-modal authentication, biometric reference adaptation and system scalability.

Regarding the integration of biometrics in mobile devices the future is promising: many companies are betting big and the technology seems to be accepted by users. According to Goode Intelligence: "*The mobile biometric security market would grow to 39 million users by 2015*" [46]. Attending to this forecast and the quantity of improvement possibilities that experts can overcome, the big amount of research works that exist in this way is justified.

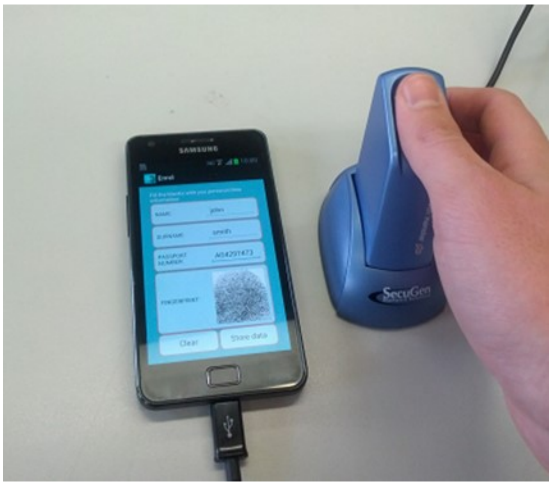

Figure 3. Biometric fingerprint obtained with a fingerprint sensor connected to a smartphone

The purchase of the company PittPatt by Google in 2011 and the posterior adaptation to Android is a clear example of the advances in this field, suggesting the facial recognition as a comfortable method for unlocking the smartphone. More recently, Apple bought the company AuthenTec in 2012, showing their clear inclination for the fingerprint authentication. Furthermore they granted a patent for a two-step unlock screen feature that has yet to be implemented.

There are several biometric modalities that fit well in mobile environments, like face recognition, ear shape recognition [47] or handwritten signature recognition [48]. At the same time, along with the migration to mobile environments, new modalities emerged, such as recognition by the touch screen input.

Probably the most attractive modality to be applied is fingerprint recognition, but as the integration of a fingerprint sensor requires industrial product development, some studies have also analysed the possibility of acquiring fingerprints with the mobile camera. That work is in addition to the obvious use of using the camera for face recognition [49]. Also, the use of the accelerometer represents a good chance to implement behavioural biometrics too [50]. Furthermore, biometric recognition is being used in conjunction with some other communication protocols in smartphones such as NFC [51]. In Figure 3 there is an example of fingerprint recognition in mobile devices using an external device.

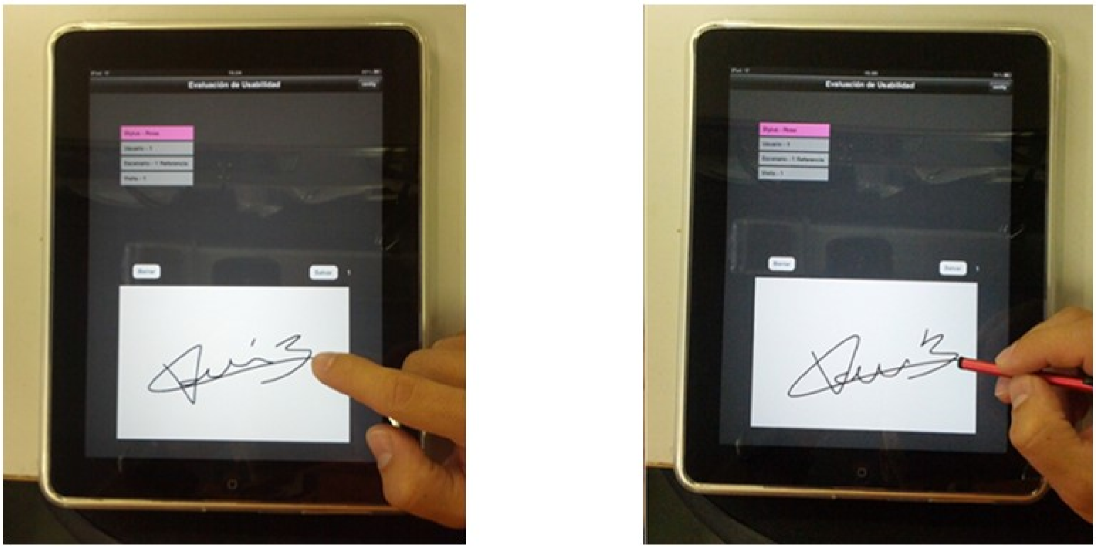

Figure 4. Handwritten signature recognition on and iPad, signing with the finger (left) and a stylus (right)

But smartphones are not the only kind of mobile devices that can be used. Mobile devices, when being integrated with biometrics, can be divided in several categories, although these three are the main, according to market trends and popularity:

- PDAs (Personal Digital Assistants) are handheld devices that combine elements of computing, telephone, Internet and networking. Typically, they are designed to be managed with a stylus, which fits perfectly for handwritten signature recognition.
- Smartphones include phone and computer functions in addition to different sensors allowing capture image, sound or positioning. This makes the biometric integration easier.
- Tablet and Tablet PCs are a type of notebook computers including sometimes phone functions. The big difference with the smartphones is the screen size that is bigger, providing users with more space to interact with the device. In Figure 4 is an example of handwritten signature recognition with an iPad signing with the fingertip and with a stylus.

- Portable devices proposed by the industry, including handheld terminals specially made for biometric recognition. These devices support common modalities such as fingerprint, face or iris recognition.

2.1.5.2 Migration requirements

The extraordinary advances that portable devices have experienced have converted them in small computers whose host processors are capable of work in the GHz range, with memory capacities larger than 256 MB and a variety of sensors (touch screen, camera, microphone, accelerometer, etc.). These sensors are suitable to capture many of the users' biometric features, but its performance or acquisition variables differ with respect to those obtained with conventional capture devices. For example, cameras included in smartphones work within the visible spectrum, unlike commercial iris capture cameras that operate in the infrared band. On the other hand, capacitive touch screens incorporated in most of the smartphones nowadays do not allow extracting the pressure exerted on the surface. This parameter is used in most of the handwritten signature recognition algorithms. Though, in order to offer reliability, migrating biometrics to mobile environments requires several modifications to address all these new constraints. The transition from PC to mobile devices is not direct or easy to deploy and it brings new challenges that have to be covered. These new challenges are:

- Adapting the device to acquire the selected biometric modality.
- Defining the application architecture to better fit the scenario (e.g. local authentication vs. remote authentication).
- Fine tuning the biometric algorithm for improving performance considering the execution platform and the acquisition properties.
- Evaluating the impact on performance of the implementation in a variety of platforms.
- Evaluating the impact on performance of the different ways and situations of using mobile devices, including not only the positioning of the user, but also the environmental conditions surrounding (e.g. light, humidity, noise, vibration, etc.).
- Evaluating usability in new applications and re-designing according to the results.

Furthermore, accessibility problems using biometrics in mobile devices can arise to some groups of users. On one hand, users with the so called fat fingers problem are sometimes not able to point accurately in a smartphone touch screen. Also users with disabilities, such as arthritis or Parkinsons' disease can find uncomfortable or impossible to handle a mobile device. Elderly would find some complications also at the time to complete repetitive procedures or understand some steps. Additionally, not all the biometric recognition modalities fit perfectly in mobile environments or work in all the scenarios. For instance, to use hand geometry recognition the only approach of using a peg-free image with variable background is viable [52] as shown in Figure 5. Also gait recognition could not work properly if the device is not in a specific place (most probable in the same pocket, and sometimes even on the same trousers). Other example is voice recognition, which would have problems in spaces where the ambient noise is loud.

Figure 5. Hand geometric recognition using a peg-free image

2.1.5.3 Inconveniences

Currently, smartphones can also access data services offered by mobile phone companies, a market that is rapidly expanding due to the growing demand from users. Through the data service a user can connect to Internet and look up information of all kinds (e.g. email, banking operations, etc.). The data security handled by applications is trusted to the operating system (OS) of the portable device.

For instance, in the Android OS, the user is responsible for authorizing the permissions (access to sensors, data or other applications activities) requested by each application during installation, so that they run in a “Sandbox” that keeps the data (contained in the memory and files) isolated from other applications.

The flexibility that these smartphones offer to users for modifying and/or updating the OS, or the installation of new applications and authorization permissions, allow access to sensitive data via several attacks like rootkits, privilege escalation, appPhishing or appJacking. As a consequence, sensitive data such as biometric references or private keys in a PKI solution are not totally safe within the mobile device, which is a major drawback in this kind of devices.

Another concern that has to be considered is the limitation of the development platforms when accessing biometric information in client/server architectures, such as through web services. Some technologies, such as Java Applets, ActiveX Controls, JavaScript or Flash that are essential to capture biometric data in PC platforms, are not fully available in some web browsers installed in mobile devices.

One of the possible solutions to these mobile devices vulnerabilities is the use of other technologies jointly with biometrics. One example is the use of smart cards and biometrics altogether. Unlike mobile devices, the smart card eliminates any possibility of reading and/or modification of the biometric traits as a consequence of the SO vulnerabilities, although does not have the computational power or sensors of the mobile device. Therefore, it can be said that, while a correctly implemented smart card is considered a tamper-proof system, a mobile device is far from achieving that status.

In other words, the need of improving security in mobile platforms is pushing forward the development of TEEs (Trusted Execution Environments), either by implementing them in software, or by using TPM (Trusted Platform Modules) that include SEs (Secure Elements), such as smartcards or other kind of hardware module that help in the accomplishment of the security requirements.

2.1.5.4 Next steps in mobile biometrics

The company Validity, which develops solutions for mobile biometrics, has recently designed a fingerprint sensor for being embedded underneath the smartphone screen. The fingerprint capture is made with a finger swipe, so this avoids having to reserve space in the smartphone housing for another sensor (and simplifies the process: unlock and fingerprint recognition can be made at the same time).

Another step forward in the biometrics mobile field is the agreement between the USA Defense Department and the company AOptix to develop a hardware peripheral and a software suite that turns a commercially available smartphone into a device that scans and transmits data from users' eyes, face, thumbs and voice. The intention is to have the possibility (soldiers, marine or special operators) to record the biometric information of suspicious people on the spot.

In addition, biometrics is being integrated in mobile devices in order to facilitate everyday tasks (automation, bank transactions, etc.) for people with disabilities.

One of the trending topics in mobile biometrics is the introduction of NFC which permits fast data wireless interchange saving time. Moreover, most of the new mobile devices being manufactured are equipped with this technology, which gained its popularity in Japan, as it started to being used for daily payments.

Chapter 3

Usability and Accessibility

3.1 Usability

This section contains the state of the art of usability from the general term to the applications in biometric recognition. The term usability has several definitions and from a very informal point of view it could be defined as "*how usable is something*". The definition provided by the ISO 9241-11:1998 [53] is: "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*". Also, this standard defines the effectiveness, the efficiency and the satisfaction:

- **Effectiveness**: "*The accuracy and completeness with which users achieve specified goals*".
- **Efficiency**: "*The resources expended in relation to the accuracy and completeness with which users achieve goals*".
- **Satisfaction**: "*Freedom from discomfort, and positive attitude to the use of the product*".

The ISO 9241-210:2010 [5] standard provides guidelines for designing usability following the UCD and the usability definition provided by the ISO 9241-11:1998.

3.1.1 User Acceptance

This concept, which is not part of the ISO 9241-11 definition, is becoming to be trendy in biometric recognition in the sense of the user willingness to use biometrics. It is important to differentiate User Acceptance from UAT (User Acceptance Test) which is a test conducted to determine if the requirements of a specification or contract are met.

Despite of factors such as performance or ergonomics are crucial in biometric recognition systems, user acceptance should be considered also as a key element: even the best device in performance will be rejected by users if they feel uncomfortable using it. Therefore, testing the user acceptance is indispensable when designing biometric systems but it is not carried out in all cases involving many times the disuse of the technology. Some of the main concerns expressed by users to not use biometric recognition are suspicion, invasion of privacy, fear of damage or linking biometrics with personal data, apprehension of "*it won't work for me*'' and many others [54].

3.1.2 User Centred Design (UCD)

The User Centred Design (UCD from now on) has been defined by the ISO 9241-210 as "*an approach to interactive system development that focuses specifically on making systems usable*''. The UCD defines an iterative design and development lifecycle for usable products but does not specify methods. Once the lifecycle outcome is the desired and the design is success, the iteration ends. The general lifecycle divided in phases is depicted in Figure 6. These phases are the following:

- **Context of use**. Identify the users who will use the product, the possible uses and conditions.

- **Requirements**. Identify all the requirements and user's goals to be met to design the desired product.

- **Design**. According to the previously specified requirements, a proper design must follow several steps from the initial concept to the final design.

- **Evaluation**. The design assessment is crucial to identify weaknesses and correct them in next cycles. The closer to real product use-cases the more useful is the evaluation.

3.1.3 Human Computer Interaction

HCI is a topic directly related to usability and it is defined as "*a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them*'' by ACM SIGCHI [55] members. The final goal of the HCI is to make the interaction with computers as comfortable and smooth as possible. The literature in HCI is extensive as well as the conferences and congresses, such as the ACM SIGCHI, being the Norman-Nielsen [56] one of the best-known groups involved. They have published several relevant works regarding the user interface design, including the 10 usability heuristics [57] and several UX (User eXperience) reports.

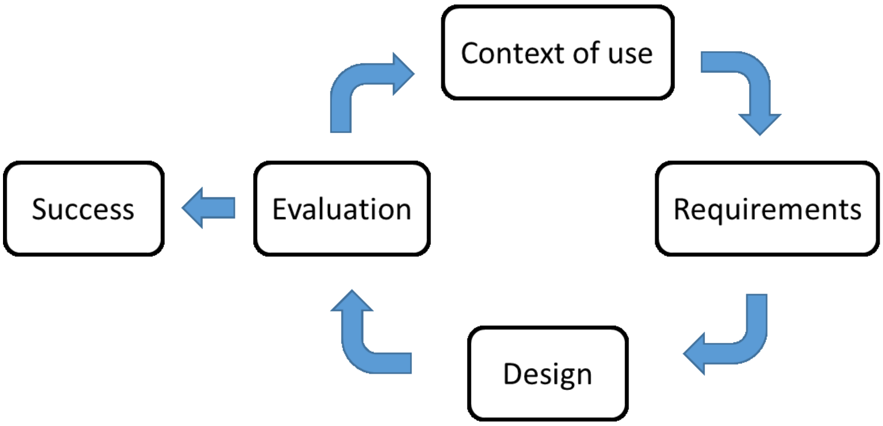

Figure 6. UCD general lifecycle

3.1.4 Usability in Biometrics

Most of the works dealing with HCI are in the line of improving user interfaces which is undoubtedly a relevant issue. Nevertheless, other important topics within HCI, also relevant in biometrics, have not been deeply studied yet by the HCI world. Then, some years ago, several research groups started to study the usability issues when interacting with biometrics, including not only interfaces but ergonomics or user's moods. The most relevant studies regarding usability in biometrics up to now are the following:

- ***NIST works***. NIST carried out a series of researches following the ISO 9241-11 definition of usability (all of them publicly available [58]). Two of the better known are "*Effects of Scanner Height on Fingerprint Capture*" (Figure 7) [59] and "*Usability Testing of Face Image Capture for US Ports of Entry*'' [60]. In order to stablish proper guidelines and help future developers, they have also published a handbook in 2008 called "*Usability & Biometrics Ensuring Successful Biometric Systems*" [61], which has been the starting point for several researches. In the NIST handbook, authors explain how to apply the UCD to biometrics and suggest different methods for evaluating usability in biometrics. Unfortunately, there are only best practices and advices in the handbook but not any proper methodology for evaluating usability.

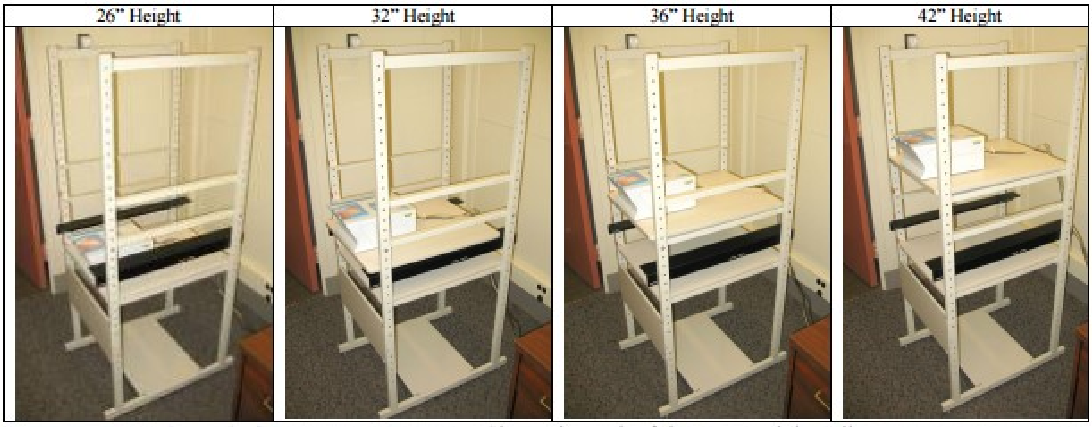

Figure 7. Different heights for placing the fingerprint scanner [59]

• ***The UK Trial***. This research published in 2005 by Atos Origin was made with the goal of "*test the processes and record customer experience and attitude during the recording and verification of facial, iris and fingerprint biometrics, rather than test or develop the biometric technology itself*" [62]. Therefore, it is one of the first works studying specifically usability in biometrics. They measured for instance processes timing, user perceptions and reactions including also people who experienced difficulties when performing the different phases. Nevertheless they only return the results but not explained the reasons of the findings nor provide suggestions for fixing the inconveniences found.

• ***The Human Biometric System Interaction model***. The HBSI model [4] is based on three kind of interactions, showed in Figure 8, namely human, sensor and biometric system. Therefore, this model is not only focused on the ISO 9241-11 definition of usability but also includes measurements to complete a usability analysis covering ergonomics and signal processing, as shown in Figure 9.

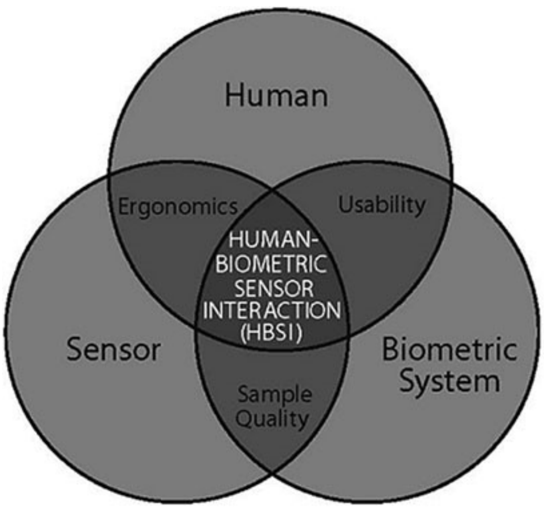

Figure 8. HBSI Model

It also includes a set of metrics to classify the presentation the user makes to the system, including the possible mistakes (both users and system errors). The current HBSI metrics are the following:

Defective Interaction (DI). The user makes an incorrect presentation that is not detected by the biometric system.

Concealed Interaction (CI). An incorrect presentation is detected by the biometric system but is not classified correctly as an error.

False Interaction (FI). An incorrect presentation that is detected by the biometric system but, is correctly handled as an error.

Failure To Detect (FTD). It is a correct presentation made by the user that is not detected by the biometric system.

Successfully Processed Sample (SPS). It is a correct presentation that is detected by the biometric system and successfully processed as a biometric sample.

The HBSI was applied to various usability evaluations including most of the best known biometric modalities such as fingerprints [63] or hand geometry [64]. Nevertheless, the approaches to dynamic modalities were only theoretical [65]. This model has been the starting point for several researches, but again it does not provide a specific methodology for a proper usability evaluation of biometrics.

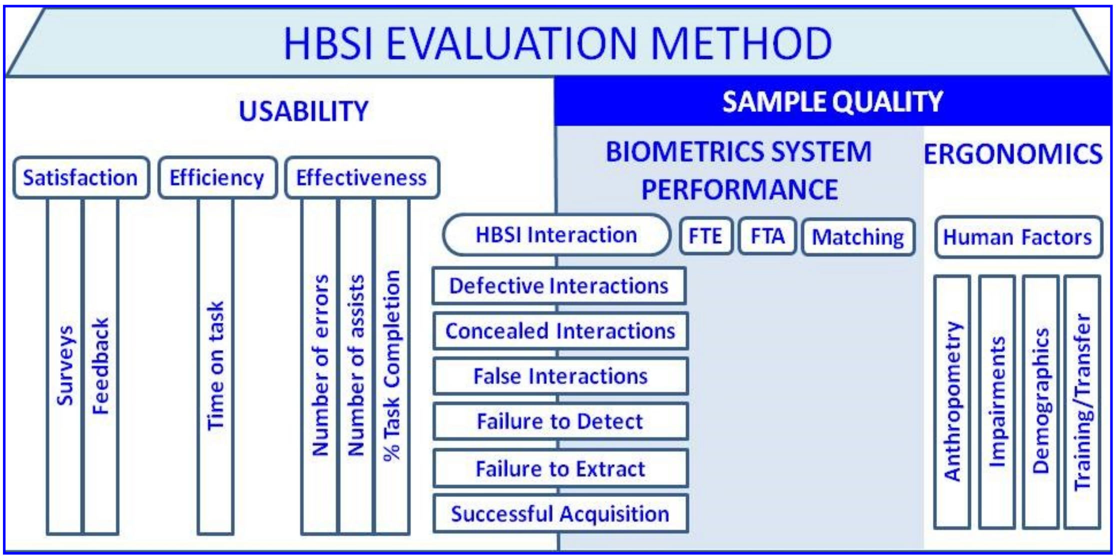

Figure 9. HBSI Evaluation Schema [4]

- ***The Evaluation methodology for H-B system interaction testing of biometric systems***.

This work developed by Fernandez-Saavedra took the HBSI as starting point to suggest a whole evaluation methodology of usability in biometrics. It includes the interaction factors that should be analysed, the proper test procedures and the most relevant metrics and measurements to quantify biometric performance variations. As the HBSI model, this methodology is focused on the usability factors and errors which could influence the system performance. Nevertheless, this work is based on the existing ISO/IEC 19795 multipart standard, focusing completely on calculating the accuracy and speed of recognition algorithms when one or more of the following circumstances occur:

1. Certain characteristics related to the biometric capture device have been modified,
2. Human beings or their biometric characteristic have certain attributes, or
3. Other factors related to the H-B interaction process itself have been modified.

This Thesis is a continuation of the H-B system interaction, which is deeply analysed in Chapter 4.

3.2 Accessibility

The term "accessibility" comes from the word "access", the action of arriving and approaching or entrance. Applying this term to the use of the space and technologies, and specially related to certain collectives with functional difficulties, this word acquires another meaning. This meaning regards to the benefits from people interacting with the environment or with other people.

Accessibility was defined at European level as "*a basic feature of the built environment. It is the way in which houses, shops, theatres, parks and places of work can be reached and used. Accessibility enables people to participate in the social and economic activities for which the built environment is intended.*" [66]

3.2.1 Accessibility of Information and Communication Technologies (ICT)

Though the term accessibility can be included into the field of usability, it is defined more accurately by the ISO 26800:2011 [67] as "*the extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of characteristics and capabilities, to achieve a specified goal in a specified context of use*''. Guarantee accessibility is relevant for ICT for several reasons. First, there is a significant percentage of population who have difficulties to interact with technology: "*about 15% of the world's population lives with some form of disability, of whom 2-4% experience significant difficulties in functioning*'' [11]. Furthermore, we are all potential users with accessibility concerns due to the physical and mental disabilities linked to the process of ageing. According to the state of the art on ICT products, there are several which provide features designed for those users who find the regular products cumbersome or impossible to use. For instance, several brand new smartphones incorporate accessible features (e.g. Talkback in Android and VoiceOver in iOS help visually impaired users to know what is on the smartphone screen) which allow people to use the technology as fluently as non-impaired users.

In the context of the IT the term accessibility is usually referred to web accessibility. In this area the W3C (World Wide Web Consortium) [68] was created to provide recommendations of accessibility to the World Wide Web (WWW). Specifically, the WAI (Web Accessibility Initiative), which belongs to the W3C, has launched a series of accessibility guidelines explaining how to make web contents accessible for web designers. Those guidelines are divided in three blocks:

- Authoring Tool Accessibility Guidelines (ATAG), directed to webmasters and indicate how to provide accessible web contents.

- Web Content Accessibility Guidelines (WCAG), directed to the developers of the software used by the webmasters in order to make those programs easy to create accessible web contents

- User Agent Accessibility Guidelines (UAAG), directed to user agents (e.g. browsers) to make that those programs facilitate all the users the web access.

There are several tools to check the accessibility of a webpage, like for instance TAW, HERA or many others developed by W3C.

3.2.2 The EN 301 549

This standard is the result of a European policy action (Mandate 376) whose goal was to have accessibility requirements for public procurement of ICT products and Services in Europe. EN 301 549 was published in February 2014 [69] and contains requirements and recommendations that apply to any type of ICT products and services.

The structure of the EN is based on the features provided by the ICT, instead of being based in product categories. One relevant aspect of the EN is that the requirements for web content, non-web documents and software are based on WCAG 2.0 [70], by applying guidance provided by the World Wide Web Consortium about how to interpret WCAG for non-web ICT.

Even though the EN 301 549 is considered a proper accessibility evaluation tool for ICT products, it has not been yet tested in a biometric recognition system. Therefore, this standard has been the starting point for the accessibility evaluation methodology presented in this Thesis.

3.2.3 Accessibility in biometrics

In the biometrics world, the design of strongly secure systems is the main concern, and other issues such as usability and accessibility are considered less relevant. Nevertheless, biometric recognition is being lately applied to smart environments where the goal is not to guarantee security at 100% but to ease some common tasks which are sometimes repetitive or annoying (e.g. unlock the mobile device). Moreover, when users gain skills using biometrics the error rates decrease: a proper use motivates better quality of samples.

Applying biometrics to ICT products may ease their use and therefore increase their accessibility. Furthermore, it is important to highlight that every individual is potentially dependent (illnesses, age, pregnancy, etc.). Improving biometrics designs would be beneficial not only for disabled people but for many others who find other technologies complicated to use. It could be thought that biometric recognition is challenging for disabled people but a correct design may ease the process for everyone.

Up to now the works in biometric accessibility are only a few and the literature in this topic is scarce [71] [72]. Regarding the standards, the ISO/IEC TR 29194:2014 *Biometrics - Guide on the inclusive design and operation of biometric systems,* describe some of the main accessibility concerns in biometrics [73]. These were summarized in [51] divided in accessibility concerns:

Eye/Sight related

The most relevant cases are:

- Aniridia (i.e. partial or full absence of iris tissue). It affect both eyes and it is genetic. It may deny the use of iris recognition and provide challenge to face location algorithms.
- Analogous effects may be present when the subject has been gone through treatments such as iridoctomy, for relieving pressure in eye chambers.
- Difficulty for users to see the sensor feedback in order to obtain an accurate alignment with the sensor for capturing biometric data, or to see or understand the structure and interaction of the interface (e.g. where the PIN-pad is located and which is the orientation of the numbers).
- People with a continuous rapid movement of the eyes may also rise challenges to biometric detectors. If such a movement affect the reaction time of the user, then timing on the interaction has also to be considered.

Auditory challenges

For those users that are deaf or have limited hearing capabilities, the challenges are:

- Acoustic feedback (e.g. beeps requiring a user action, such as place a fingertip on a sensor) may become useless, and a visual feedback shall always be available.

- Difficulties in hearing may also impact speech properties, and therefore speaker recognition systems may not be used.

Physical or mobility related

Some of the most relevant challenges are:

- The lack of body mobility may impact the possibility of interacting with the system (e.g. moving the hand to the height of the PIN-pad or a biometric sensor). In some cases this may not seem too serious as the user may be able to interact, but such interaction being not conformant with sensor requirements or process timing.
- Illnesses such as eczema, may prevent for certain biometric modalities to be used, as the quality of the sample acquired could be below the desired threshold.
- Any kind of surgery process that changes significantly the face structure may require the user to go through enrolment again.
- Lack of body parts may prevent the use of certain authentication methods. This is particularly critical when hands or arms are not available for the citizen.
- The use of mobility aid devices, such as crutches or wheelchairs, may impede the correct access to the user to the system, or a proper alignment with it.
- Certain illnesses, such as multiple sclerosis, muscular dystrophy, or any other kind of illnesses that affect motor neurons, may prevent citizens to have full control of their movements, and therefore not being able to interact correctly with the system, or requiring extended time for the interaction.

Cognitive difficulties

When the disability of the person is related to a limited cognitive capability, some of the main challenges are:

- Dyslexia. Difficulties in learning may prevent citizens to use the authentication system in a constant and correct way. It may also raise issues when requiring repetitive processes.
- Degenerative diseases, such as Alzheimer or Parkinson, can impede the understanding of the guidance given, or even the correct interaction with the system.

Other disabilities

But not only are the commonly understood disabilities present. There are other challenges that impact accessibility such as those derived from cultural differences, age, labour record, etc. Here are some of the most common ones:

- Different cultures present different understanding on the use of a system, the instructions given, the symbols used, or even they can provide limitations on the use of certain authentication methods. For example, in certain countries an X-mark is understood as a negative action (e.g. cancellation) instead of a selection. Another typical example is that in certain cultures, the population do not show the whole face publicly, which means the prevention of using face recognition systems, or even challenging iris recognition.
- Elderly population usually present higher error rates in biometric systems, due to several factors, such as a non-accurate interaction, or the gradually deterioration of their biometric trait.
- Elderly people, usually, present more difficulties to adapt to new technologies and processes, particularly if these are considered as not-natural.
- Manual workers may have their fingerprints very eroded, preventing the use of this biometric modality.

Temporal disabilities

But what is more important is that those considered nondisabled people, may present temporally disabilities equivalent to the ones mentioned above. For example a person may break his leg and then needing mobility-aid devices. If the system is not adapted to this kind of situations, those citizens won't be able to use the system until the disability disappears. In other cases, some uses may present a condition that dive some difficulties which may be considered equivalent to the disabilities mentioned. Examples are: mobility of a 7-9 month pregnant woman, difficulty in hearing due to otitis, existence of temporal bandage due to some accident, high degree muscle contractions, etc. All these situations are extremely common in our daily life and may minimize the universality of systems, and may be easily covered when the system is accessible. Therefore not only the disable people, but the whole population, benefits for this kind of approach.

3.3 Summary

This part has introduced the state of the art of the main topics of this Thesis: biometrics, usability and accessibility. The main works in those areas have been written down and analysed in the following chapters. Next part deals with usability in biometric recognition systems, analysing the main methodologies up to now and suggesting improvements through the experiments carries out.

Part III: Usability Evaluation Methodology

Chapter 4

Initial evaluation methodologies

The main goal of this Thesis is to complete the H-B interaction testing of biometric systems, as the most relevant usability evaluation methodology. Moreover, the most common usability evaluation tools are included. Therefore, this work has several initial inputs, including the previous methodologies for assessing biometrics (e.g. the performance evaluation methodology –ISO/IEC 19795-), previous biometric usability models (e.g. the HBSI) and finally, the main works on usability (e.g. NIST, Norman Nielsen, HCI, etc.).

The objective of this section is to analyse the H-B interaction, taken as the baseline, and suggest improvements. To overcome this, we have carried out several experiments which also validate the new methodology updates. As in the UCD design, the followed method is based on cyclical processes, namely: design, develop and evaluation. Once each iteration ends, the outcomes obtained are applied to validate the methodology and a new cycle starts.

In the following subsections, ISO/IEC 19795 and the H-B system interaction testing are introduced in order to understand what is already done and what is not yet covered. Next, all the experiments performed are explained in UCD style. Finally, we explain the suggested modifications to the current methodology and the conclusions extracted. Thus, we divide these sections in Current Methodology, Experiments and Modifications and Conclusions.

4.1 The H-B interaction

4.1.1 The ISO/IEC 19795

According to its definition, the ISO/IEC 19795 "*is concerned with the evaluation of biometric systems in terms of error rates and throughput rates. Metrics for the various error rates in biometric enrolment, verification and identification are unambiguously specified. Recommendations and requirements are given for the conduct of performance evaluations through the steps of planning the evaluation; collection of enrolment, verification or identification transaction data; analysis of error rates; and the reporting and presentation of results*''.

Therefore, the ISO/IEC 19795 is a complete methodology for assess biometric performance, including how to proceed with metrics, requirements and reports. It consists of these parts:

- **Part 1: Principles and Framework.**

This part establishes general principles for testing biometric performance in terms of error rates and throughput, specifies performance metrics, requirements on test methods, recording of data and reporting of results. Furthermore, it provides a framework for developing and describing test protocols

- **Part 2: Testing Methodologies for Technology and Scenario Evaluation.**

Part 2 provides requirements and recommendations on data collection, analysis, and reporting specific to two primary types of evaluation: technology evaluation and scenario evaluation.

- **Part 3: Technical Report on Modality-Specific Testing.**

The purpose of this part is to present and define methods for determining, given a specific biometric modality, how to develop a technical performance test.

- **Part 4: Interoperability Performance Testing.**

This part prescribes methods for technology and scenario evaluations of multi-supplier biometric systems that use biometric data conforming to biometric data interchange format standards.

- **Part 5: Grading scheme for access control scenario evaluation.**

This part of ISO/IEC 19795 is concerned solely with the scientific ‘technical performance testing’ of biometric systems and subsystems to be used for access control.

- **Part 6: Testing Methodologies for Operational Evaluation.**

This part is focused on operational test which evaluate complete biometric systems in the targeted operational environment with the target population. Tests may encompass performance monitoring of operational systems or assessment of performance in operational trials.

- **Part 7: Testing of on-card biometric comparison algorithms.**

This part of ISO/IEC 19795 establishes a mechanism for measuring the core algorithmic capabilities of biometric comparison algorithms running on ISO/IEC 7816 integrated circuit cards.

4.1.1.1 Performance evaluations

The ISO/IEC 19795 provides performance methodologies for 3 kinds of evaluation: technology, scenario and operational evaluations. To better understand this standard, we provide a brief definition of each kind of evaluation:

- Technology evaluation: It is carried out under a controlled environment, generally in a laboratory. The testing is performed offline and it is applied most of the times for testing algorithms. The experiment repeatability and traceability are feasible.

- Scenario evaluation: This evaluation is performed with realistic groups of users, using products close to real situations in real time. The scenarios are similar to real scenarios and the repeatability of the experiments is considered possible (keeping environment and other variables). Nevertheless, changes in the participants may vary the results.

- Operational evaluation: Are carried out in real environments under real conditions. Parameters of the evaluation are measured but not controlled. Therefore, this kind of evaluation is not repeatable.

All the experiments carried out during this Thesis are scenario evaluations: we have replicated real environments and controlled all the variables when possible. Further usability evaluations in real time and real scenarios (operational evaluations) would also be helpful.

4.1.2 Evaluation methodology for Human-Biometric system interaction testing of biometric systems

In this section, the main reference of this current Thesis is analysed in detail. First, an introduction and the objectives of the methodology are given. Then, the main points of the evaluation are explained and finally, the points to improve are cited. As long as the complete evaluation methodology is extensive [2], only the highlights are explained in this Thesis.

4.1.2.1 Introduction

In 2013, Fernandez-Saavedra et al. published an evaluation methodology of the environment in biometric recognition systems [2] following ISO/IEC 19795 and Common Criteria. Within the same document, an evaluation methodology for usability in biometric systems was also published, following the same guidelines. This methodology includes:

- Relevant factors
- Relevant metrics and measurements
- Evaluation conditions specification
- Requirements for planning, executing and reporting

Experiments for validating the methodology and conclusions are available in [2]. The starting point of the methodology (H-B interaction from now on) has been the HBSI model, which had previously split up the usability analysis in 3 parts: human, biometric system and their interaction (Figure 10 shows the H-B interaction conceptual model). As a differentiation point, Fernandez-Saavedra has included the Environment as a key concept in the H-B interaction with respect the HBSI. Therefore, the importance of the environment during the user-biometric system interaction is highlighted.

Even when this evaluation model "*does not provide either which influential factors have to be analysed or the specific procedures to carry out such tests*" as Fernandez-Saavedra claimed, HBSI suggests several usability measurements which may influence performance in biometrics. Thus, the HBSI model has been tested during the first stages of this Thesis.

Figure 10. H-B interaction conceptual model [2]

4.1.2.2 Objective

The main objective of the H-B interaction is to discover which usability factors influence in performance and to what extent. To do this, performance results under different conditions are obtained. For each different condition, one factor is isolated and the rest fixed. Therefore, measuring that factor's influence in results is feasible.

This procedure is not new and was also used by NIST and HBSI, but not providing enough information (specific details needed to fully understand the experiments: e.g. user's characteristics or environment).

In short, this H-B interaction methodology was developed in response to the lack of a formal methodology which provides means to conduct a proper usability evaluation of biometrics.

4.1.2.3 H-B interaction Factors

The methodology defines three kinds of factors which may influence usability in biometric recognition systems:

1. Factors depending on the **biometric capture device**. These factors are those related to the design, position or condition of the capture device

2. Factors depending on **human** beings. These factors have to do with the user's characteristics including biometric traits.

3. Factors depending on the **human-biometric system interaction**. These factors depend on the interaction of users with the biometric system.

These factors lists were completed taking into account only fingerprint recognition, leaving the rest for future works.

4.1.2.4 H-B interaction Metrics

The metrics used in the H-B interaction were extracted from the HBSI evaluation method. Then divided in 3 blocks according to the main 3 factors of the methodology: usability, ergonomics and signal processing. The HBSI model shows in a diagram which measurements are taken into account and how they are divided (Figure 9). In Figure 11 are the metrics suggested in the H-B interaction. Both uses the traditional usability metrics, namely: satisfaction, efficiency and effectiveness. Furthermore, they include HBSI metrics and ergonomics. The H-B interaction includes also signal processing.

4.1.2.5 Specification of the evaluation conditions

The first step in the evaluation design is to fix the evaluation conditions (factors to be assessed). Here the terms REC (Reference Evaluation Conditions) and TEC (Target Evaluation Conditions) are introduced. The REC are the baseline conditions and the TEC are the target conditions where one or more H-B interaction factors have been changed. Then, the results from TEC and REC are compared in order to check the influence of the chosen factors.

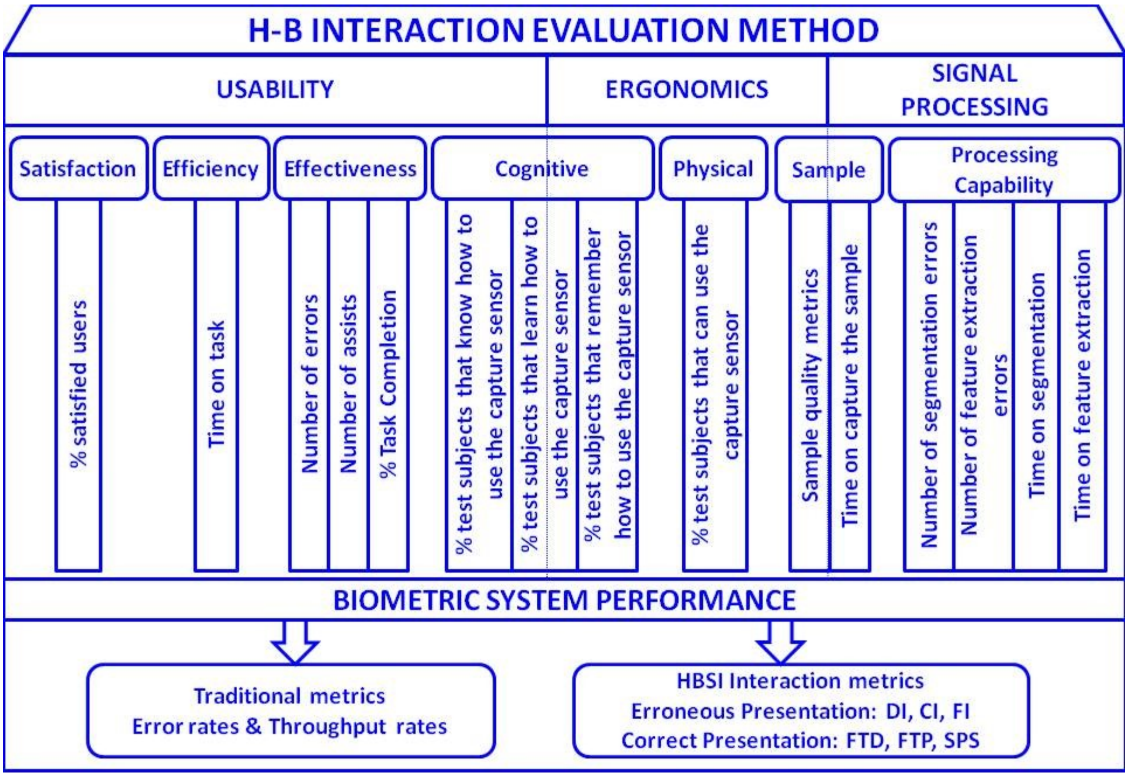

Figure 11. H-B interaction Evaluation method [2]

4.1.2.6 Fundamental requirements for planning an H-B interaction testing

The requirements included for planning a usability evaluation are those belonging to the ISO/IEC 19795 parts 1 and 2. Moreover several factors related to the H-B interaction were added. In Figure 12 is a schema of those requirements.

4.1.2.7 Fundamental requirements for executing an H-B interaction testing

This part of the methodology includes the necessary activities to perform when executing an H-B interaction testing (once the planning has been completed). It includes pre-test activities, test activities and post-test activities.

4.1.2.8 Fundamental requirements for reporting an H-B interaction testing

In this last section are the requirements for reporting the results obtained during the execution phase. These results include comparisons with the baseline, modifications, inconveniences, errors, relevant comments and final conclusions among other factors.

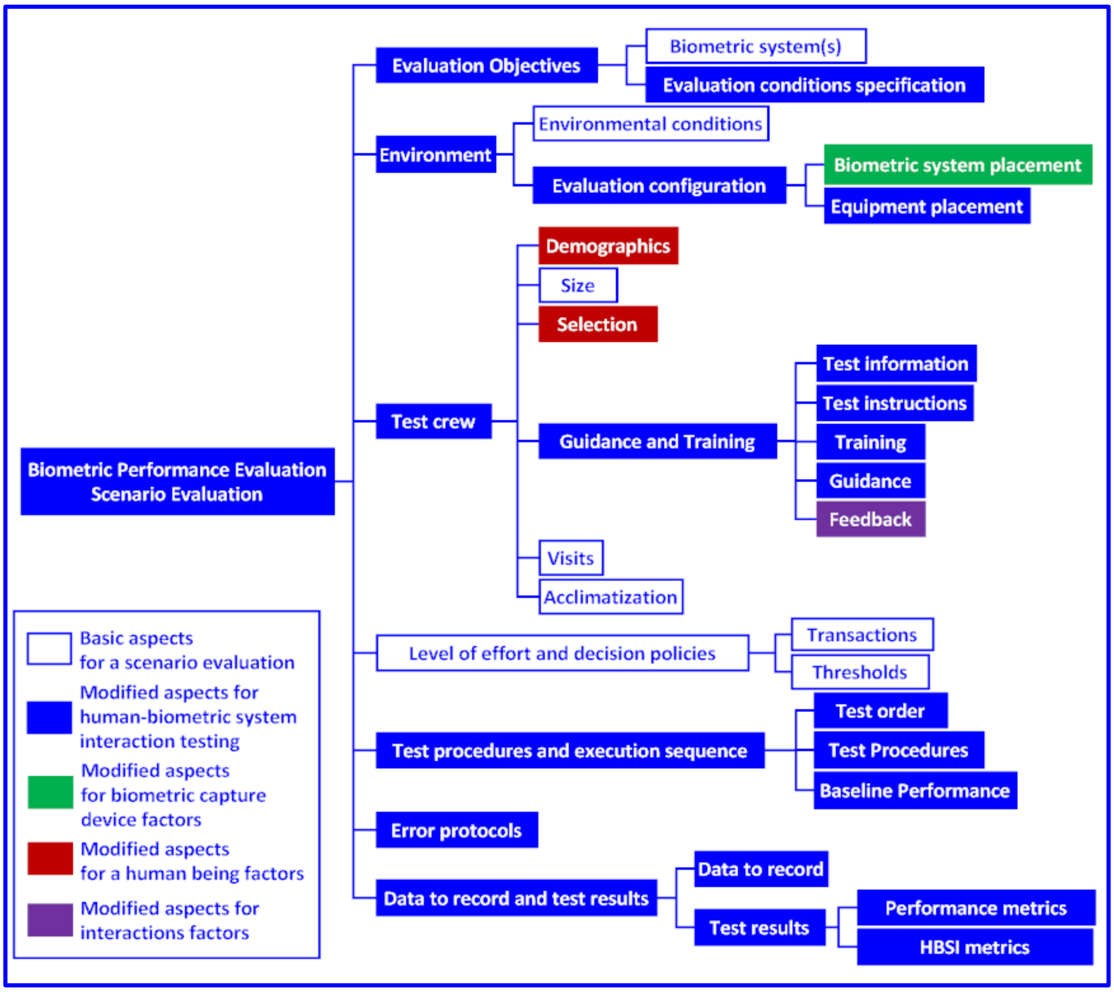

Figure 12. Scenario evaluation specification in the H-B interaction testing [2]

4.1.3 Preliminary points of improvement to the H-B interaction testing

Once the H-B interaction testing methodology has been defined, some improvements are suggested. The rest of the suggestions are obtained directly from the continuous validation of the methodology (experiments carried out during this Thesis). This section covers the initial points of improvement:

A) Extending the methodology to all modalities. The H-B interaction as it currently is, is basically designed following a typical fingerprint recognition scenario. Most of the factors and measurements are focused only on fingerprint and other modalities are left out.

Including new modalities would involve also the inclusion of several new kinds of user-system interactions. For example, dynamic modalities require a higher extent of user participation, so that, the usability evaluation tends to be harder than in static biometrics. This is because presentations require a bigger effort from user's side and the quantity of possible misuses increases substantially. Furthermore, in many dynamic modalities there is not a defined "correct" way to present the trait (e.g. there are not specific, complete and commonly approved guidelines to sign in DSV, apart from accomplishing some criteria such as performing the signature within the singing area boundaries). Thus, the range of possible user-system interaction errors is bigger than in static modalities, where the process is more straightforward.

Changes in the H-B interaction: Including other modalities when testing usability, would involve modifications in some parts of the methodology as a consequence of having several changes in the interaction. New influential factors and measurements are a direct consequence of this. One clear example is the measurements taken from the HBSI, which are modality dependant.

How is this addressed in this Thesis: Experiments carried out during this Thesis include other modalities (static and dynamic) which allow us to derive further updating to the methodology. We have mainly worked with DSV during the whole Thesis, which is widely known by users. Some of the experiments were performed following the HBSI schema and metrics. Then, appropriate modifications are suggested.

B) Extending the methodology to the new biometrics paradigm. Biometric recognition systems are not isolated and fixed anymore: biometrics is getting mobile. Several parameters in the H-B interaction are based on the traditional scenarios (e.g. border controls or ATMs). Moreover, the use of biometric recognition has also changed. Biometrics are used now to unlock devices, sign documents or for continuous authentication among others

New evaluations under different conditions should be carried out. Biometrics should be assessed in mobile devices, such as smartphones or tablets, to better understand the new usability concerns. Furthermore, evaluating usability under the new biometric usages is convenient.

Changes in the H-B interaction: Increasing the number of scenarios would involve changes in the H-B interaction in terms of environment and influential factors (weather, lightning, other people, etc.). Moreover, new ergonomic challenges come along also with the new mobile devices uses (different positions, extra elements such as styluses, etc.). This includes also the use of the mobile devices features used to collect biometric data: camera, microphone, GPS and so on and so forth.

How is this addressed in this Thesis: Most of the experiments performed deal with biometric recognition in mobile devices, including ergonomics in all of them. This include experiments with embedded biometrics (fingerprint) and using the devices touchscreen (DSV). Extra elements have also been studied (styluses).

C) Applying the traditional usability assessment. Usability evaluations are carried out since many years ago and there are various methodologies developed [74]. One of the main points of traditional usability is evaluating the system several times before the final design. In fact, best practices in usability assessment advice to evaluate the usability from the early stages. The H-B interaction methodology starts only once the product is finished.

Both H-B interaction and HBSI have targeted the repercussions in performance when evaluating usability. Then, no matter the user is comfortable or not, when the performance is adequate. Nevertheless, the final goal of usability is to make systems easy and comfortable. Then, more attention should be put in user's acceptance and satisfaction.

Changes in the H-B interaction: Modifications in the H-B interaction may include the main usability evaluation tools. Some of the most relevant are prior usability evaluations (from the early stages of the product), further analysis of user's feelings and reactions by means of surveys, previous interviews, etc. Interfaces assessments, like the Norman Nielsen heuristics [57] should also be included in the H-B interaction. The user acceptance factor is also added to the methodology.

How is this addressed in this Thesis: Several user's opinions are gathered by means of video recordings, notes, interviews and surveys in all the experiments. This leads us to collect several subjective information from users and real time information of the interactions. Then, critical points of the interactions may be located and the observed problems overcame (ergonomic improvements, different guidance, etc.).

Moreover, a user acceptance study on fingerprint recognition was also carried out. The choice of fingerprint recognition is motivated by the high usage of this technology. As the goal of the experiment is to study the user acceptance of biometrics, the most well-known and widely used modality was considered as the most appropriated. Surveys in this experiment were fulfilled before (pre-surveys) and after the experiment (post-surveys) to better understand the user's opinions once they have tried the technology.

D) User's behaviour and moods are relevant. There are so many variables when analysing user's feelings in HCI or UX. Biometric recognition, which is always joint to security environments, involves a wide range of reactions in users (in accordance with the user's mood, the relevance of the situation, etc.). Many inconveniences in usability and performance may occur due to user's distrust, fear, nervousness and so on. The H-B interaction covers only a few behavioural aspects. Also the biometrics evolution to mobile scenarios may cause different reactions in users, as mobile devices are deeply integrated into the society (smartphones, tablets, etc.).

Even when measuring user's behaviour and moods is not straightforward (and may cross the border between engineering and psychology), this analysis is necessary. Biometrics should be tested inducing different moods in users as in real situations. User's behaviour may be more visible in dynamic modalities, where users are required to participate in a higher extent.

Changes in the H-B interaction: These changes affect mainly to influential factors related to the user. Even when this change seems to be not really remarkable in the methodology, user's mood could involve several variations in final results in terms of performance and usability. For example, unmotivated and non-cooperative users may bias a whole evaluation.

How is this addressed in this Thesis: During this Thesis we have controlled user's moods as much as possible. Pauses were made during the experiments and there were always several days between sessions to not make the users feel tired or bored. Moreover, we conducted an experiment where the stress was induced in order to test the real influence in biometrics. As biometric recognition is always joint to security environments, stress is one of the most common side effects.

F) Accessible biometrics. Nowadays, biometric recognition systems are thought mainly for experienced users or users having certain skills. People who do not have technological knowledge may struggle a lot when using biometrics. This is especially true for people with accessibility concerns, who instead of benefit from biometrics (which is supposed to be transparent and straightforward), find several difficulties or cannot use it.

Designing final inclusive products would ease the use of biometrics and make it accessible for a wider percentage of people. This must be carried out including users with accessibility concerns in the products design from the very early stages.

Changes in the H-B interaction: The inclusion of accessibility involves many changes in the H-B interaction at many levels including factors and metrics. The European Standard EN 301 549 is the major contribution to the H-B interaction, including hardware and software measurements for testing accessibility.

How is this addressed in this Thesis: We strongly believe that biometric recognition could help people with accessibility concerns to perform common tasks requiring authentication (e.g. ATM transactions, border crossing, etc.). Having this in mind, we have decided to deeply analyse accessibility in biometric recognition systems and therefore, in the H-B interaction.

There were not accessibility factors included in the H-B interaction, then the basic requirements are not satisfied. In this Thesis we have completed the H-B interaction including accessibility concepts at user and system levels. Once this is accomplished, the accessibility is also guaranteed during the user-system interaction. The modifications made are the inclusion of the EN 301 549 and implicitly, the inclusion of the WCAG 2.0. Further modifications are the inclusion of accessibility metrics and the re-adjustment of various traditional usability terms.

According to these new modifications we have decided making further evaluations to validate the methodology. Further study and experiments were carried out and are in Part 4: Accessibility in Biometrics.

All the experiments were carried out with the goal of including all of these suggestions to the H-B interaction methodology. Therefore, all of these preliminary suggestions were applied to the experiments done. In the following section, the experiments carried out are described.

Chapter 5

Experiments for searching improvements in the initial methodology

In this section are all the experiments carried out in this Thesis regarding usability evaluation of biometric systems. As above mentioned, these experiments were made in order to validate the H-B interaction methodology. Furthermore, the preliminary improvements cited in 4.1.3 were applied when designing the experiments. The experiment's order is as follows:

1. *Performance evaluation of behavioural biometrics in mobile devices.*
2. *Ergonomics in behavioural biometrics.*
3. *Ergonomics in behavioural biometrics including styluses.*
4. *Analysing moods in behavioural biometrics.*
    a. *Time efficiency*
5. *User Acceptance in biometric recognition.*

The experiment's order responds to the following reasoning: first, we measured the performance of a current DSV algorithm in mobile devices because the planned researching line involves the use of mobile biometrics. Then, this analysis became our performance benchmark. As mobile biometrics involves changes in the HCI, ergonomics studies were scheduled as posterior experiments. Other important factor is the proper analysis of the UX and the User Acceptance, then last experiments are intended to better analyse user's feelings and reactions.

Notes of the experiments

Before going through the experiments it is important to highlight various points which clarify some characteristics of them:

- We remark the importance of behavioural biometrics in the experiments set, according to the preliminary improvements suggestions. DSV was the main modality used during this Thesis phase due to its popularity in mobile biometrics and easiness. Moreover, most of people is used to sign on paper, so they already have skills.
- The H-B interaction has been applied to all the experiments. Nevertheless, the application of the whole methodology involves taking into account a high number of variables. In order to make the results easier to understand, only a limited number of factors are left variable while the rest are fixed to not influence results. Each of the experiments focus on different parts of the methodology. Then, several parts of the H-B interaction are omitted but the valuable information is shown. The valuable information allowed us to validate the methodology and derive usability conclusions.
- We tried to give the same importance to all the experiments to maintain the document consistence but this was not possible as long as we have extracted more conclusions from some of them. Different factors are highlighted in each experiment (e.g. in the user acceptance experiment we give further details of users characteristics), though all of them follow a similar roadmap.
- None of the experiments was made under real conditions, but most of them are based on commercial products. Furthermore, the designed scenarios and several evaluation characteristics are really close to real environments, which encouraged us to believe that similar results would be obtained in daily use.
- In terms of the performance results obtained, there are not statistical data analyses in these experiments because the majority of them has an insufficient number of users as to derive broad conclusions. Nevertheless, in the usability field, some experts argue that in quantitative studies, the use of at least 20 users allows confident results [75].

5.1 EXPERIMENT 1: Performance evaluation of DSV in mobile devices

This first experiment [48] analyses the performance of a DSV algorithm over a dataset of signatures gathered through mobile devices. One of the big challenges of this research was to discover if the handwritten signature modality in mobile devices should be split into two different modalities, one for those cases when the signature is performed with a stylus, and another when the fingertip is used for signing. The results include interoperability, visual feedback and modality tests.

Motivation of the experiment

This experiment was not designed for obtaining usability results. When we started this research line and decided to focus on DSV, we did not have a clear reference to match with further experiments. Through this experiment, we have a benchmark in DSV which allow us to derive comparisons with further usability testing where the performance is part of the results.

Influence on the H-B interaction

Thought this experiment is not focused on usability, some of its points have to do with the changes suggested in 4.1.3., as long as it incorporates a dynamic modality to mobile devices. The variability of devices and the use of the fingertip and styluses for signing are both factors to be added to the H-B interaction (having influence on the system and user-system interaction). Other factor to remark is the use of non-biometric devices to acquire the biometric traits, using the device's touchscreen to collect the signatures.

5.1.1 Evaluation set-up

This section contains information about the evaluation characteristics, including hardware, software, users, visits, biometric features and further details.

Mobile devices

Five different devices were used in this evaluation (Figure 13), with one of them (Samsung Galaxy Note) used in 2 different modes (i.e. signing with the fingertip and signing with its own stylus). Devices have been divided according to the object used for signing, being one group of those using a stylus (Asus Eee Pc T101 MT and Samsung Galaxy Note), while the other group used the fingertip (Blackberry Playbook, Apple iPad 2, Samsung Galaxy Note and Samsung Galaxy Tab) to perform the signature. This division have been done with the preliminary hypothesis of considering that we may be talking about two different biometric modalities, as the signing object is changed.

It is important to note that, in order to reduce the number of external variables that may impact the analysis of the results to be obtained, it was decided to fix the signing environment. Therefore all devices were placed on a table, and all participants were requested to sign on them. But in order to not impact the way the user was signing, they were allowed to rotate the device at their own wish, so in order for them to feel as comfortable as possible. Also the software for capturing the signatures in each of the platforms used has been developed in the same way, as to minimize the variability of the user during the process of signing. Therefore all platforms had a software that only had white rectangular area to sign, plus two buttons, one for accepting the signature, and the other for deleting it.

As a reference is needed for comparing results, two extra devices were used to establish the baseline system (Wacom STU-500 and the Wacom Intuos 4). Actually only the STU-500 was used as a reference, as the Wacom Intuos was used without inked-stylus and paper (obtaining blind signatures).

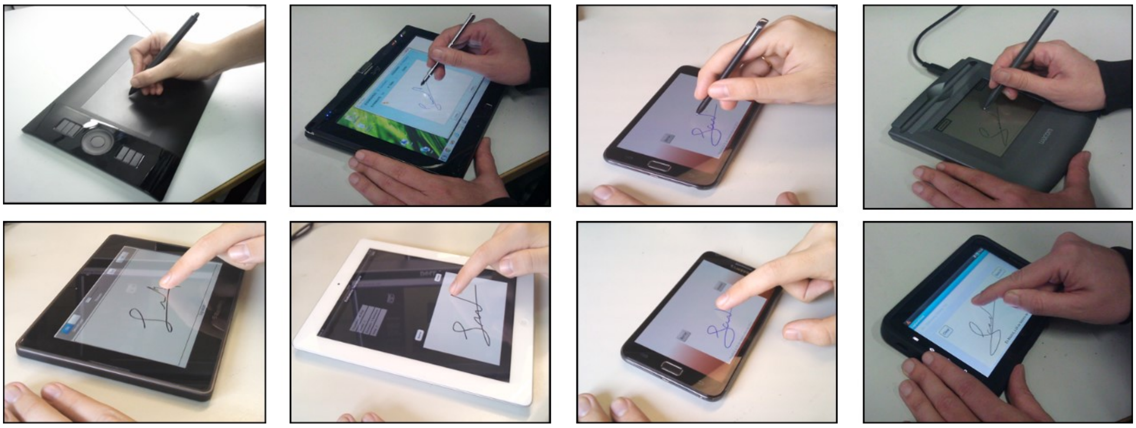

Figure 13. Devices used during the evaluation. In the first row are the stylus-based devices. From left to right: Wacom Intuos 4, Asus Eee Pc T101 MT, Samsung Galaxy Note and Wacom STU-500. The fingertip-based devices are in the second row. From left to right: Blackberry Playbook, Apple iPad 2, Samsung Galaxy Note and Samsung Galaxy Tab.

<u>Users</u>

There was not any a priori rule for selecting participants for the data acquisition apart from rejecting users younger than 16 years old due to the claim that handwritten signature is not persistent in people below that age. Test crew age distribution is divided in ranges: 25-30 years old (22 users), 30-40 years old (13 users), 40-50 years old (5 users) and over 50 (3 users). The total amount of users was 43, being 30 men and 13 women. All test crew members have a university degree except 3 that have elementary studies. The nationalities of the users were Spanish (36), Colombian (5) and Ecuadorean (2). Colombian and Ecuadorean signature style is the same as the Spanish style (e.g. is not performed only with the initial letters). Participants signed on the devices in their chosen most comfortable position, being assured that they were sat on a chair and placing the devices on a table. None of the users had previously signed with the fingertip, so that, for most of them it was a strange process as they commented during the evaluation. Furthermore, all users considered their signature less accurate and the procedure slower with the fingertip than with the stylus. There were pauses during the process when users felt tired (at least one after finish each device).

Evaluation protocol and Biometric processing

The evaluation protocol is composed of sessions, scenario, capture process, user guidance and biometric processing. It is important to note that this evaluation entails sample acquisition in real time, as well as off-line signal processing. The signal processing was decided to be done after the whole set of samples is captured, as not to influence the user act of signing depending on the results being obtained. The chosen scenario consists of the user sitting on a chair and signing with the capture device resting on a table. Users did not sign in all cases in the same physical place but in all those places where the acquisition process took place, the scenario was modelled maintaining the same characteristics: i.e. table and chair height and shape were kept equivalent, while environmental conditions were set as traditional indoor conditions.

Regarding the data acquisition process, the order in which the devices where taken for signing was randomized in each session for each user in order to avoid results biased due to habituation or tiredness (e.g. there can be differences between the signatures provided at the beginning of the session compared to those at the end). Each user went through 3 sessions separated at least 1 week between them and 20 signatures were collected in each device per session, i.e. a total amount of 480 signatures per user (60 signatures x 8 devices). Then, the evaluation reported a total of 20 640 signatures, stored in ISO/IEC 19794–7 format [76].

The whole process was supervised by an operator who was on charge to assure the proper operation of the evaluation. The operators' tasks were: checking the correct state of devices and their applications, explaining how to use the applications to participants, staying alert in case of any application misuse and to randomize the order of acquisition devices. As it was said, the objective was to make the process as realistic as possible, so no advice on how to sign was given to the users.

Once all users have completed the data acquisition process (the 3 sessions), the biometric processing started with all the signatures gathered. The experimental results were extracted from the algorithm executed in Matlab using a desktop computer. It made more sense than execute the algorithm in a mobile device due to the multiple comparisons among different platforms that have to be made (this process would take too much time in a mobile device due to processing constraints).

For the enrolment, the first 3 samples from each user-device were used to create the each of the biometric references for each user-device combination, and the rest were taken to obtain the distribution of the intra-class and the inter-class comparison scores, obtaining the FMR, the FNMR and the EER. The number of 3 samples to enrol was chosen as to approach user acceptance during the enrolment phase, as a larger number of acquired samples could mean a certain level of user rejection to the system. In order to calculate the FMR, signatures from other users as random forgeries were used as no skilled forgeries were applied to this evaluation.

5.1.2 Signal processing and feature extraction

In this section, the process from the signatures acquisition to the feature extraction is explained. According to the evaluations' nature, the verification process is carried out once all the signatures were gathered, so that, no verifications are done until the end of the capture process of all the users. Following the standard ISO/IEC 19794-7 several signals can be extracted from the signatures' capture process, but the more signals are obtained the more time is spent in processing, so that, considering potential processing limitations, as well as acquisition limitations from some of the devices, only four signals were obtained: time, S and X and Y coordinates. The S signal indicates whether the pen/finger is in contact with the screen or not offering two pressure levels. It is important to highlight that pressure, which is considered to be an important discriminative signal, has not been considered in this evaluation, as some devices were not able to provide reliable data regarding this characteristic, or even no data at all. During the signature performance, all the points that the user touch in the screen are collected (X, Y and time) and when a signature is finished and accepted by the user, the data is converted to a ISO/IEC 19794-7 biometric data record in order to complete a database as interoperable as possible.

The algorithm applied for obtaining error rates after the evaluation process was a DTW-based algorithm. This algorithm allows an optimal alignment between two sequences of vectors of different lengths using dynamic programming. From this alignment a measure of distance between two temporal patterns is obtained. The signals used as input from signatures are the X and Y time-series coordinates. From this X and Y signals, also their first derivative signals have been calculated, VX and VY signals. In order to achieve interoperability between the different devices used to acquire the signature data, several normalization steps have been adopted:

1. Equi-spacing by linear interpolation. This normalization avoid problems for DTW algorithm due to high difference between the number of sample points acquired by the different devices. The X, Y and time signals are transformed to an equispaced 256-point temporal sequence, by linear interpolation.

2. Filtering of X and Y signals. Smoothing of the X and Y channels by a low pass filter to remove the noise introduced by the capture device during the data capture process.

3. Calculation of Speed X and Y signals (VX and VY). A regression formula [77] have been used to obtain an approximation of the derivative signal, providing softened waveforms, removing slight noise variations.

4. Location and Size normalization. X, Y, and its derivative signals VX and VY, are normalized using their mean and standard deviation:

$$S^* = \frac{(s - \bar{s})}{\sigma_s}$$

The algorithm error rates achieved in previous experiments with portable devices were EER = 1.8% for random forgeries and EER = 7.6% for skilled forgeries. As the algorithm performance is out of the topic of this research, its detailed description and its previous performance results are not included here, although that information can be obtained from [78]. This algorithm is used in the rest of experiments (where DSV is applied) included in this Thesis.

5.1.3 Experiments and Results

This section shows the results obtained for four different kind of experiments. The first one consists of the intra-device error rates, containing the comparison with the baseline. The second experiment is based on studying the inter-device/intra-modality case, by performing two different experiments, one for the devices using stylus, and the other for those using the fingertip. In all these cases, the samples of each device were compared against the biometric reference obtained with another device of the same modality. After that, and using the device that is able to acquire signatures both by the use of a stylus and by the use of the fingertip (i.e. Samsung Galaxy Note), the 'inter-modality' analysis was done. With these results, the analysis was continued by obtaining the inter-device/inter-modality error rates for all different combinations of enrolment device and verifying device. Finally, a potential operating scenario was tested, where the enrolment is done with the reference device (STU) and the comparison is done against the rest of the devices. It is important to note that, in order to save space, the data analysis will be presented here in terms of EER, whereas the comprehensive set of evaluation data can be found in the following webpage (*http://guti.uc3m.es/Graphics*).

Intra-device performance analysis

Table 1 shows the EERs achieved by each of the devices individually (in addition to their main features). It is important to note that within these results, the Intuos digitizer is also included, getting also information about the changes in the error rates when no visual feedback is provided to the user when signing. The first four devices are used with stylus to sign, while the rest are used with the fingertip.

The best performance was obtained with the iPad2, over performing the stylus based devices. Being this a surprising result, two main hypotheses are proposed for such a performance. The first one is the size of the signing area, and the second is the input resolution, which is the biggest among all mobile devices, although this second hypothesis becomes invalid when comparing the results with the STU device, whose input resolution outperforms the screen resolution, being 2450 dpi. Also, the Intuos has a really high screen resolution (5080 lpi) but the visual feedback supposes a big concern and the performance is affected. Nevertheless, according to the Table 1, there seem to be no main features which suppose a major parameter of performance (e.g. devices with big screens have a good performance, but not in all cases; the same occurs for devices with small screens as well). These hypothesis need to be further studied, and this is left as future work.

The best performance in stylus based devices was obtained with the reference device (STU) and Note-Stylus (i.e. Samsung Galaxy Note using its own stylus), which was described by most of users as a comfortable device for signing at the end of the evaluation. Another important result is that there is a significant EER difference between Intuos and STU, showing that although it is possible to use the technology without visual feedback, the performance is negatively affected. The influence of the resolution in the visual feedback is proposed as another hypothesis, because better visual feedback is given by the devices with better resolution (STU and Note). Even thought that hypothesis is well supported by most of the results obtained, in the case of iPad2 we have found a contra-hypothesis, as it results in better performance with a slightly worse visual feedback. It seems that cause-effect relationships could not be extracted from individual and separate causes, so the analysis of the effect obtained by the combination of causes shall be addressed. This is left for future work. A last important conclusion is that the EERs between the fingertip-based devices and the stylus based devices do not differ significantly in average. This means that the solutions using the fingertip present an acceptable performance, even though it is considered by users as somehow uncomfortable and not natural.

Table 1. Devices features and performance. (*) Software under Windows 7

Modality	Device	Visual feedback	Screen size	Signing space	Screen resolution	O.S.	EER %
Stylus Based Devices	Note Stylus	Excellent	5,29”	3,74”	1280x800 px (320 dpi)	Android	0,58
	STU	Excellent	5”	5”	640x480 px (2451 dpi)	Windows7*	0,63
	Intuos	Not available	10,81”	10,81”	223x139 (mm) (5080 dpi)	Windows7*	1,45
	Asus	Bad	10,1”	4,41”	1024x600 px (133 dpi)	Windows7*	1,10
Finger Based Devices	iPad2	Good	9,7”	5,4”	2048x1536 px (132 dpi)	iOS	0,19
	Note Finger	Excellent	5,29”	3,74”	1280x800 px (320 dpi)	Android	1,62
	Playbook	Bad	7”	6,2”	1024x600 px (169 dpi)	BlackBerry OS	1,87
	Tab	Good	7”	4,92”	1024x600 px (169 dpi)	Android	0,52

Intra-modality analysis

When comparing the samples of each device with the biometric reference from any of the other devices of the same modality, the EERs obtained do not differ much from the intra-device results seen previously. These rates can be seen in Figure 14 for fingertip-based devices and Figure 15 for stylus-based devices. In each of these figures, the EERs that are closer to a certain device represent the EER when that device has been used for enrolment and the linked one for verification. As it can be seen, in the case of fingertip-base devices, the EERs obtained is in the range of the intra-device rates and therefore, the conclusion is that there is interoperability within fingertip devices. Obviously there are some cases with slightly lower performance, such as enrolling with Playbook. Further analysis is left for future work in order to get a reason for that lower performance.

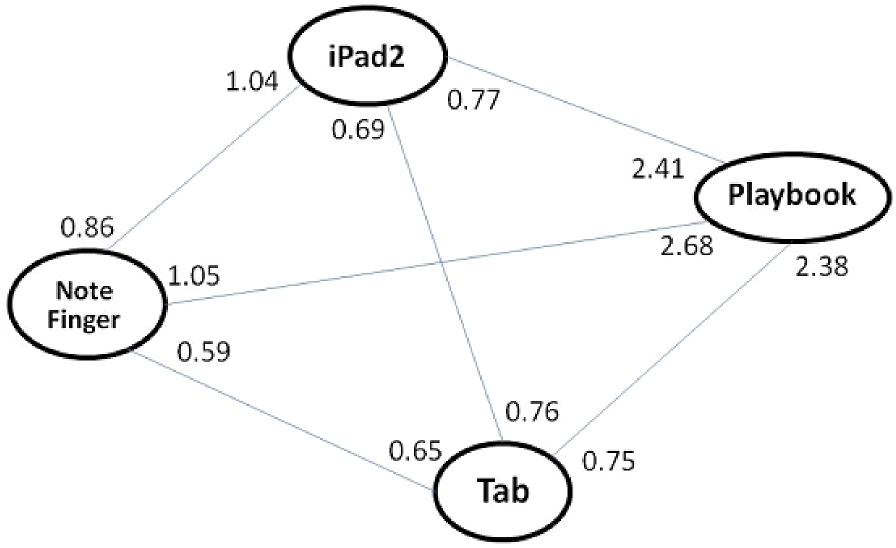

Figure 14. Performance interoperability results in fingertip-based devices

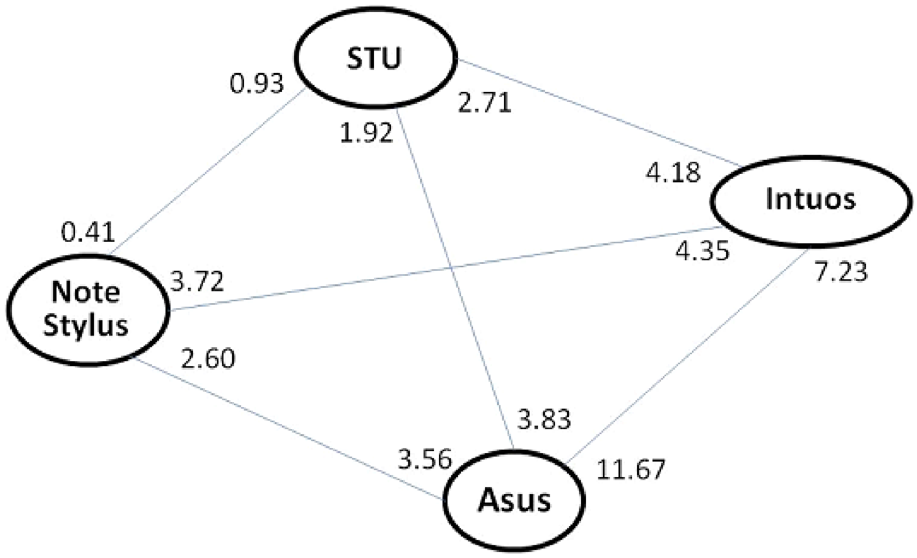

Figure 15. Performance interoperability results in stylus-based devices

Surprisingly, although users felt more comfortable with stylus-based devices, interoperability results are worst in average. Intuos and Asus devices provide EERs double than the ones for intra-device, not only when enrolling with them but also when being used for verifying. But EERs remain in the same order of magnitude, so, although the results are not as good as with fingertip-devices, we can determine that an acceptable level of interoperability has been achieved. It is also important to note that enrolling with Note-Stylus over performs the results when the enrolment is done with the reference platform. But considering the overall performance achieved, it can be determined that the most interoperable performance is achieved when enrolling with the STU digitizer.

Inter-modality analysis

According to the acceptable results obtained from both modalities isolated, the next natural step is to combine all the devices between them considering just one modality. Thus, through these experiments, the interoperability study is completed.

Enrolling with stylus based devices

In this case, two studies have been carried out. One enrolling with Note-Stylus and the other one enrolling with STU were chosen to be templates, due to their better performance in the intra-modality experiments. The comparison is done against all samples from all the fingertip-based devices. Figure 16 and Figure 17, represent the ROC representation of each of these two experiments.

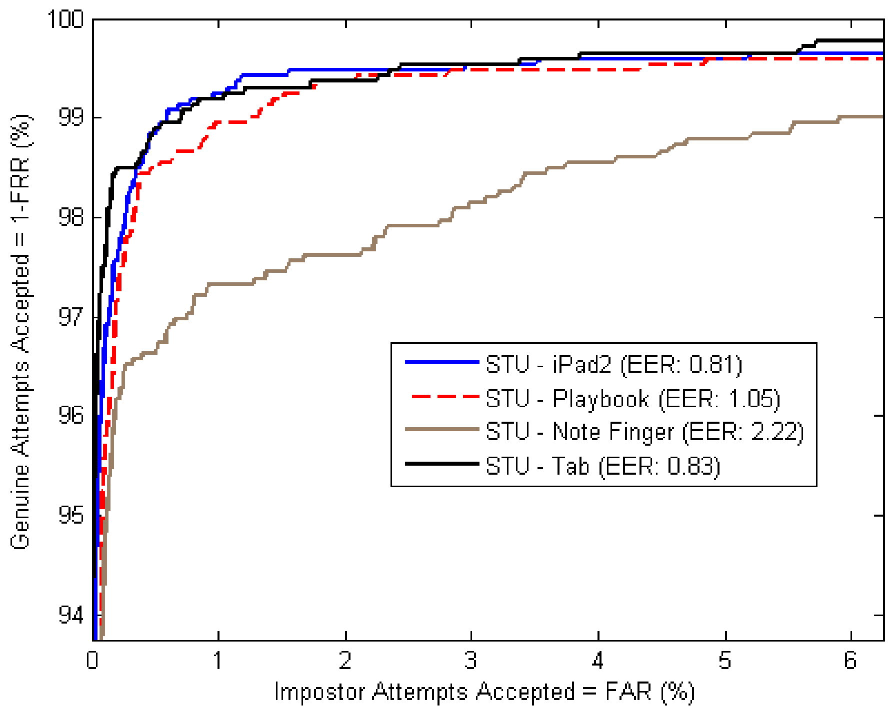

Figure 16. STU enrolment vs. Fingertip-based devices

In each of these two figures, not only the ROCs are represented but also the EER achieved in each of the cases is stated in the legend. As it can be seen, the error rates are in the same order of magnitude as the ones obtained for the intra-modality case. This enforces the previous claim that one single modality shall be considered. It is left for further work the analysis of why verifying with Note-Finger provides the worst results. The EERs obtained enrolling with Note-Stylus comparing with Note-Finger and vice versa are shown in Figure 18.

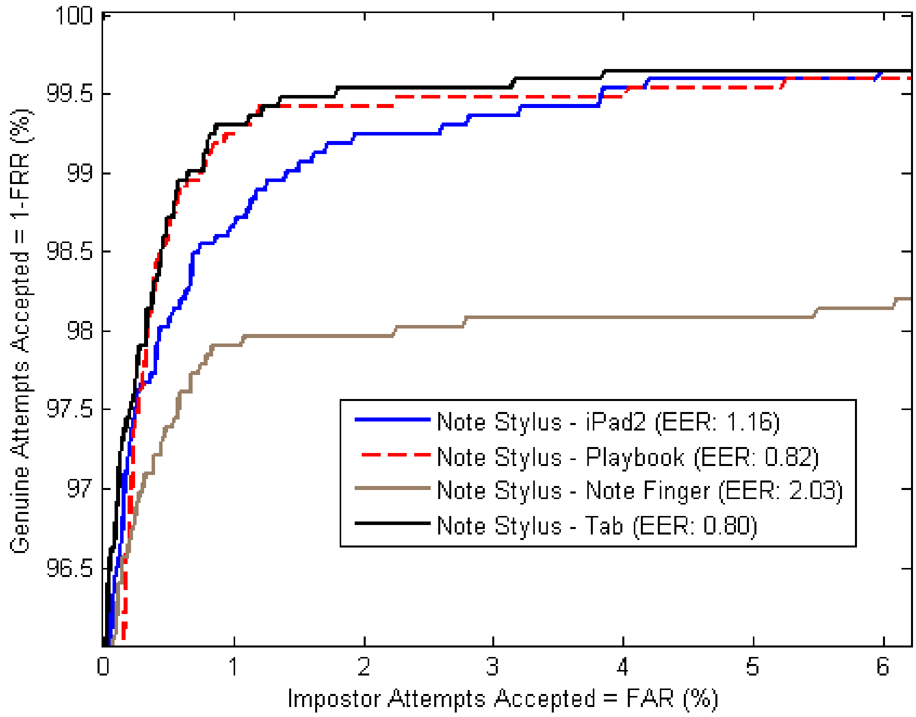

Figure 17. Note-Stylus enrolment vs. Fingertip-based devices

Enrolling with finger based devices

The same experiment was done but this time enrolling with fingertip devices. iPad2 and Tab were chosen as enrolment devices, and all the stylus-based devices where used for comparison. Figure 19 and Figure 20 show the results obtained in ROC curves, also providing the EERs achieved in both figure legends. As it can be seen, once again the error rates are in the range of the intra-modality values, so the single-modality conclusion is again enforced. It is also noticeable that verification with Intuos provides the poorest performance, which is to be related with the lack of visual feedback. This hypothesis has been left for future study. Within that future study, also the performance obtained with Asus shall be considered, as it also obtains lower performance rates than the other devices. In this case, the lower performance may be derived from the slower reaction of the screen during the acquisition process (i.e. for some users, the screen refresh was slower than desired, noting the user a retarded visual feedback, which was noted to be quite uncomfortable by some users).

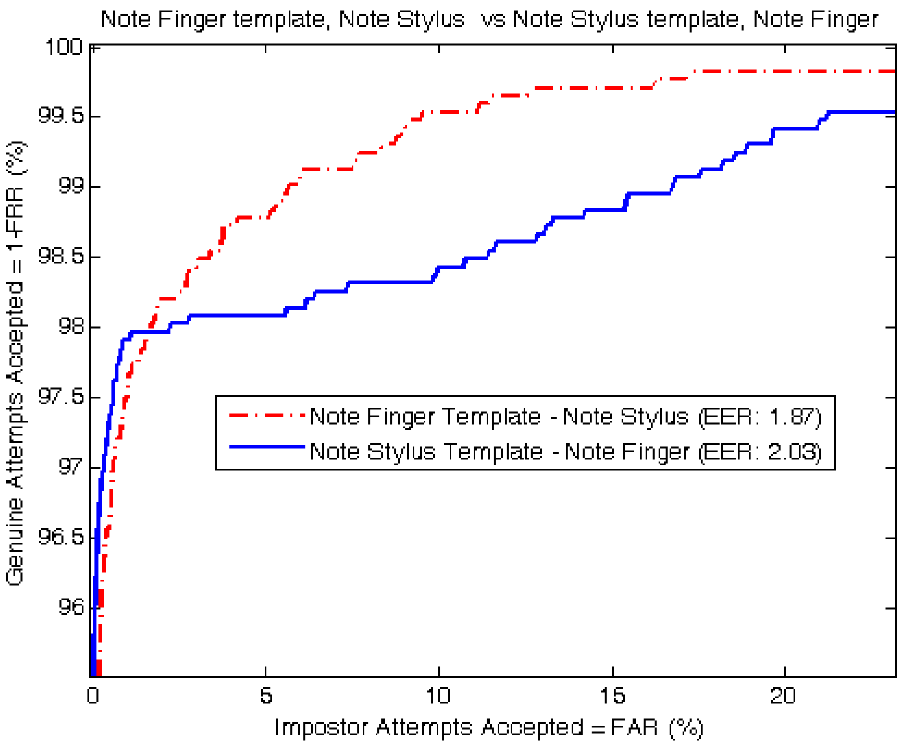

Figure 18. Note-Stylus enrolment vs. Note-Finger enrolments

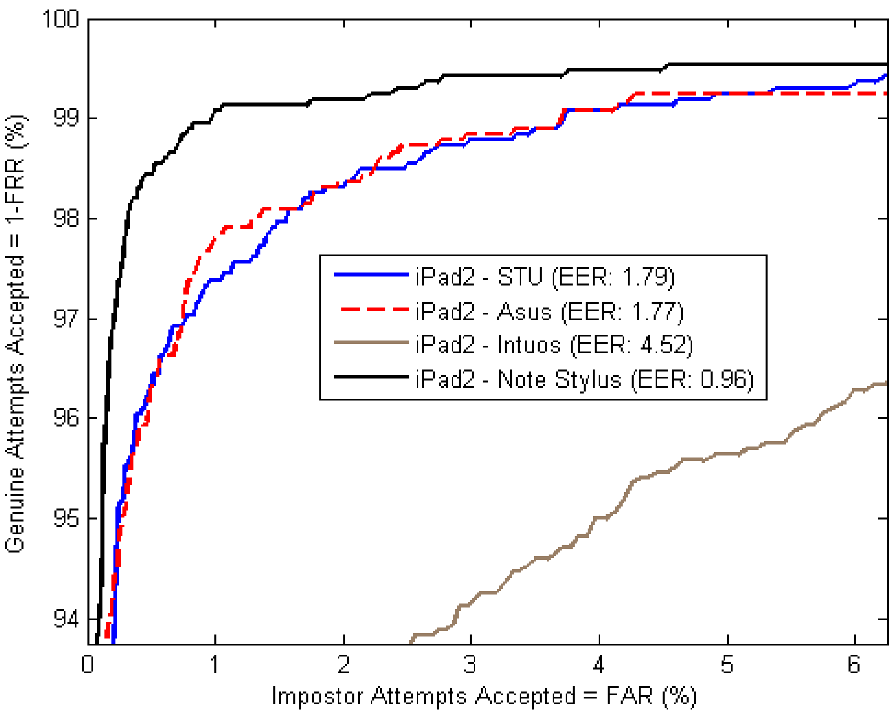

Figure 19. iPad2 template vs. Stylus based devices

5.1.4 Conclusions obtained

The results obtained are optimistic enough for having reached good interoperability levels in handwritten signature recognition in mobile devices. According to the different characteristics and nature of the devices involved in the evaluation, the outcome is positive. Several conclusions were gathered from this evaluation:

-Finger-tip based devices are the less preferred by users due to the lack of habituation to make the signature with the fingertip. Nevertheless, results reached with some of these devices show a performance good enough, even in the line of stylus based devices in average. Furthermore this kind of devices is the most common in the market nowadays.

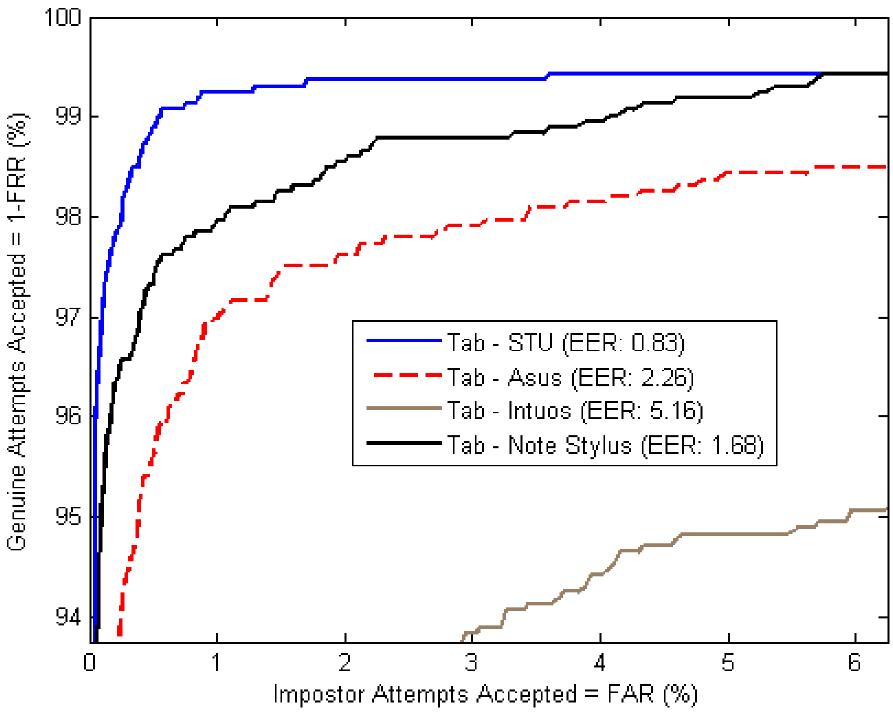

Figure 20. Tab template vs. Stylus based devices

-The results obtained with the fingertip based devices are satisfactory considering that this is one of the first experiments signing this way. It is important to note that the EERs reached with the iPad are within the best ones.

-Receiving a visual feedback is really important for users. Intuos offers worse results than STU (both are the reference platform devices) and the interoperability EERs obtained are higher than the rest of the devices. Furthermore, users felt less comfortable without such visual feedback.

-Under interoperability conditions, some devices present better performance than others when used for enrolling. It has been shown that Note (both with stylus and with fingertip), iPad2 and STU reach the best EERs.

-Most of the EERs obtained in inter-modality experiments are comparable to the intra-modality ones, showing that a single modality can be considered instead of two different modalities.

The good performance results obtained within interoperability environments encourages us to get more involved in improving systems relating biometrics and mobile devices, although there are still drawbacks to solve. Not all the devices employed show acceptable results, and then it is necessary to know where the problem is and solve it. These cases have been highlighted along the experiment. Furthermore, increasing the range of devices used will help us to improve the algorithm throughput and increase the performance under interoperability conditions.

5.1.5 Outcomes to improve the H-B interaction

During this experiment, even when it is not a usability assessment, various factors not included in the H-B interaction have arisen and must be included as improvements:

Factors depending on the system. Users interacted with mobile devices with separable parts (hardware) and received different feedback modes from the interfaces (software). Those factors should be better categorized in the methodology. At the same time, there were users who found the interface buttons very little in some cases. This is an accessibility concern which is not considered into the H-B interaction.

Factors depending on the user-system interaction. There were errors in the app due to the user-system interaction, many of them due to the buttons size. System errors have a clear influence on users and usability, so this must be added to the methodology. Other important factor extracted from this experiment is the use of the fingertip or the stylus for signing. As shown during this experiment, this factor influences the user-biometric system interaction.

5.2 EXPERIMENT 2: Ergonomics in behavioural biometrics

In this second experiment [79] a usability evaluation of biometrics was carried out. As abovementioned, the use of dynamic modalities is needed to better validate the H-B interaction methodology, so that, we continue using DSV. This evaluation is focused on ergonomics, signing the participants in different positions and using different devices (some stylus-based and other fingertip-based). The usability factors and metrics used are the traditional, provided by NIST guidelines [61]. Further conclusions about training and tasks order were also obtained during this experiment.

Motivation of the experiment

Continuing with the use of DSV on mobile environments, we have made a first usability evaluation. The intention of this experiment is to apply the ISO 9241 measurements of usability as starting point. The application of these well-known usability metrics is a novelty in DSV. The main targeted factor is ergonomics, clearly influenced here by the use of biometrics in mobile devices: the variability of new scenarios brings new challenges. We expected to extract significant outcomes from the comparison of usability and performance results.

Influence on the H-B interaction

The validation of several of the suggested improvements to the H-B interaction is made with this experiment: the influence of the user's position when signing as a consequence of using mobile devices, the interfaces who may have repercussions in user errors (deleting signatures, for instance) and the devices order which is modified in each session are the main targets.

5.2.1 Evaluation set-up

In this section, a complete description of the evaluation parameters is provided, including: users, scenarios, devices and guidance of the users. The evaluation process is divided in three sessions at least one week apart for avoiding some undesirable usability consequences such as tiredness and habituation due to the number of signatures per session. There is no time limit for completing the process and users can take the necessary time for making pauses and rest. The number of signatures required per device in each scenario is 12 but it is not required to sign on all the devices in all the scenarios (the distribution of devices and scenarios is shown in Figure 21).

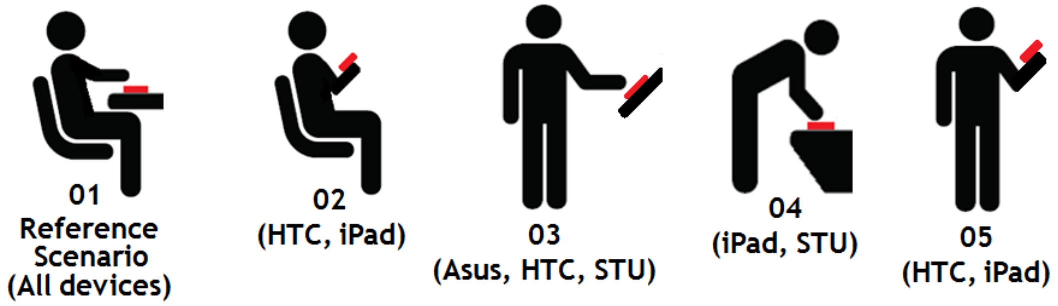

Figure 21. Scenarios and devices used during the evaluation

Users

All the users have to perform 3 steps in the evaluation. First, they have to sign a declaration of acceptance for participating in the process. Then, they have to complete the 3 sessions of signatures and finally, they have to fill a satisfaction form. The participants are not required to have technical formation or any previous skill about signing on mobile devices. For all of them it was the first time signing with the fingertip. That is the desired crew in order to make the evaluation as realistic as possible. Usability factors were obtained from the forms in order to correlate users' parameters with the scenarios, devices and error rates. The number of participants in the evaluation was 20: 7 women and 13 men. The 70% of them are between 18 and 30 years old and almost all are familiarized with smartphones. Only one of them was used to biometric devices (fingerprint scanners) and the rest have used digitizers for signing with stylus (post office or packets delivering). All of them have at least minimum studies: 65% has finished high school and 35% has finished a degree.

Guidance and Training

One of the experiment's purposes is to let the user as free as possible to complete the process. That means that the users' guidance was scarce and provided mostly by the acceptance sheet delivered at the beginning. Only if a systematic error was made, the guide stopped the process and warned the user about it. All the questions related to the signature appearance were answered with "*the signature should be as similar as possible to the one made on a paper*". The intention was to make the process realistic.

Even though most usability guides advice to carry out a previous training [80], there was not any instruction process at the beginning of the evaluation in order to test the users' signatures evolutions. It was considered an important experiment to check out how much effort takes to develop a stable signature in a mobile device, not having previous knowledge or practice. Nevertheless, the software allowed users to delete signatures if they were not happy with them, therefore, a kind of training process was actually performed. For obtaining efficiency rates, the time spent has been accounted.

Scenarios

The evaluation done is a scenario evaluation, so that, for all the users, pattern scenarios were recreated with similar characteristics, modifying the elements slightly in case the user was not comfortable. There were 5 scenarios (showed in Figure 21) representing the most common situations where mobile devices are used. The scenario 01 is considered the reference scenario: the user is sat on a chair and the device rests on a table. Users have to sign on all the devices in this scenario. The scenario 02 represents the user signing sat on a chair and holding the device. The user decides how to hold the device (placed on the knees, held with both hands, etc.). The scenario 03 consists of the user signing on foot and the device is totally resting over a slope surface (users are allowed to modify the angle). The utilization of this scenario is motivated by some entrances and security accesses where biometric traits are presented over a slope surface for ergonomic reasons. In the scenario 04, users sign on foot and place the devices on a table. It is up to the user to put the free hand on the table or not.

Finally, in the scenario 05, the user is on foot and without devices' support. Regarding the scenarios' order for signing, it changes from one session to another but it is constant for all the users (the order is shown in Table 2). It is a critical factor that may influence the final results due to there may be errors involved by changing the posture.

Devices

Four mobile devices were used in the evaluation representing some of the the most common in the market: a smartphone (HTC Desire), a tablet (Apple iPad) and a tablet-PC (Asus EeePC MT101). Moreover, users have to sign on a specialized digitizer (Wacom STU 500) taken as the baseline as in the previous experiment.

The signing order changes from one scenario to another in order to avoid the habituation and not make the same mistakes in the same devices involving bias. Regarding the software, the same program was developed for each device as to keep the application as a constant parameter in the evaluation. Only 3 elements are shown to the user in the screen: a blank space for signing, a button for deleting the signature and another button for accepting the signature.

5.2.2 Evaluation protocol and biometric processing

The performance results are given in terms of EER, which even not offering the complete information about the results, offers a good trade-off between the FAR and FRR, measure enough for comparing the accuracy between the experiments. The parameters used as input data to the algorithm are X and Y time-series coordinates. Although the use of pressure information improves DSV error rates, only the STU and the Asus returns pressure, then this parameter is not used. The three first signatures from each combination user–device-scenario (and not the five first, as usually used [81]) were applied for obtaining the users' template. This decision was made due to the intention of the authors of reproducing realistic scenarios were the user could not be willing to sign a high number of times to get enrolled. The FRR is obtained matching the users' template versus the remaining signatures from the same combination user-device-scenario. To compute the FAR, templates are matched versus all the signatures of all the other users in the same device-scenario.

5.2.3 Experiments

The experiments done are mainly of two kinds. First, error rates are calculated to compare results between sessions, scenarios and devices. Usability rates, such as satisfaction, effectiveness, efficiency and learnability are obtained also. Combinations of both experiments permit to draw a complete usability map.

Table 2. Order of the devices and scenarios

Session 1		Session 2		Session 3	
E1	Asus, STU, iPad, HTC	**E5**	HTC, iPad	**E3**	Asus, STU, HTC
E2	iPad, HTC	**E4**	iPad, STU	**E1**	Asus, STU, iPad, HTC
E3	Asus, STU, HTC	**E3**	HTC, STU, Asus	**E5**	iPad, HTC
E4	STU, iPad	**E2**	HTC, iPad	**E4**	iPad, STU
E5	iPad, HTC	**E1**	HTC, iPad, STU, Asus	**E2**	HTC, iPad

<u>Performance-related experiments</u>

As no comparisons were done until the end of the evaluation, these experiments were made once all the signatures had been gathered. There are several parameters for comparing the performance of a biometric system, though in this case the parameters used are the ones related to usability. According to the EERs obtained, a comparison between the devices' performance can be done. This result is not totally definitive as in this case it is influenced by some other factors such as device order and scenarios. In order to provide these results, all the signatures gathered are used, enrolling with the first 3 signatures obtained in session 1.

<u>Ergonomics</u>

Various experiments were done in order to test which scenario-device combination fits better and in comfort for users. These tests are highly important for representing users' preferences and point out the way for developing usable systems.

- Devices and scenarios. The differences between the scenarios are basically based on the position that users have to acquire. The scenarios were designed for representing the most common postures where users are expecting to sign: sitting on a chair, on foot, holding the mobile in the hands or with it placed on a table (as showed in Figure 21).
- Device and Scenario order. Getting used to sign in mobile devices and using the fingertip is not trivial and takes time, so that, users do not perform the best

signatures at the beginning. Then it is remarkable that error rates in the first devices are thought to be higher a priori due to the lack of train. This points out the importance of the order, not only with devices but with scenarios also.

Efficiency, Effectiveness, Learn-ability and Satisfaction

Most of the experiments carried out in this evaluation were made to predict and improve usability in future system developments. In this section the experiments of the evolution through the 3 sessions are shown.

- Efficiency. For computing the efficiency the time spent per signature was calculated.
- Effectiveness. The number of times that users sent a wrong signature (signing out of bounds or in the air, repeating strokes, etc.) was measured also and represented as the FTA.
- Learnability. Worst EERs are expected to be obtained during the first session and are also expected to be improved during the other sessions, because of the habituation. It is interesting also to compare the error rates obtained in the second session and in the third, because the difference should be smaller than between the 1st and the 2nd session (due to habituation also).
- Satisfaction. This is a subjective parameter reached from the satisfaction forms. Users were asked about 5 usability factors: comfort, time, easiness, intrusion and the global experience.

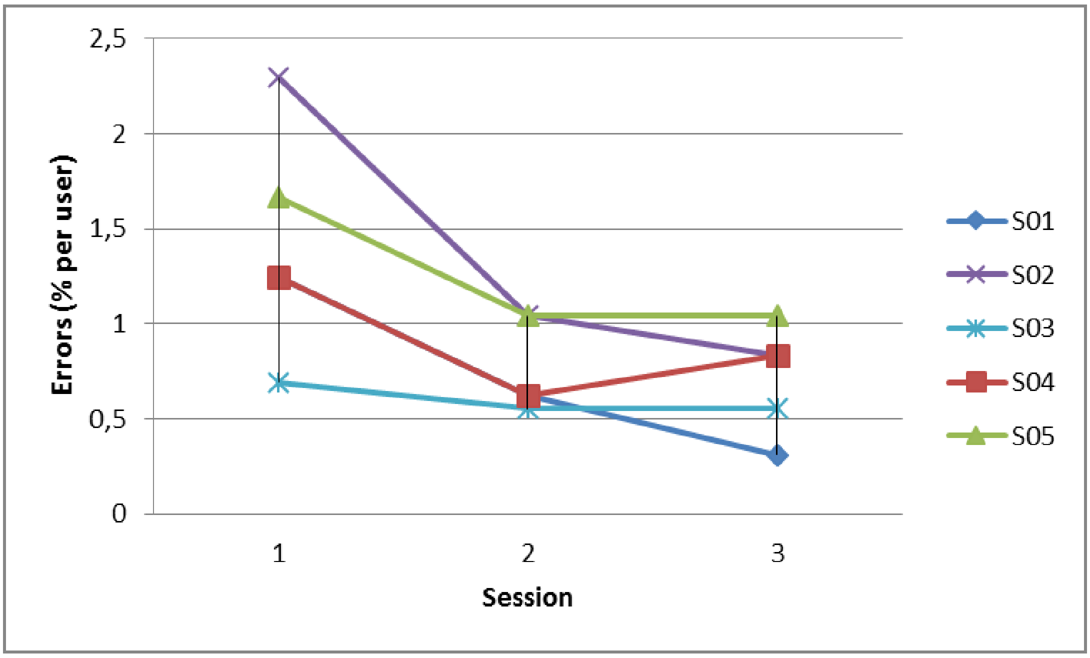

Figure 22. Effectiveness (in % per user), divided in sessions and scenarios

5.2.4 Results

Performance results are shown in Table 3 divided in scenarios and devices and in Table 4 divided also in sessions in terms of EERs. The learnability is obtained from the performance, the effectiveness (Figure 22) and the efficiency (Figure 23) as the time spent. Finally, the satisfaction was measured through the forms filled by users. They were asked to score (from 0 to 5) some evaluation features: comfort, time, easiness, intrusion and the global evaluation (Figure 24).

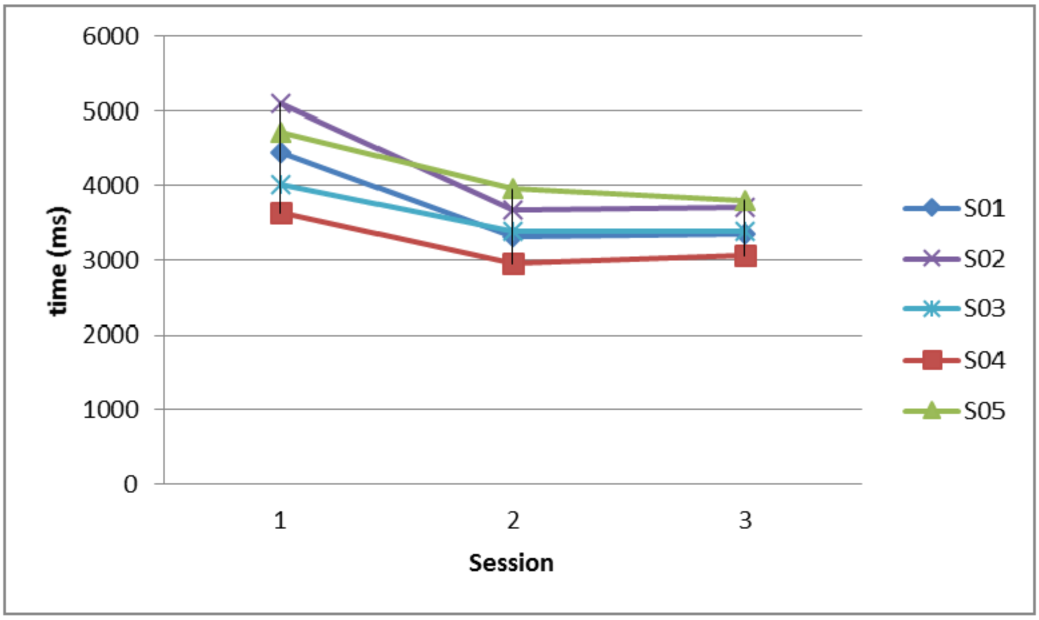

Figure 23. Efficiency (in time -ms-), divided in sessions and scenarios

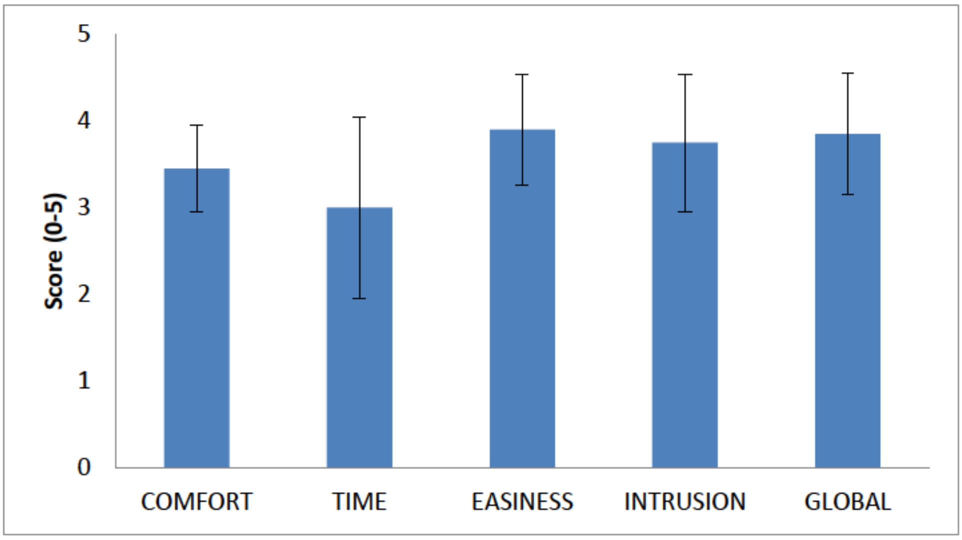

Figure 24. Satisfaction Results in terms of comfort, time spent, easiness, intrusion and global experience. Black sections are the variances

Table 3. Global performance results (in terms of EER) by scenario and device

	S01	S02	S03	S04	S05
STU	1.06%	-	1.97%	2.43%	-
ASUS	1.97%	-	3.18%	-	-
IPAD	**7.27%**	0.49%	-	0.92%	0.6%
HTC	3.63%	2.89%	3.01%	-	1.36%

Table 4. Performance results by session scenario and device (in terms of %EER)

	Session 1					Session 2					Session 3				
	S01	*S02*	*S03*	*S04*	*S05*	*S01*	*S02*	*S03*	*S04*	*S05*	*S01*	*S02*	*S03*	*S04*	*S05*
STU	0.56	-	1.67	2.2	-	0.55	-	**4e-4**	0.43	-	0.01	-	**1.11**	7e-3	-
ASUS	**1.67**	-	2.78	-	-	**0.04**	-	2.65	-	-	0.56	-	1.66	-	-
IPAD	**3.88**	0.56	-	0.55	0.47	**1.2e-4**	4e-4	-	0.44	1.11	0.56	0.56	-	0.55	**0.01**
HTC	2.29	1.66	2.22	-	0.99	0.56	0.55	1.11	-	1.11	1.11	1.12	0.56	-	1.06

5.2.5 Conclusions obtained for future improvements

Performance

According to the results (Table 3 and Table 4), the STU returns the most constant performance, as expected (it is the reference device). Nevertheless, the best result is reached by the iPad in the scenario 2. At the same time, the iPad has obtained an EER over 7% in the reference scenario. It demonstrates the high importance of the order of the scenarios-devices in performance.

Ergonomics

The results are better in some scenarios than in others using the same device and in some cases the differences are really noticeable and influenced also by the order and other factors that are summarized below.

- STU: As a device do not thought for mobility, the best results are reached in the scenario 01. Due to a habituation effect, the results improve gradually from one session to another but in the third session an EER increase occurs due to users signed first in this device-scenario combination.

- ASUS: The same conclusions are extracted for this device, the results improve gradually and it is more noticeable when users do not sign first in this device (for instance, in the session two the EER decreases more than in the third). The best performance takes also place in the scenario 01.
- iPad: The EER varies highly according to the order. As it is shown, the scenario 01 was the first chance for users for signing with the finger and it involves really bad results. A good performance is obtained in the scenarios where users have not support for the device.
- HTC: A big improvement can be seen from the first session to the second, though the order did not affect the results as much as in other devices.

At first, users found strange to sign with the fingertip and as a consequence the results with the iPad and the HTC are worst in session 1. A better conclusion of the best relation device-scenario can be obtained from Table 3 because the effects of the order are less noticeable. According to the results the lower EERs are obtained in the scenarios without devices support and not in the reference scenario as it could be thought.

Traditional usability metrics

There is a clear relation between the efficiency and effectiveness: a big decrease (error rates and time) is observed in all the scenarios between the session 1 and 2. Nevertheless, it is not as noticeable between session 2 and 3. This effect means that a training process is highly necessary in order to achieve better results. Once the users have habituated to the system these usability factors become constants and the EER improves also as demonstrated.

Users learned to sign in session 2 in average. They scored the global of the evaluation (the conjunction of the usability factors) with a 3.85 over 5, being the preferred factor the easiness (3.9) and the less favourite the time (3). This could happen due to the session's duration and the number of signatures to complete. Other realistic approaches (where users are required to sign less times) may improve their opinion regarding timing. 80% of the users would use handwritten signature recognition on mobiles again in the future (one of them only with stylus). The preferred devices were the STU with stylus and the iPad with the fingertip. Furthermore, the favourite scenarios were (from best to worst): S01 > S03 > S04 > S02 > S05. This shows the preference for the most common scenarios and at the same time, is noticeable that the scenarios that do not provide support return sometimes the best results even being the less preferred.

5.2.6 Outcomes to improve the H-B interaction

Outcomes obtained after this experiment suggest the following improvements to the H-B interaction:

Factors depending on the user. Emotions have clear influence during this experiment. When interacting with the devices users proceed different according to their moods (e.g. if users are in a hurry, they tend to sign faster and if they had a bad day, they tend to sign slowly). Pauses, rests and the time span among sessions involved active and collaborative users.

Factors depending on the system. In this case, the devices' location shows a clear influence in usability and performance. Therefore, the location of the system components should be included in the H-B interaction within the ergonomics. Interfaces have a significant repercussion on the interaction. In this case, the simplicity of the screens (two buttons and a blank space), as well as the interface consistence among the different platforms helped to smooth the processes.

During these experiments, users interacted with non-specific biometric devices. According to the differences obtained (in performance and usability) between signing on a specific digitizer and on a tablet or smartphone, this factor is relevant when analysing biometric systems. Also in this case, the devices used have separable parts (styluses), which must be included to in the H-B interaction

Factors depending on the user-system interaction. Users are not equally motivated in all the experiment phases. The enrolment, for instance, is shorter and the user is required to perform less actions. Then, users proceed more fluently than in the rest of the phases. This factor must be reflected in the H-B interaction.

H-B interaction metrics. More than one session involves that users may learn how to use the system during the first session and may remember it in the following. These well-known usability metrics (learnability and memorability) must be included in the H-B interaction.

5.3 EXPERIMENT 3: Ergonomics in behavioural biometrics including styluses

In this third experiment [82] we have also evaluated ergonomics in biometric systems. The biometric modality was DSV too, but in this case signing on an iPad. Styluses are ergonomic elements convenient for signing on electronic devices (as shown in the previous experiment). The procedure is similar to the previous experiment regarding the measurements and we have also took into account user's positions when signing. Moreover, in this work we have categorized the possible errors when signing as an effectiveness measurement.

Motivation of the experiment

Previous experiments have highlighted the user's preferences in mobile devices and the iPad was pointed as very comfortable for signing, even when it was used with the fingertip, which is something new for users. At the same time, the iPad has a convenient size for real life tasks including biometrics as reading and signing contracts. We believed the inclusion of the stylus for signing would benefit both performance and user's satisfaction: people is used to sign with a pen on paper. We took the opportunity of testing different styluses as the iPad does not have a proprietary one. The scenarios were those considered as the most comfortable by users when signing on an iPad. Again, the measurements used were the well-known usability metrics.

Influence on the H-B interaction

To carry out this experiment, we have gathered previous user's comments and opinions, designing as a result the iPad-stylus scenario. This way to proceed is in accordance with the current usability evaluations methodologies, which encourage designers to receive user's feedback from the early stages of the design. Then, traditional usability testing was also applied. Some other factors for updating the H-B interaction and studied in this experiment (e.g. ergonomics, devices order, etc.), has been added in the previous experiment.

5.3.1 Evaluation set-up

The evaluation was split into 3 sessions with at least 1 week of separation between each one in order to avoid tiredness: users will not be eager to complete the evaluation in three days in a row and signatures would be done apathetically.

Users

There were 21 participants in the experiment, 13 men and 8 women. They were chosen randomly but representing all the main different age groups. Almost half of them have used electronic digitizing tablets to sign before but for all of them it was the first time they signed in a mobile device. There was not any special condition for joining the evaluation. An explanation of the evaluation was made before starting and users filled a satisfaction form at the end of the third session. This form included questions about:

- *Preferences of scenarios and devices.*
- *Comfort, time-spent, easiness, intrusion and global satisfaction.*
- *Familiarity with biometrics and previous experiences.*
- *Diseases that could modify the signature (i.e. Parkinson or Sclerosis).*
- *Age, gender, laterality and level of studies.*
- *The received instructions during the evaluation*
- *Possibility of using DSV in the future.*

Figure 25. Evaluation scenarios.

Scenarios

Three different scenarios were completed by users in the evaluation as it is shown in Figure 25. These scenarios were chosen in order to represent the most common situations where users may use this technology. The scenario A, which is taken as the reference scenario, consists of the user signing sat on a chair with the device resting on a table. In scenario B the user is sat on a chair handling the device with the only support of her/his hands. The user herself/himself decides how to grab the mobile device. In scenario C the user signs standing up and with the device resting over a lectern. The height of the scenarios' elements was up to users' preference and they could modify it whenever they wanted to feel comfortable. The scenarios order was changed from one session to another, being in the 1st session A-B-C, in the 2nd session C-B-A and in the 3rd session B-C-A. This was done to minimize the effect of order in performance results as well as in users' habituation.

Device and styluses

The device used is an Apple and three styluses were used in the evaluation in all the scenarios. The differences between them are their tips' shape, length and diameter. A colour code is used to identify them (black, white and pink). An illustration of the styluses can be seen in Figure 26. The pink stylus is made of plastic, its length is 10.7 cm and its diameter is 6 mm. Its' tip is made of sponge and ends in a plane shape. The white stylus is made of plastic also. It has 11 cm of length and 8 mm of diameter. Its tip is composed by soft rubber and has rounded form. Both styluses (pink and white) have a deformable tip. Finally, the black stylus is made of steel. It has 12.6 cm of length and 8 mm of diameter. Its tip is made of hard plastic and steel, so it is not deformable but its position changes when it contacts with the device and therefore it is always plane to the touched surface. The 3 styluses are conductive and made to write over capacitive devices. The order in using the different styluses was changed in each session, being the order in the 1st session pink-white-black, in the 2nd session black-pink-white and in the 3rd session white-pink-black.

Figure 26. Styluses utilized in the evaluation. From left to right: Black, White and Pink.

Guidance and training

There was not any training process scheduled before the evaluation and users started the experiment without previous knowledge. This was designed to test how the users' signature changed through sessions and time (habituation) although users were previously told about the evaluation structure and proceedings. Also, users had the possibility to delete the signature made as many times as desired; so this can be considered as a self-training process. An operator responsible for the evaluation offered support to users if needed, though no instruction about how to sign was given.

5.3.2 Experiments

It is important to differentiate the two parts of the evaluation: data collection and its posterior processing. Thus, all the results gathered are calculated in the processing step. In this section all the usability and performance experiments are shown.

Usability

- Effectiveness: The effectiveness error is the ratio of wrong signatures divided by the total. If users are not able to complete the whole process they feel frustrated and this influences directly in the satisfaction and has a bad impact over the signature quality. Accordingly, when users deliver low quality signatures (bad task completion) it has a direct effect on the algorithm performance. A signature is considered as wrong in the following cases:

  -The signature is totally or partially out of the devoted space for signing.

  -User has to repeat at least one stroke (e.g. due to latency problems).

  -Part of the signature is made without touching the device (signing in the air).

  -The accept button is touched being the signature incomplete or not desired.

  -User' wrist touches the screen.

- Efficiency: The efficiency error rate is calculated by the average time employed to complete the signatures.
- Satisfaction: It was measured through the satisfaction forms, the operator notes and the video recorded. Furthermore, both the efficiency and the effectiveness influence this parameter as abovementioned.

Apart from these three factors, learnability was measured due to users' habituation to signal acquisition through the sessions. Thus, time spent and errors made are expected to be decreasing from one session to another as the user acquires skills signing in these scenarios.

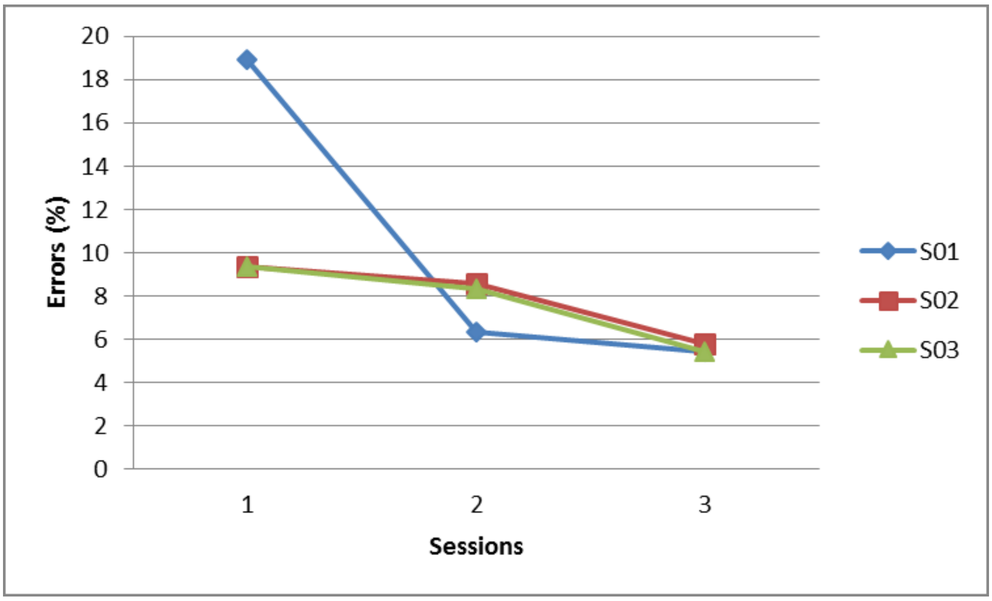

Figure 27. Effectiveness rates obtained by session and scenario

Performance

Performance error rates were calculated for each scenario and stylus. Through these error rates it is possible to measure habituation effects and the influence of scenarios and styluses' order in users. To obtain performance results a previous template to compare with (enrolment) is necessary. Once the enrolment is done it is feasible to carry out the verification process.

The error rates obtained from the algorithm were the FAR, the FRR and the EER. The comparisons are offline (done after all the acquisitions have been finished) and 3 signatures gathered from each user were taken for obtaining the templates. Thus, FRR is calculated comparing each template with the rest of the users' genuine signatures from the same user-device-scenario combination. FAR is obtained by comparing each template with all the signatures from the same combination device-scenario containing all the other users.

5.3.3 Results and conclusions

Usability

In Figure 27 are the effectiveness rates divided in sessions and scenarios. In Figure 28 time spent averages are provided divided in styluses and sessions. The satisfaction factors obtained from surveys (comfort, time spent, easiness, intrusion and global experience) were scored from 0 to 5 (Figure 29). The global experience score is 4.28 and the average of all factors is 3.87.

The effectiveness rates show a big decrease of the errors from 1st session to 3rd session, especially in the scenario A. It occurs due to the scenario A is the first to be completed and users had no previous experience signing in mobile devices. By the end of the evaluation this rate becomes stabilized. This error variation could be solved by completing a previous training process. The time average in the scenarios shows a high dependence on the order. Thus, in 1st session (A-B-C) and in 2nd session (C-B-A) the first scenarios took more time. In 3rd session the time becomes almost equal for the three scenarios, showing that users have not problems to sign in any scenario in particular.

Regarding the styluses, the black is the faster followed by the pink and the white. Users considered the evaluation globally good and non-intrusive. In fact, 14 of them would use this kind of system habitually. The worst satisfaction parameter was the time spent. Nevertheless in a real scenario this should not be a problem because the number of signatures required would be much lower. The order of preferred scenarios is A > B > C and black>pink>white for styluses. Users felt comfortable singing on the iPad and considered that the device response was good enough. For example, one of the users said "*the visual feedback is not very different to the one perceived with pen and paper*". Regarding the learnability, users acquired part of the skills in the 2nd session and the rest in the 3rd session: according to effectiveness and efficiency, users acquired habituation in the 2nd session (in Figure 27 these rates become stable from the 2nd session). Though, according to the time spent (Figure 28) it is descendant until the end of the evaluation, so that, users are still gaining ability signing. Analysing these results it is feasible to conclude that participants get used to the system in the 2nd session, once efficiency and effectiveness become steady (they suit better the space to sign and commit fewer errors). The time spent, always decreasing, shows that users do not stop improving their habituation to the signing process.

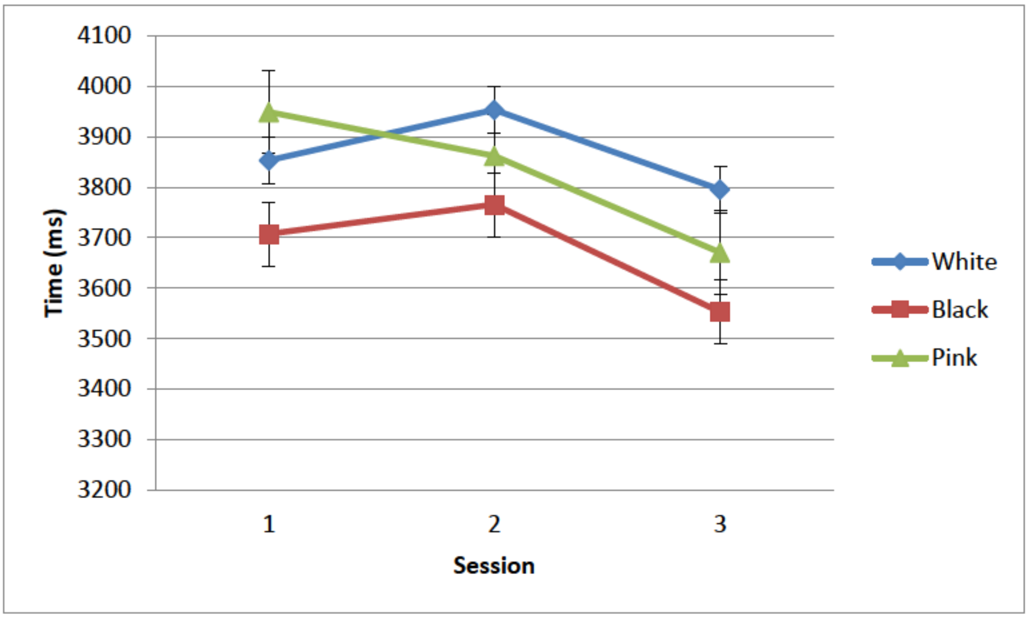

Figure 28. Average time spent as an efficiency measure by session and stylus. The segments are the variances

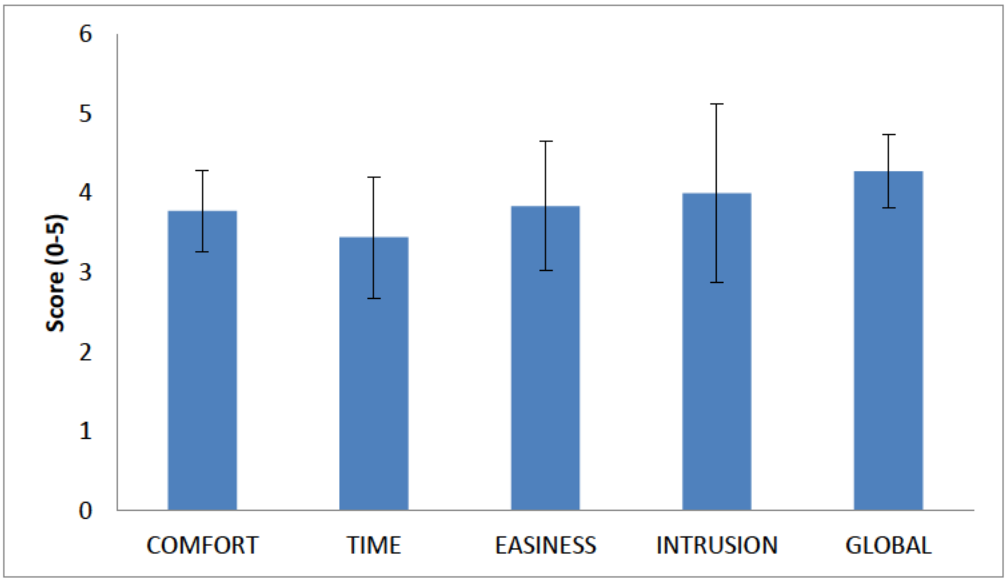

Figure 29. Satisfaction rates regarding comfort, time, easiness, intrusion and global opinion. The segments are the variances

Performance

In Figure 30 a comparison between the EERs obtained from the 3 styluses (within the legend box) is done along with the ROC curves, taking all the signatures gathered (from all the scenarios). In Table 5 the EERs obtained with the different styluses in the different scenarios are revealed. Previous results obtained signing with the fingertip are included also. It is noticeable the EER=3.96% of the pink stylus in the scenario A. This high error rate is due to be the first stylus-scenario combination in the evaluation. As in the usability results, the lack of previous training is remarkable and the order seems to be highly important also. For the 3 devices the best performance is obtained in scenario C which is the least preferred by users. This result can be due to users had not to handle the iPad with the hands (it was placed in a lectern).

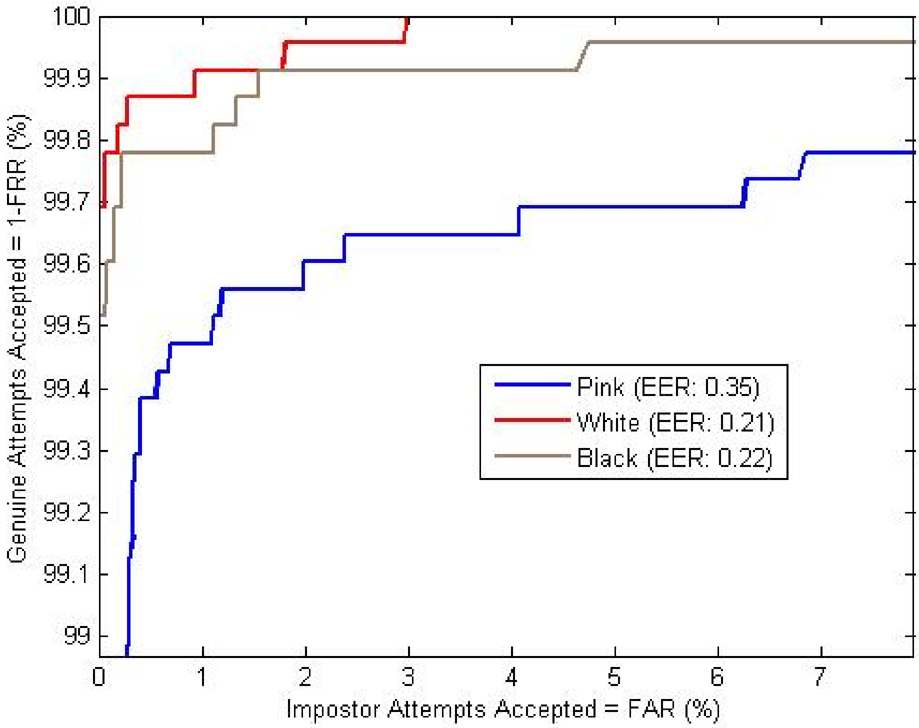

Figure 30. ROC curves of styluse's performance (in terms of % EER)

The fact of having the device resting on a surface avoids bad effects of pressure changes produced when is the user who handles the device. This shows also that a small tilt of the surface makes the signing process easier (results are better in scenario C than in scenario A). In order to test which stylus returns better performance, the templates for calculating the ROC curve were chosen randomly (minimizing the effect of order). In Figure 30 it is shown that the EER of the white stylus is the best (0.21%), followed closely by the black stylus (0.22%) being the pink the worst one (0.35%). This is almost in accordance with users preferences. In addition these results improve previous results obtained (EER = 1.8%). This improvement can be due to improvements in the capture software or iPad features such as screen quality or processor (different from digitizing tablets' features used to design the algorithm).

Table 5. Styluses and fingertip EER divided by scenarios

Scenarios	Pink	White	Black	Fingertip
A	3.96%	0.26%	0.78%	7.27%
B	0.52%	0.93%	0.52%	0.49%
C	0.26%	0.13%	0.13%	-

Table 6. Usability comparison between stylus and fingertip

	Effectiveness errors	Satisfaction
Styluses	0.72%	4.3/5
Fingertip	1.49%	3.9/5

Fingertip versus stylus

One interesting point in this research is the comparison between signing with the fingertip and signing with a stylus. Results regarding signing on an iPad with the fingertip are extracted from previous works [79]. The performance and usability results are given in Table 5 and Table 6 (the iPad was not used with the fingertip in the scenario C). Regarding the performance it is noticeable that signing with the fingertip is less straightforward for users than with the stylus (7.27% fingertip and 3.96% stylus) and the signatures have different appearance, in particular at the beginning. This is due to the lack of habituation of users (none of them had ever signed with the fingertip). Once the user is habituated to the system both results become similar (2nd session) involving similar performance in both approaches. The number of effectiveness errors is bigger with the fingertip due, in most of the cases, to the different position in which users sign with the finger (the gesture is very different for some of them). Satisfaction rates (4.3/5 stylus and 3.9/5 fingertip in average) and users' opinions reveal the general preference for stylus based systems as expected. In short, both sub-modalities (fingertip and stylus) offer similar results in performance and styluses are better rated in usability. Nevertheless, as long as handling mobile devices is more common with the finger only, getting used to sign with the fingertip requires some previous training.

5.3.4 Outcomes to improve the H-B interaction

Outcomes obtained after this experiment suggest the following improvements to the H-B interaction:

Applying the common usability evaluations prior to final designs. Previous interviews, notes and video recordings drove us to the final design. The application of these methods was successful as the user's satisfaction was high in terms of the surveys and in offline conversations.

In this case, usability testing in low prototypes, showed the convenience of using certain mobile devices in certain scenarios (according to user's preferences). Prior usability evaluations are not in the H-B interaction.

Factors depending on the user-system interaction. The evaluation's phase is relevant for the interaction. Users do no performed signatures equally in the first session than in the others (both in usability and performance as shown in final results). Being conscious of the current phase of the evaluation influences user's eagerness to proceed.

5.4 EXPERIMENT 4: Analysing moods in behavioural biometrics

The fourth experiment deals with moods in behavioural biometrics (DSV). In this case, the HBSI method is applied to analyse usability, as this model has not yet been tested empirically in dynamic modalities. The errors assessment in the interaction have been made by users without video recording. Furthermore, this study includes stress tests where users sign under pressure conditions. The inclusion of these tests in the evaluation is motivated by some common scenarios where users are indirectly encouraged to sign quickly and carelessly (e.g. post offices, banks or supermarkets).

Motivation of the experiment

This experiment has three main motivations. First, our intention is to measure the influence of stress in the biometric recognition process, as this is one of the major concerns regarding usability and performance. We suggest this is especially true in dynamic modalities, where the interaction is longer. Second, we wanted to carry out a validation of the HBSI for DSV in mobile scenarios and test if the model has a proper coverage of this kind of interactions. A proper HBSI evaluation involves the video recording of the whole evaluation, therefore the required time (of several people) to process all those videos is considerably long. Our third intention with this paper is the introduction of an automatic procedure to process the interaction avoiding to spend time and resources.

Influence on the H-B interaction

The H-B interaction incorporates the HBSI metrics and in this experiment we suggest modifications to those metrics. Moreover, the study of the stress reveals weather this factor is influential and must be included in the methodology. Further outcomes of the user-system interaction were expected from this experiment as the user is who indicates the system the reasons of the interaction errors (in the case of errors).

There are several interfaces in this evaluation, increasing the possibility of errors during the interaction (apart from the biometric interaction). The inclusion of a proper interface analysis in the H-B interaction is convenient.

5.4.1 Evaluation set-up

This evaluation set-up is in accordance with the conclusions obtained in previous works and the necessity to accomplish users' requirements in order to develop usable systems.

This experiment was divided into 2 sessions one week apart. Training, enrolment and Verification 1 (V1) were done during session 1 (with pauses between them). Verification 2 (V2) and Stress-Influence Tests (SIT) were done in session 2 (also including pauses between them). Users could pause at any time to rest, except during the stress-influence testing.

Users

The evaluation crew was composed of 56 users (37 men and 19 women) chosen without any special requirement. The only condition for taking part in the evaluation was age (over 16 years old as it was demonstrated that the handwritten signature is not stable in children [83]). There were 54 right-handed and 2 left-handed. Most of them (41) were between 18 and 35 years old, 8 were between 35 and 50, and 7 were older than 50. Regarding the level of studies, only one user did not have a minimum qualification. The rest of them had primary education studies (2), secondary (2), high school (13) and university degrees (38). Several (29) did not have previous experience of signing on mobile devices. Almost all the users completed the evaluation successfully (54).

Hardware

One of the best sellers in the smartphones market, the Samsung Galaxy Note, was chosen for this experiment. One of the reasons for using this smartphone is its proprietary stylus used in the signing evaluation. It also covers the initial requirements fixed by previous works: capacitive screen, stylus and considered comfortable by users.

User guidance and training

None of the participants received any previous information from the operator before starting the experiment. Nevertheless, the application offered guidance during the process. At the beginning, a video explaining the whole process was shown. In addition, reminders (e.g. text messages) were shown at all the stages. The evaluation was supposed to be completed without an operator, so users were encouraged not to ask for help unless they could not continue with the evaluation.

Before starting the enrolment, a training process was to be completed by users. The application required at least one accepted signature to move forward to the next stage, although the user could stay in training as long as desired. During training, all deleted and user-accepted signatures were accounted for to measure the training that each of the users needed.

5.4.2 Biometric processing

The algorithm applied was translated into Java for Android and embedded in the smartphone. It was used in the mobiles' application to calculate similarities between the signatures as feedback for users just after performing a signature and pressing the "Accept" button. The use of the algorithm differs from the enrolment in the verification.

During enrolment, the algorithm compares each signature with all of those previously acquired, starting with the second one (2nd vs 1st; 3rd vs 2nd and 1st; 4th vs 3rd, 2nd and 1st; and 5th vs 4th, 3rd, 2nd and 1st) and returns the best result. This verification is performed in order to check if the user is making a different signature or if the acquisition process has not captured a significant number of sample points. As a result, the application shows a red or green square above the signing space as feedback to users. The feedback square becomes red if the similarity score is under the threshold 3; otherwise it becomes green. If the provided signature returns a similarity score below 3, users are required to repeat the process. If after the third attempt to acquire a signature for the enrolment, the similarity score is not at least 3, the enrolment ends and the user is not allowed to complete the evaluation, increasing the rate of Failure to Enrol (FTE).

After enrolment, the verification stages occur. In these phases, the algorithm compares each signature with the five signatures obtained through the enrolment and returns the best result. In this case, if the signature is under the threshold, the error is stored and the user shall not repeat the attempt, but continues with the next one.

App characteristics

The application guides the user through the different menus intuitively, so the user is supposed to complete the experiment without additional help. As the application has been executed in Spain, all guidance is written in Spanish, and therefore the captured screens are in that language. These screens fit with the evaluation roadmap pictured in Figure 31. All the signatures and data gathered were locally stored in the mobile device until the end of the evaluation. Once the experiment was finished they were transferred to a PC in order to obtain the final performance and usability rates. As in a previous approach to a real application, critical concerns for a final implementation such as security have been untreated because the main target is the usability analysis.

Session 1

As previously mentioned, the first session starts with a video explaining the whole process in detail. Then, users have to accept the evaluation conditions, covering National Data Protection Law requirements, as well as a participation agreement. After this, the user introduces his personal data, including name, surname, age range, profession, gender and laterality. A personal number is generated for each user, keeping personal data unlinked to signatures.

The next step is the training process where the user can practise until he feels ready to start the evaluation. The training screen has 3 buttons (accept, delete and continue) and a blank space to sign. All the deleted and accepted signatures are counted at this point. Once the user feels comfortable with the system, the enrolment starts and he is required to provide 5 signatures correctly (as mentioned above). Next, V1 starts and 12 signatures are requested. The verification screen is similar to the enrolment one: 2 buttons (accept and delete) and a blank space for signing. A screenshot is shown in Figure 32. This is the end of session 1.

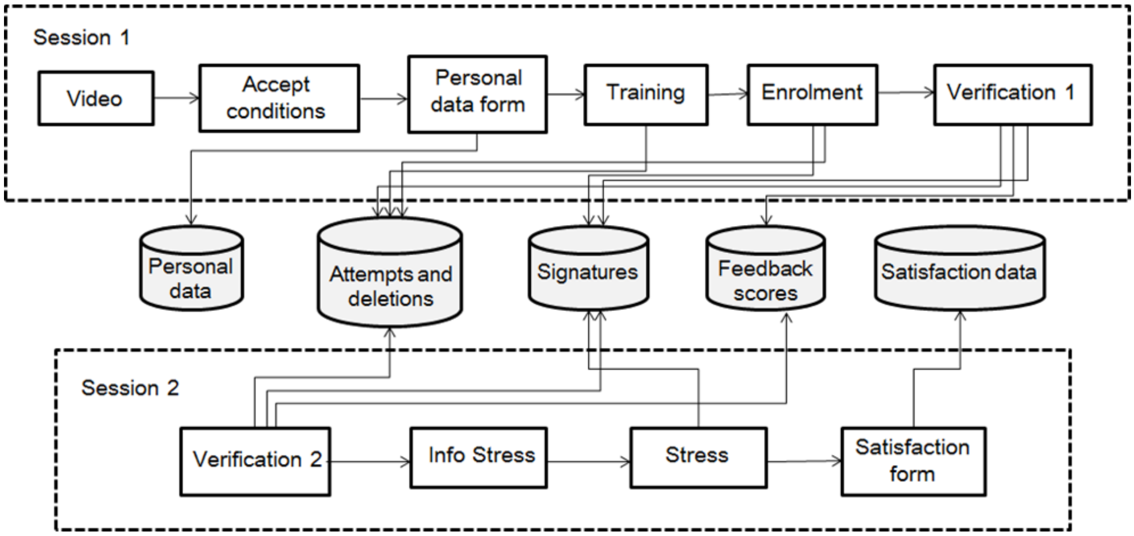

Figure 31. Evaluation roadmap

Session 2

This session starts with the V2 phase, which follows the same process as the V1. Once the user completes 12 signatures again, the SIT starts, by showing a new screen with the message "*the interface will change slightly, please keep signing normally like up to now*". Then, the interface starts blinking, changing between yellow and red. At the same time, an intermittent annoying sound is played loudly and a countdown from 5 to 0 starts. When the countdown reaches 0 a text message appears saying "*you are so slow, please go faster*". Loud sounds [84], blinking images [85], encouraging text messages and countdowns are intended to provoke stress in the users. At this stage, no feedback about the signature similarity is provided and only when the user completes 12 signatures does the stress test finish. When all the signatures are completed, a satisfaction questionnaire is provided to the user to express his opinion about several aspects of the test and the evaluation ends.

Signature decision

A user can finalize each signing attempt at any time either in V1 or in V2. The signing process finishes by accepting or deleting the signature performed. Once the delete button is pressed, a pop-up window with the following options is shown:

-Why did you delete the signature?

- *I did not like it*
- *I repeated strokes*
- *I made it partially in the air*
- *I placed my wrist on the screen*

If the user presses the accept button, the algorithm works as explained before. Furthermore, if the similarity score obtained is 1 or less (in the 0–5 scale), a new pop-up window appears, asking the user whether the signature is correct or not. By pressing yes, the user continues with the process and pressing no means repeating the signature. These steps were made in order to categorise both errors and deleted signatures into the HBSI metrics.

The intention of this work was to save resources during evaluation, such as video recording and data processing, making the system automatic. Recording the whole evaluation on video is a method used in the HBSI to better understand the user–system interactions. In such a case, at the end of the evaluation, operators have to review the videos carefully to categorise interaction errors. This process takes a very long time and requires several personnel to minimize categorisation errors during video replaying. In this work, it is the user who decides whether a signature is correct or not (deleting or accepting the signature), and also the reason for deleting. Therefore, the categorisation is automatic. This decision process proposed by the HBSI was modified and it is shown in Figure 33.

5.4.3 Experiments

The experiment involves measuring several usability parameters. This evaluation returns 3 kinds of outcomes: HBSI rates, stress-influence and system performance. As these outcomes are quite inter-related, this analysis is done separately first and the correlations found are then analysed in the conclusions section.

HBSI - Usability

The usability analysis included in HBSI considers the three main parameters proposed by the ISO 9241, satisfaction, efficiency and effectiveness:

Satisfaction: This parameter is measured and studied through the satisfaction questionnaires. The questions concerning satisfaction included in the experiment questionnaire are:

- *Would you use this system in your daily life?*
- *Do you consider the received instructions enough?*
- *Time spent. Score 0-5 (0-very annoying to 5-very satisfactory)*
- *Easiness. Score 0-5 (0-very annoying to 5-very satisfactory)*
- *Privacy. Score 0-5 (0-very annoying to 5-very satisfactory)*
- *Global opinion of the evaluation. Score 0-5 (0-very annoying to 5-very satisfactory)*
- *Intrusion (how intrusive the application is). Score 0-5 (0-very annoying to 5-very satisfactory)*

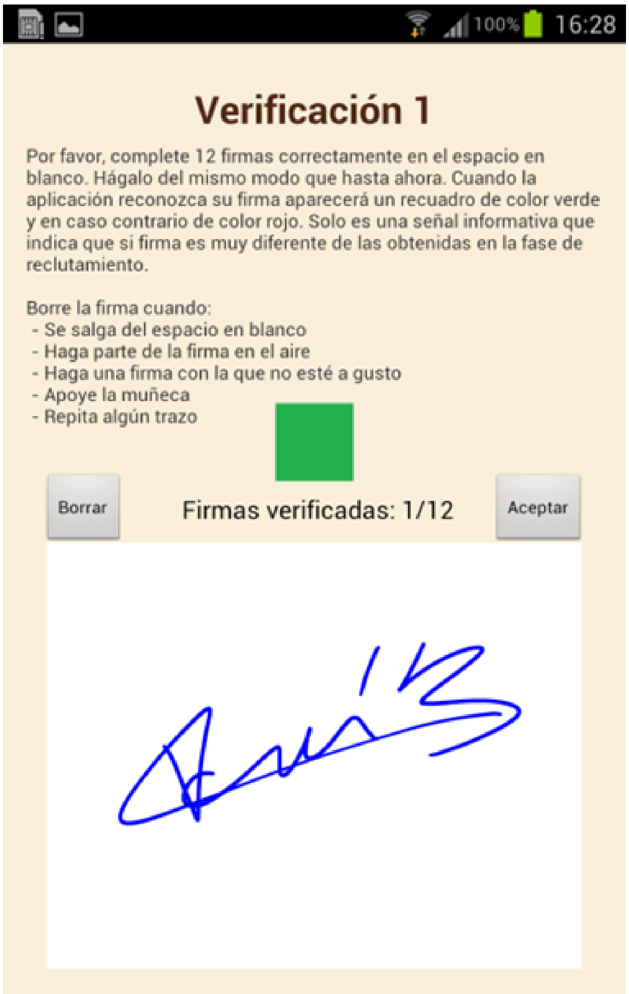

Figure 32. V1 screen capture in the application. The signature shown is not part of the database captured

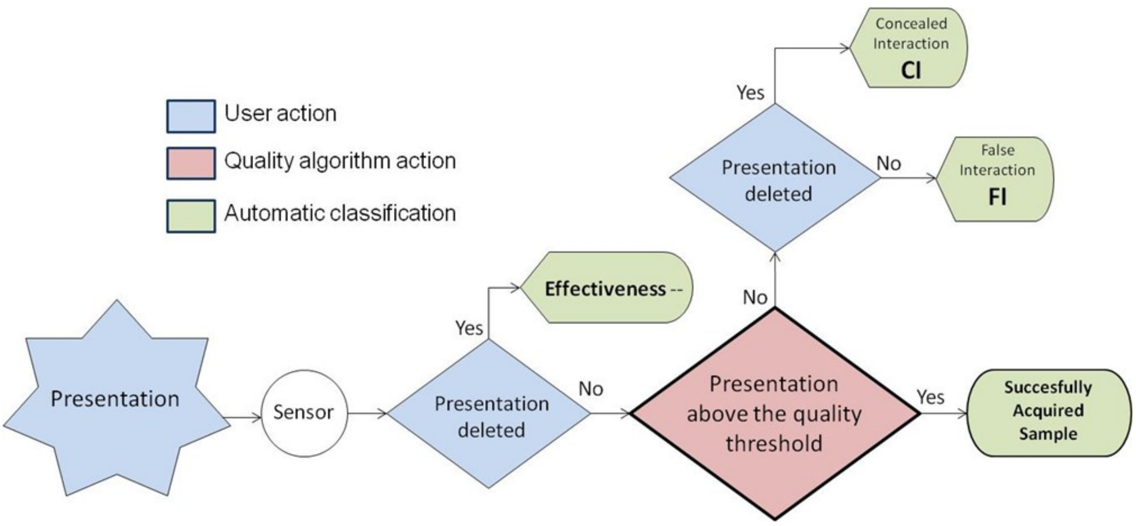

Figure 33. Modified HBSI metrics decision flow chart

Efficiency: The time spent on performing tasks. When a user deletes a signature, it also decreases the efficiency as the time performing that task is increased. For that reason, in this particular case the efficiency is calculated with two parameters: the time spent signing and the rate of non-deletions:

$$Non\ Deletions\ Rate = \left(1 - \frac{Number\ of\ deleted\ signatures}{Total\ number\ of\ signatures\ x\ Users}\right)\ x\ 100$$

This involves a change to the HBSI proposal, as the possibility of deleting an acquired sample is not considered in HBSI.

Effectiveness: Defined as the task completion rate by users.

$$Task\ Completion = \left(\frac{Number\ of\ users\ able\ to\ complete\ the\ process}{Total\ number\ of\ Users}\right)\ x\ 100$$

These three measurements are inter-related as the decrease of any of them usually involves the decrease of the rest of them (i.e. when a user deletes a signature several times the efficiency decreases and at the same time the user feels annoyance so the satisfaction decreases too).

<u>HBSI - Ergonomics</u>

Ergonomics in the HBSI includes cognitive and physical categories.

- Cognitive: Defined as the percentage of users that

  - Know how to use the capture sensor. Obtained in session 1
  - Learn how to use the capture sensor (also known as learnability). This is obtained in session 2 once the users have acquired skills previously in session 1
  - Remember how to use the capture sensor. This parameter is also observed in session 2.

- Physical: The percentage of users that can use the capture sensor.

<u>HBSI - Signal Processing</u>

This measurement includes sample quality metrics and processing capability (number of segmentations and feature extraction errors). In the handwritten signature recognition state-of-the-art there is no method to measure the sample quality, so this parameter was not measured in this evaluation. However, as aforementioned, a similarity metric between signatures was given to users as feedback by the mobile application. This work is therefore a novelty in the sense of returning a similarity metric of the handwritten signature.

HBSI - Interaction Metrics

The HBSI metrics are divided into two types depending on whether the presentation is successful or not. As already mentioned in the previous section, the application allows the user to delete signatures, and also asks for a reason for doing so. This, together with the newly defined flow chart, limits the possibilities of HBSI metrics to the following two:

-FI (False Interaction): "*Incorrect presentation is detected by the system and classified as correct*".

-CI (Concealed Interaction): "*Incorrect presentation is detected by the system, but not classified as correct*".

In the handwritten signature recognition state-of-the-art there is no method to measure the sample quality to determine its correctness, so new definitions for FI and CI are needed. It is important to note that the user is aware of whether the signature meets the similarity requirements or not through the feedback provided by the application. Therefore we propose the following definitions:

-FI (False Interaction): "*Presentation with a similarity score below the defined threshold that is classified by the user as correct (accepted)*".

-CI (Concealed Interaction): "*Presentation with a similarity score below the defined threshold that is classified by the user as incorrect (deleted)*".

Stress influence tests

This kind of measurement is not included in the HBSI model, as there are no measurements related to the user's mood, so it is necessary to include them in the evaluation. It is important to remember that during the acquisition of signatures under stressful conditions, no feedback on the similarity of the signatures is provided. Then, these signatures are used to calculate performance, being compared with the ones obtained in V1 and V2, extracting from these the new metrics for the influence of stress on the biometric behaviour.

Performance

Most of the main outcomes of the evaluation are extracted from the performance results, which are the following:

-Error rates in V1, obtaining the FRR when comparing each of the 12 genuine samples with the corresponding template, and doing this with each of the users enrolled in the system. The

FAR is obtained by comparing the template from each user with the whole set of 12 signatures of the other 55 users, which are considered here as impostor signatures.

-Error rates in V2. Learnability information is obtained by comparing these rates with the ones obtained during the V1. The FAR and FRR are calculated in the same way as in V1.

-Error rates under stress conditions. Information about the influence of stress is obtained by comparing these rates with the two previous rates. The FAR and FRR are calculated in the same way as in V1 and V2.

-Error rates of the whole system in normal conditions: considering the signatures of the V1 and the V2 jointly. The FRR and the FAR are obtained in the same way as the V1, V2 and SIT, but considering 24 signatures instead of 12.

-Error rates considering the signatures from the V1, V2 and SIT jointly. This reports a result of a possible real environment where the stress factor is present sometimes. The FRR and the FAR are obtained in the same way as the V1, V2 and SIT but using 36 signatures instead of 12. The complete set of error rates is summarized in Table 2.

5.4.4 Results and discussion

Usability and HBSI results

One of the strong points of this research has been to design a usability evaluation where users could complete the whole process with the minimum external help. Then, the aim at the development phase was to include several items of information in the application, including video and text guides. Most of the users considered the information received enough, but two of them (over 60 years old) could not complete the experiment, indicating that either usability or accessibility for elders may currently be one of the major weaknesses (i.e. text messages are not big enough). These two users also found complications when introducing their personal data due to their lack of skills in the new technology. In addition, six users asked for help during the process, having doubts related also to the smartphone accessibility (keyboard and writing) rather than to the main application objective.

After the evaluation of the different responses to the questionnaires filled in by the participants, the overall satisfaction can be considered successful as the quantitative parameters measured were all close to the highest possible values. The parameters, scored from 0 to 5, are shown in Figure 34. The global opinion score of the evaluation (from 0 to 5) was 4.13. The best scored feature by the users was easiness (4.38), while the worst one was intrusion (3.72). Furthermore, 87% of the users would use this kind of system daily. Half of the evaluation crew had not had experience of handwritten signatures in mobile devices and only 12.5% would not use it daily. Only 1.79% considered that the instructions given were not clear enough. Efficiency is measured through the time spent signing and the number of deleted signatures. As shown in Figure 35, the time employed to sign decreased from one phase to another, with the stress section being fastest. The time spent signing was not high on average (although it is user-dependent) and it was considered good by users (3.96/5) and it decreased also with familiarity.

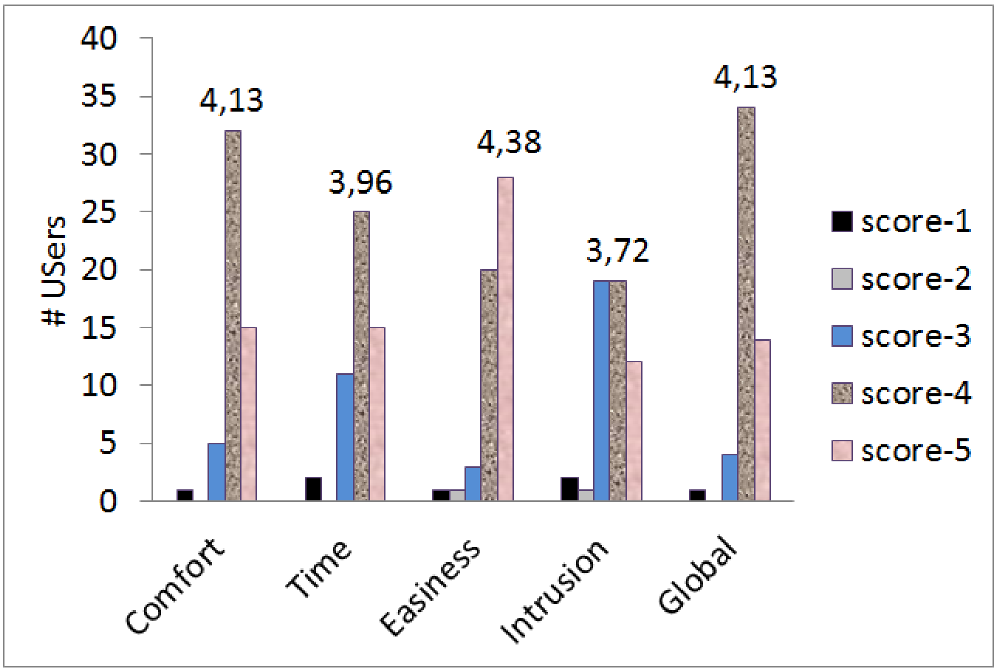

Figure 34. Results obtained from satisfaction questionnaire. Numbers over the bins are average scores

The numbers of deleted signatures are shown in Figure 36, divided into the different phases of the evaluation where users could delete their signatures, i.e. enrolment, V1 and V2. Accounting for all the signatures made in these three phases (1537) and the number of deletions (119), the non-deletions rate was 92.26%.

The results show that most of the deleted signatures were due to disliking the signature performed (74.29% of all deletions) and that this tendency did not noticeably decrease throughout each of the phases. After interviewing the participants, most of them commented that they decided to delete the signature because it was slightly different from the original one made on paper. The authors consider as a hypothesis for a future study that the number of deleted signatures will decrease once users get used to the application, the way the signature is shown in a digital device, the improvement of those devices, and its daily use.

Effectiveness was computed using the number of errors, assists and the percentage of task completion. Two fatal errors occurred, meaning that 2 users could not complete the evaluation (at the beginning of the evaluation there were 56 users and 54 finished it): one user (over 65 years old) did not manage to follow the instructions and complete the process by himself and another user introduced a wrong identity at the beginning of the second session. Thus the task completion factor was 96.43%. Regarding the errors made when signing, they did not increase significantly with the reduction of time needed for signing: by the end of the evaluation users were signing faster than at the beginning but the number of errors remained constant. This means the users' training was successful and they learnt quickly to use the application. We conclude that efficiency was satisfactory as the measured factors rates did not have a negative influence in the experiment.

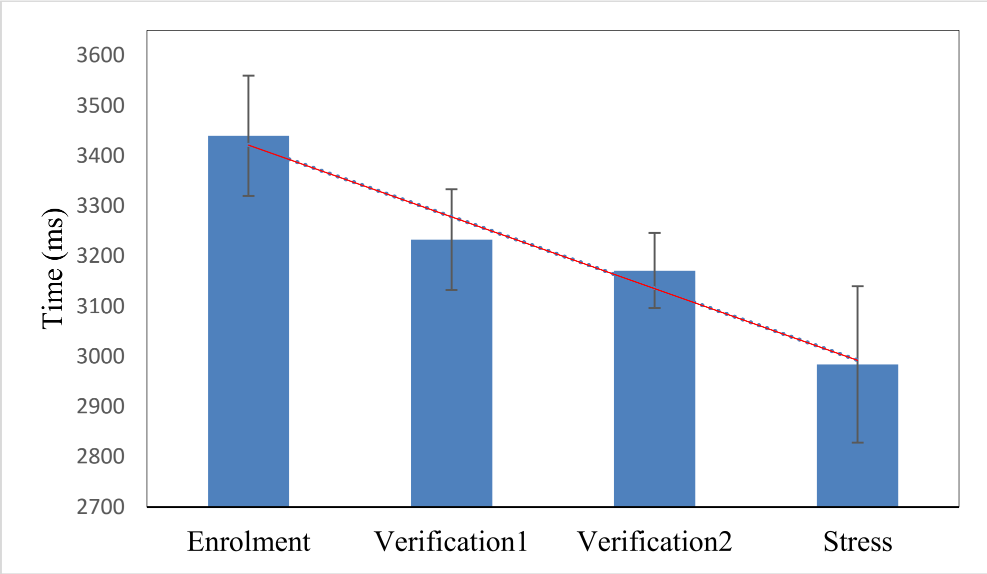

Figure 35. Average time spent by signature in each of the experiment phases. In red is the tendency line and segments are the variances

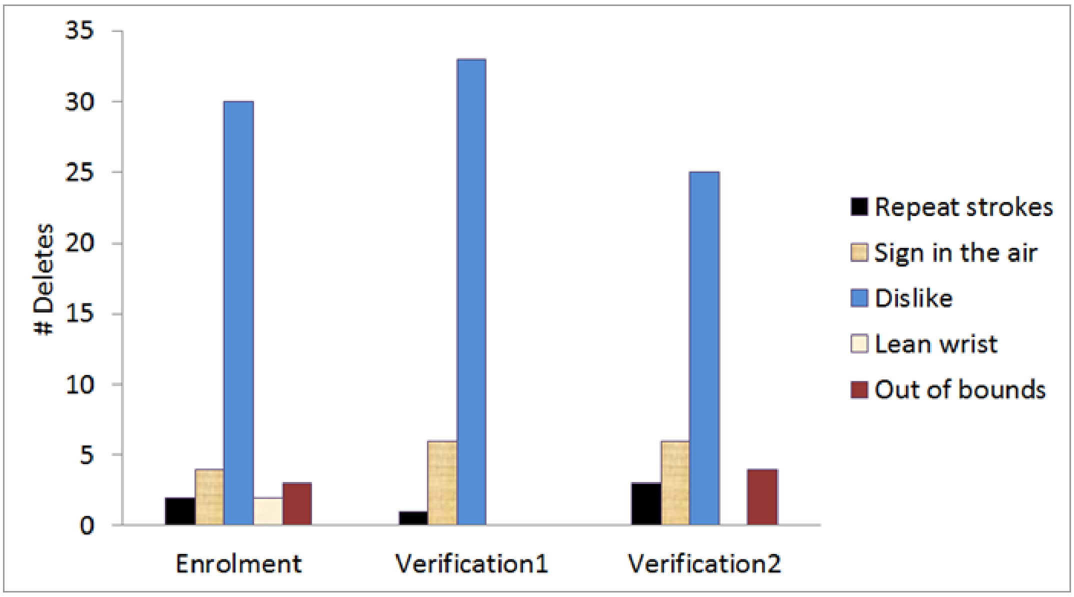

Figure 36. Deleted signatures by phases and reasons for deletion

Regarding ergonomics, the percentage of users that knew how to use the sensor equated to the users who did not need help. This was 87.30% because there were 6 users (10.70%) who needed help to complete the experiment due to misunderstandings with the information received. The percentage of users that learned how to use the capture sensor was the same because the two people who failed the evaluation did not start the second session. This is also applicable to the percentage of users that remembered how to use the capture sensor: 87.30% of users showed good skills at completing the second session. There was no user who was unable (physically) to use the capture sensor. As a complement to the ergonomics measures, most of the users (87.50%) signed seated and only 16.07% decided not to place the smartphone on a table. 7.14% signed while standing and 5.35% held the device themselves, which indicates that the users' preferred position was to be seated with the device resting on a table.

Regarding signal processing, there were no segmentation or feature extraction errors. The scores obtained from the users' feedback as a similarity measure between samples were high, as shown in Table 7.

Table 7. Similarity scores on average by phases. The ± specifies the variances

Phase	Similarity Score Average	Under Similarity Threshold (3/5)
Enrolment	4.41 ± 0.32	0%
V1	4.49 ± 0.20	0.31%
V2	4.42 ± 0.16	0.47%
SIT	4.01 ± 0.41	0.63%

There were only two similarity scores under 3 (red squares) in the V1 and three in the V2. All the squares were green during the enrolment. The SIT scores (calculated but not shown to users as feedback) were worse than the V1 and V2 similarity scores as expected, but only four signatures obtained similarity scores under 3 (the similarity threshold) in the stress section.

Only two signatures had similarity scores under 1 during the evaluation (in V2) and those users did not decide to delete them, increasing the FI rate. Thus the obtained rates were CI = 0% and FI = 0.10%, showing that the HBSI error rates were not influential.

Performance

Once all the signatures were gathered, the algorithm was applied to process them in order to obtain performance results. In Figure 37 various DET curves are shown: V1, V2, stress, AS (all including stress) and AWS (all without stress). A summary of the details including EERs is in Table 8.

The best performance was reached during V1, while in V2 the EER was clearly worse. The possible explanation is that several days had passed since the first session and there were no training sessions on the V2 day. Two solutions arise at this point: to include another session to better analyse the learnability or to extend the training session.

Those tests where the stress signatures were included returned the worst results. The results with the whole data set except for stress (AWS) were not far from the V1 in terms of EER, but according to a wider view of the whole performance (DET curves in Figure 37; the dashed black line is the EER) there was a clear difference. The better results in V1 may be due to the several signatures that users provided just before starting that phase (training and enrolment), thus acquiring habituation in terms of the consistency of the signature and therefore reducing the intra-class variation, as can be seen in Figure 37 (FRR decreased much faster in the case of V1).

Table 8. Performance results

Signatures set	# Genuine Signatures	# Impostor Signatures	EER
V1	12	648	0.31%
V2	12	648	0.94%
AWS	24	648	0.55%
SIT	12	648	1.26%
AS	36	648	0.84%

During the SIT, users tended to sign faster than in a regular environment, as shown in Figure 6. The resulting similarity score in the SIT was worst on average and the number of signatures under the similarity threshold was higher (Table 7). Also, the performance obtained from the signatures of the SIT was worse than the rest (EER = 1.26% with SIT signatures isolated, EER = 0.83% with the complete set (including SIT), and EER = 0.55% considering only non-stress situations). Even though the results under stress proved to be worse, the outcomes obtained from this test show an acceptable system performance under stressed scenarios. The average of the similarity score when providing feedback was 4.44 and this fell to 4.01 when removing feedback and subjecting the users to stress conditions. Furthermore, as can be seen in Figure 37, the DET curve of the SIT is below the V2 curve and close to the AS for several threshold values, indicating that stress in handwritten signatures is a non-major influential parameter on performance. Nevertheless, it is important to remark that the SIT was carried out a few minutes after the V2, so the users had been signing just before starting the SIT. This suggests that the order in the procedure is another influential parameter that should be measured in future experiments.

5.4.5 Conclusions obtained for future improvements

HBSI conclusions

The whole model was applied to this evaluation including the proposed metrics. However, because this is the first empirical approach to the analysis of a behavioural modality, the metrics adopted had to be adapted to this new context. These changes are impacted by the fact that the application allowed users to interact freely with the system (i.e. deleting signatures) and by the impossibility of knowing whether a signature is correct or not. Then, it is noticeable that the HBSI flow chart to decide the result metric is highly dependent not only on the modality but also on the characteristics of the evaluation. The ergonomic results demonstrate that the application is easy to use and the procedure is easy to learn and remember, as only two users could not complete the experiment. The sample similarity metrics applied as users' feedback provided some interesting outcomes in various aspects: users proved to be comfortable with an acknowledgment of their signature and there were not too many errors. Furthermore, similarity scores on average were quite good in all the phases.

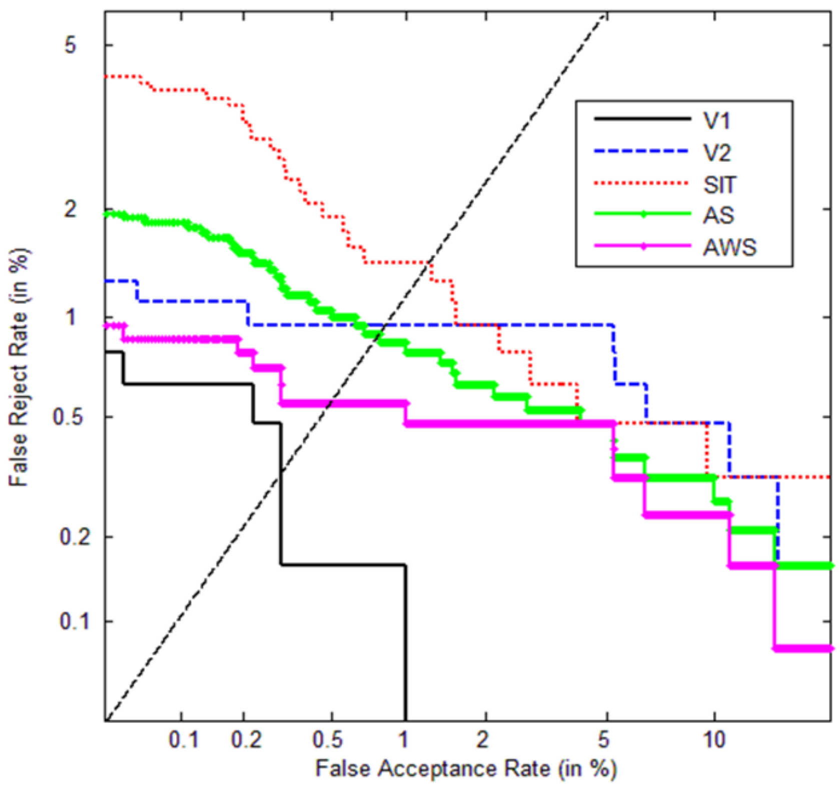

Figure 37. DET curves. The broken black line is the EER.

Stress influence and performance

The aim of causing stress in users was achieved, as all of them were anxious to finish this phase (as they expressed at the end of the experiment). In addition, the time spent completing the signatures was reduced ostensibly and the similarity scores were worse than in other phases. Nevertheless, the error rates obtained in the SIT reveal better performance than the V2 at several operating points in the DET curves, indicating that the stress factor is not a major drawback for recognition.

Regarding general performance results, these are similar to those obtained in our previous works [86] with the same mobile device (EER = 0.17%). Note that, in this case, most of the users had no previous knowledge of signing on mobile devices.

5.4.6 Outcomes to improve the H-B interaction

Outcomes obtained after this experiment suggest the following improvements to the H-B interaction:

Factors depending on the user. This experiment, focused on the users' stress, shows a clear influence on the usability. Users' conditions, both temporal (stress, fear, etc.) and permanent (e.g. patience), influence clearly in final results. Also in this case, accessibility concerns may involve misuses (e.g. deaf or blind users not receiving the proper feedback).

Factors depending on the system. In this case, the feedback provided by the hardware is highly relevant and must be included in the H-B interaction.

Factors depending on the user-system interaction. User's displacements motivated by external factors (annoying sounds and lightning in this case), may be noted as variations in the presentation of the biometric characteristics. Also the presence of other users may distract the user interaction with the biometric system (factors depending on the environment).

H-B interaction metrics. Cognitive concerns may cause the impossibility to follow the instructions of this app as it requires understanding several instructions. Also the presence of non-common signals could cause problems to some users. Accessibility characteristics are required to properly measure the usability of the system: users who can/know use the app.

Updating the HBSI. The HBSI has been tested only in a few biometric modalities. In all of them, changes were suggested because the interaction differs. In this case, we have also suggested changes according to the specific experiment features. The HBSI must be updated also regularly in order to cover the modality dependant factors.

5.5 EXPERIMENT 4a: Time efficiency in behavioural biometrics under stress

This experiment [87] is intended to better analyse the time influence (or the efficiency) when interacting with biometric systems under mood changes. Specifically, we have used all the data gathered during the previous experiment, so the biometric capture process has been already explained. In this part, we show the methods and results obtained to analyse the time influence.

Motivation of the experiment

Time is a major factor of influence when interacting with biometric systems: long interactions may cause impatience on users. This is clearer in mobile devices where interactions are supposed to be quick. Therefore, in this experiment we wanted to know how the time influences the performance when users are under the effects of the stress and also contrast several hypotheses regarding this topic. This experiment involves no changes on the H-B interaction, apart from those cited in the previous experiment.

5.5.1 Hypotheses and data processing

As shown in the previous experiment, it is undeniable that stress is a factor that modifies human behaviour, but in which extent does the stress influence DSV in terms of performance over time? In this section we define the initial hypotheses and how we have processed the gathered data in order to contrast those hypotheses. The hypotheses are:

i) Signatures under stress are shorter in time than signatures under normal conditions.

ii) Signatures under stress are worst in performance.

iii) The longer the signature in time the worst genuine rates.

iv) Performance in Verification 2 is best than in Verification 1 because users already got skills.

v) Last signatures under stress are worst in performance than first signatures.

vi) Last signatures under stress are longer in time than first signatures.

vii) The performance is more stable in short signatures than in long signatures

Data Processing

The data processing was made once all the signatures were collected and all the users have finished the evaluation. First of all, in order to contrast the initial hypotheses, it was necessary to split the database in groups according to the different processes and length of the signatures.

The first division was made by processes: Verification 1, Verification 2 and Stress tests. The enrolment signatures were used as templates. To differentiate between long and short signatures (in average) we have divided the users in 4 subsets (focusing on the stress tests) afterwards:

A- Users who spend less than 2 seconds signing (15 users)

B- Users who spend between 2 and 4 seconds signing (27 users).

C- Users who spend between 4 and 6 seconds signing (8 users).

D- Users who spend more than 6 seconds signing (2 users).

Our intention is to contrast hypothesis vii) and possibly stablish further differences between groups. Then, all the genuine comparisons were made, matching the signatures obtained during the 3 verification processes with the templates obtained during the enrolment.

Algorithm

The performance results were obtained applying the DTW based algorithm (same as in the other experiments) to the signatures, using the X and Y time series coordinates of the signature as inputs. In this case, returning a similarity score between two signatures as output: the closer to zero, the more similarity between signatures. It was used in the mobiles application to calculate similarities between the signatures and give feedback to users just after performing a signature and pressing the Accept button. It was also used once all signatures had been acquired from all the subjects in the test crew, in order to obtain global performance.

5.5.2 Results and discussion

In this section are all the obtained results regarding time and performance. Most of the graphs shown include polynomial regression curves (2nd order) fitting the results to better understand the evolution of the time-performance variables. First, we compare among all the verification processes and then we focus on the stress tests.

Verification 1 vs Verification 2 vs Stress tests

The genuine scores obtained in the Verification 1, Verification 2 and Stress tests are in Figure 38 and the time employed signing is in Figure 39. So, it is feasible to contrast hypotheses i) and

ii) because as shown in Figure 38, stress signatures are worst in performance and as shown in Figure 39, they are shorter in time than the rest, as expected.

The differences is time are meaningful but not the differences in performance. This means that users indeed felt stressed (they made the signatures as quickly as possible), but at the same time, this also points out that the signature recognition is still reliable under stressful scenarios (differences in performance are ~ 0,1 between Verification 2 and Stress tests).

The scores of all the signatures in the three verification processes with respect the time are in Figures 40, 41 and 42. The blue crosses represent each comparison and the red lines represent the regression curves fitting the results. Those regression curves show a clear increasing in genuine scores (the closer to zero the better performance) along with the time for the three verification processes, answering to the hypothesis iii): longer signatures return worst results in genuine performance than shorter. We can also deduce that the performance decreasing during the stress test is not due to the less time spent signing (as shorter signatures are better in performance than longer) but due to faster movements, different strokes, etc. Figures 40-42 also show a performance decrease no matter if the users are under stress or not.

Another conclusion is that performance in Verification 2 is worse than Verification 1. So that, the hypothesis iv) is false, showing that users did not get enough habituation in session 2 to outperform session 1 results. One possible reason is that in session 1 users had made the enrolment and training first and in session 2 (one week apart) they did not train previously in the same day.

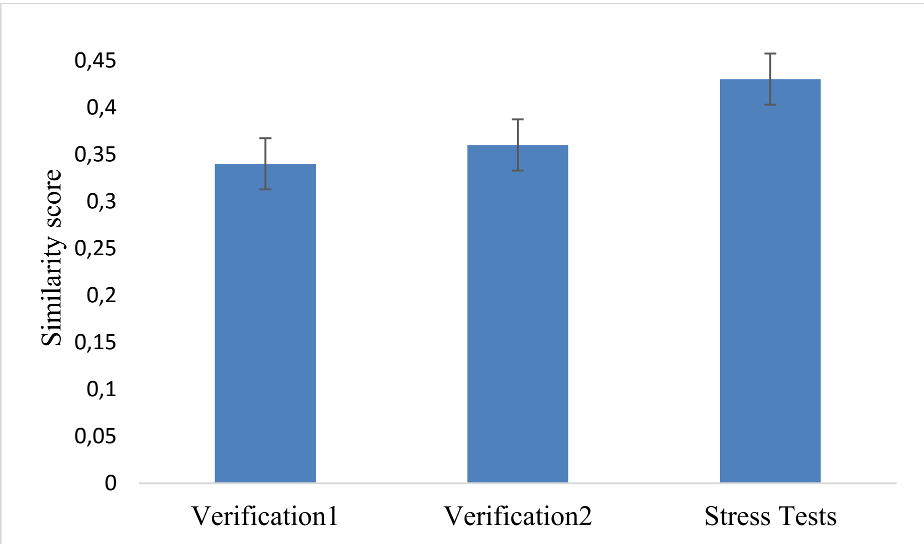

Figure 38. Average genuine rates in all the verification processes. Segments are the variances

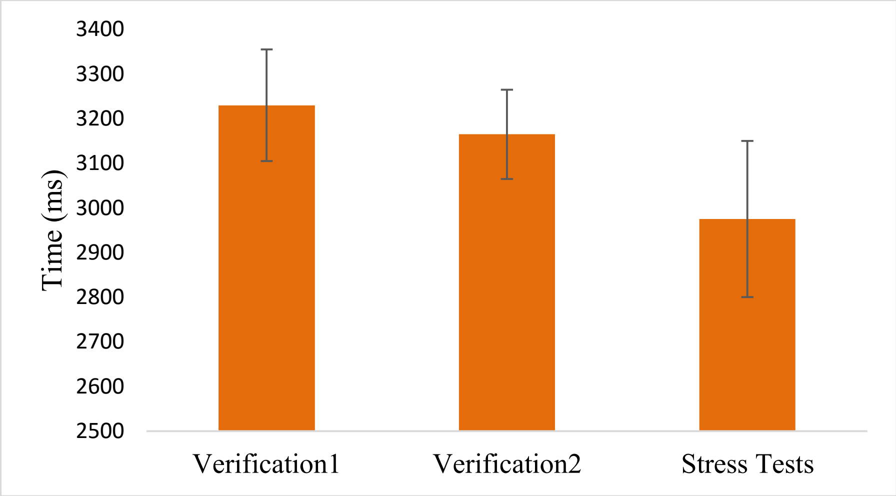

Figure 39. Average time in all the verification processes. Segments are the variances

Differences by time spent signing in the Stress tests

We have divided the evaluation crew in 4 groups according to their signatures length in average as explained in section 2. According to Figure 42, users who spent less time signing obtain better results in performance, but as shown in Figure 43, this performance fluctuates more in groups A and D. Then, those users who spent less than 2 seconds signing and users who spent more than 6 have bigger variations in genuine rates than the rest. This involves that hypothesis vii) is false a priori, because the most stable signatures in performance are those of groups B and C (between 2 and 6 seconds of length).

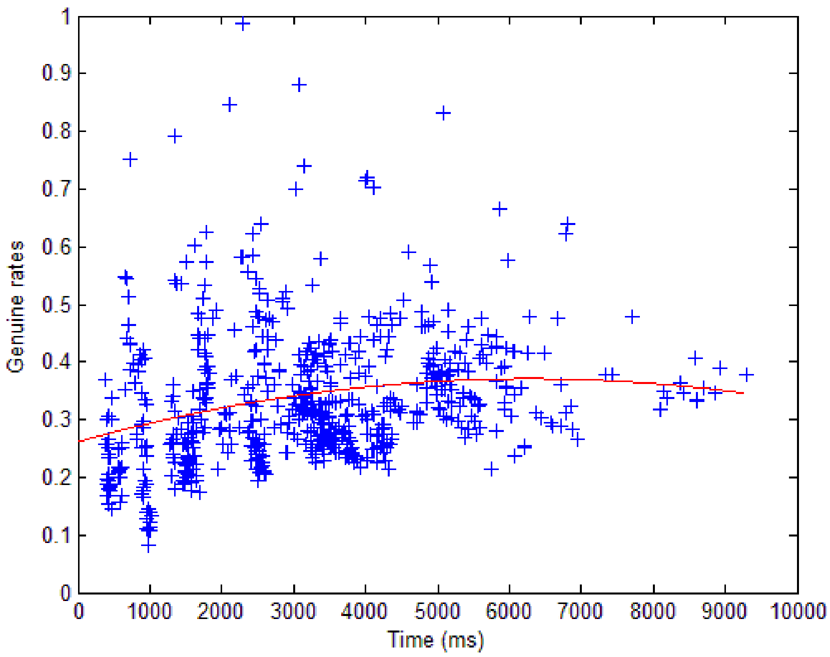

Figure 40. Verification 1. Evolution of the genuine rates in time

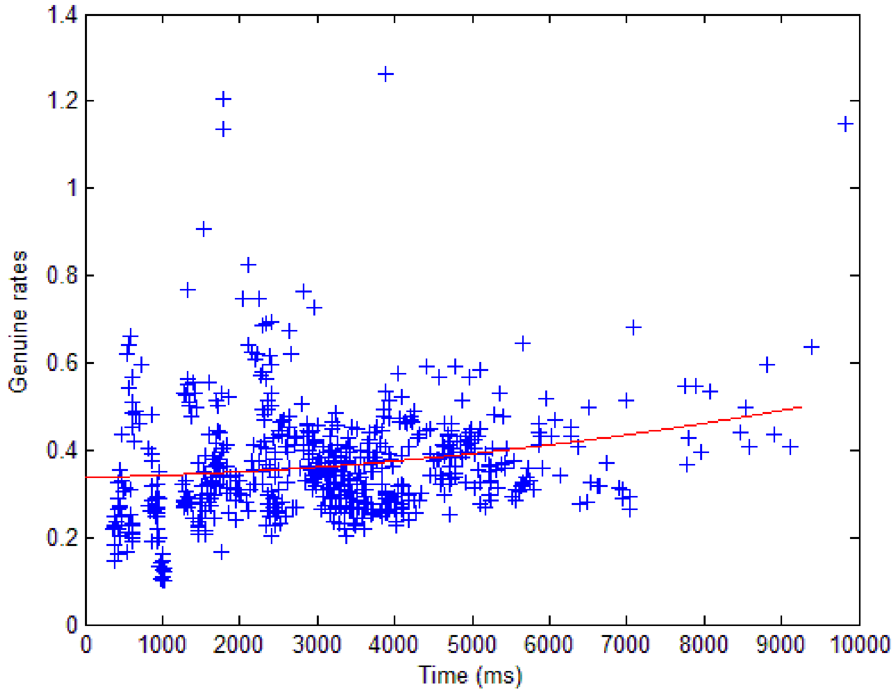

Figure 41. Verification 2. Evolution of the genuine rates in time

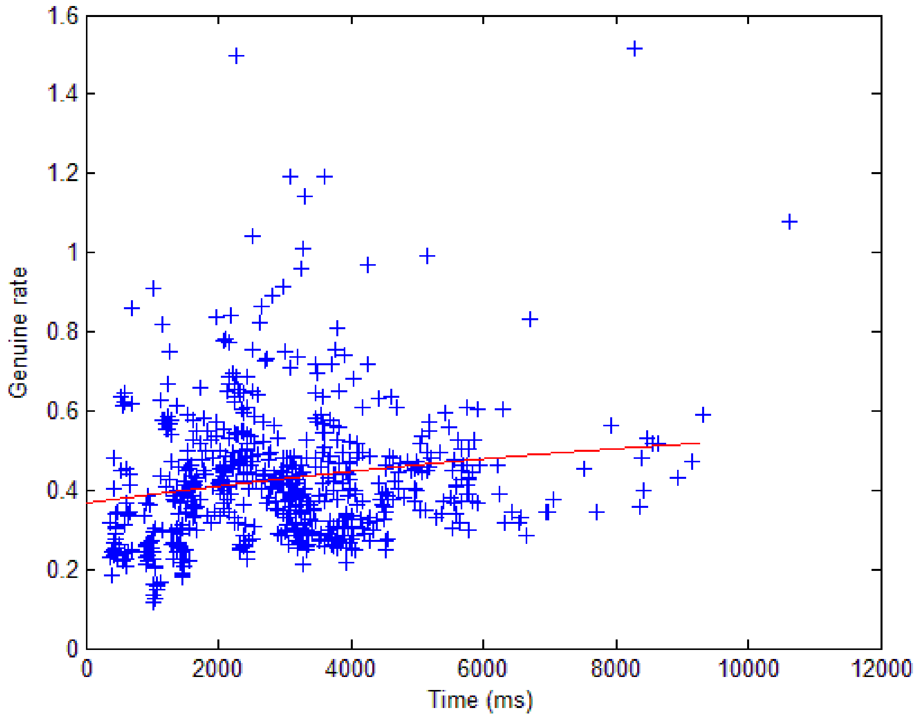

Figure 42. Stress tests. Evolution of the genuine rates in time

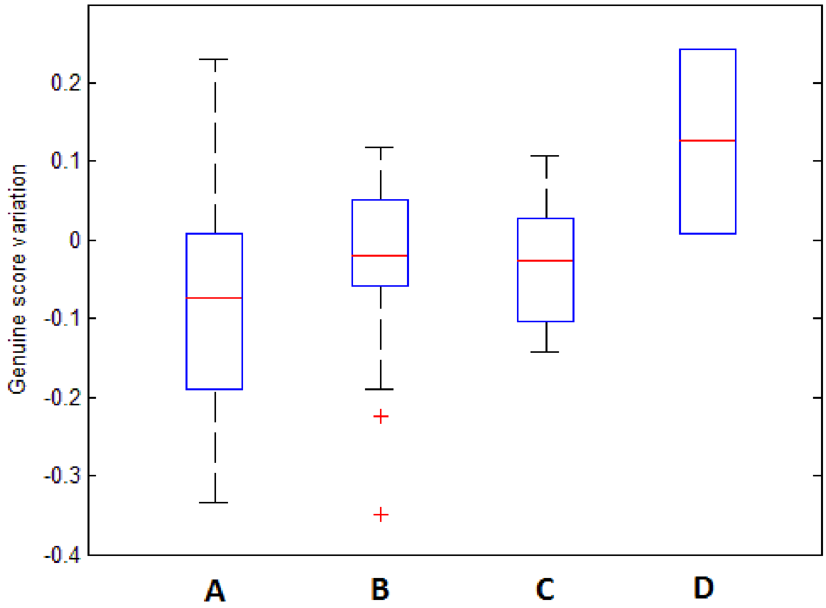

Figure 43. Genuine scores variation divided in groups by signatures length. In the boxplots: the central box represents the central 50% of the data. Its lower and upper boundary lines are at the 25%/75% quantile of the data. A central line indicates the median of the data. Two vertical lines extending from the central box indicate the remaining data outside the central box that are not regarded as outliers. These lines extend maximally to times the height of the central box but not past the range of the data. Outliers (+): these are points indicating the remaining data.

The last two hypothesis have to do with the differences in time and performance between the first and the last signatures. Once we have gathered the final results, we have found several inconsistences between users of the same group, so it is not possible to stablish conclusions in any group or even compare between them. In Figure 44 (performance) and Figure 45 (time), the signatures ordered (from 1 to 12) from one user of each group are shown. According to both figures there are sharp changes in time and performance from one signature to the next, without any distinguishable pattern. For instance, in the Figure 44 – user A, the signature 10 has scored 0,36 (~0,1 worse than the first one), but the signature 12 (the last one) has scored better than the first one. Regarding the time, the results are similar than the performance results in terms of differences among signatures, the first and the last signatures are close in time but there is a variation of almost 5 seconds between the third and fourth signatures.

This tendency changes from one user to another, so it is not feasible to stablish any correlation or tendency. These results point out that the stress influence in users involves unpredictable behaviours.

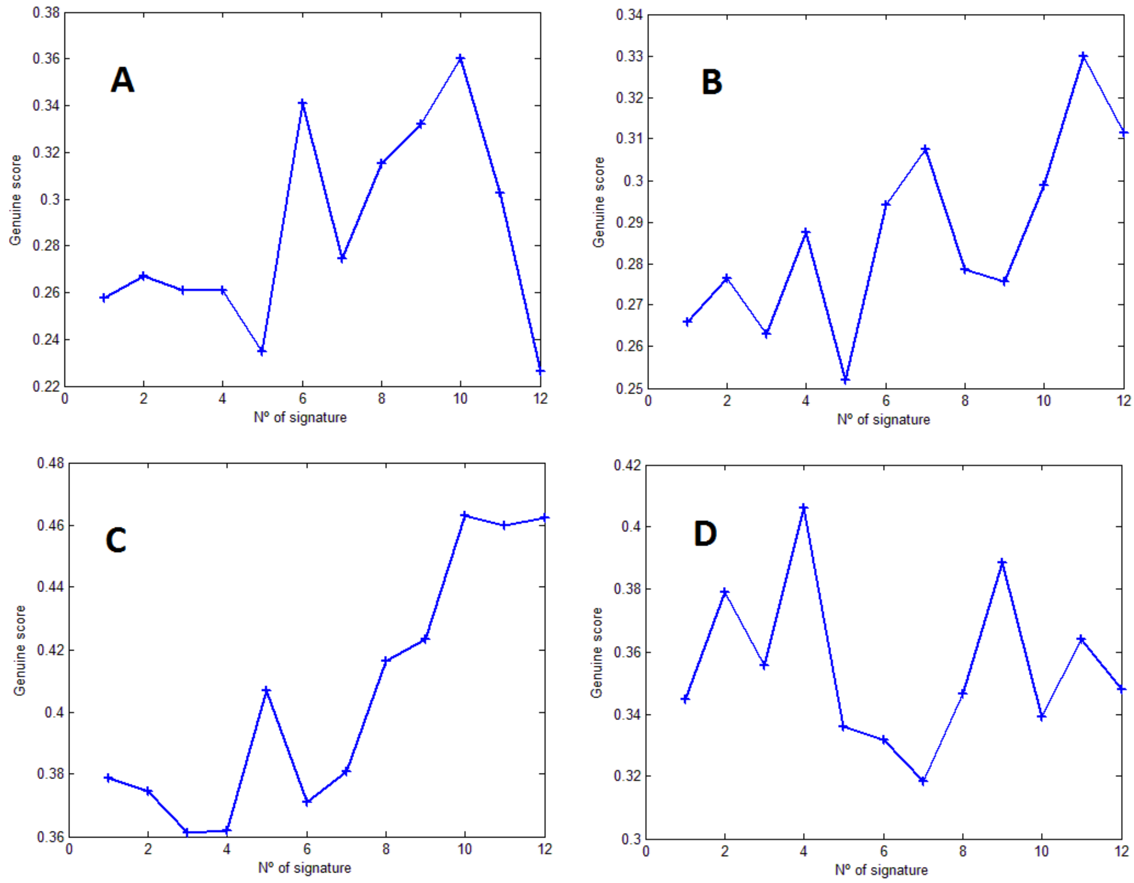

Figure 44. Genuine scores from the first to the last signature in users of the four categories randomly chosen

5.5.3 Conclusions obtained for future improvements

Once all the results were analysed, we determine that stress is an influential factor in DSV which affects both performance and efficiency in terms of time. Tough the genuine results obtained during the Stress tests are close to the Verification I and II results, showing that DSV is still reliable under stress conditions.

Another important conclusion is that the signature length is determining, both under normal and stress situations. Our findings suggest that the longest signatures (group D) are the worst in performance (genuine) and highly variable under stress. The best in performance are the shorter signatures (group A), but again the variability between them is high. This fact indicates that signatures from groups B and C, which are the less variable, are probably the more stable and less influenced by stress (although worse in performance than group A).

The aim of causing stress in users was achieved, as all of them were anxious to finish this phase (as they expressed at the end of the experiment). Stress conditions make the user feel unsecure and involve unpredictable results. Therefore, the hypotheses regarding the differences in time and performance between the last and first signatures (v) and vi)) are false, because this results vary significantly depending on the user. Other factors must be taken into account at the time to predict the stress influence extent in a particular user, such as tiredness, time of the day, personal situations, etc.

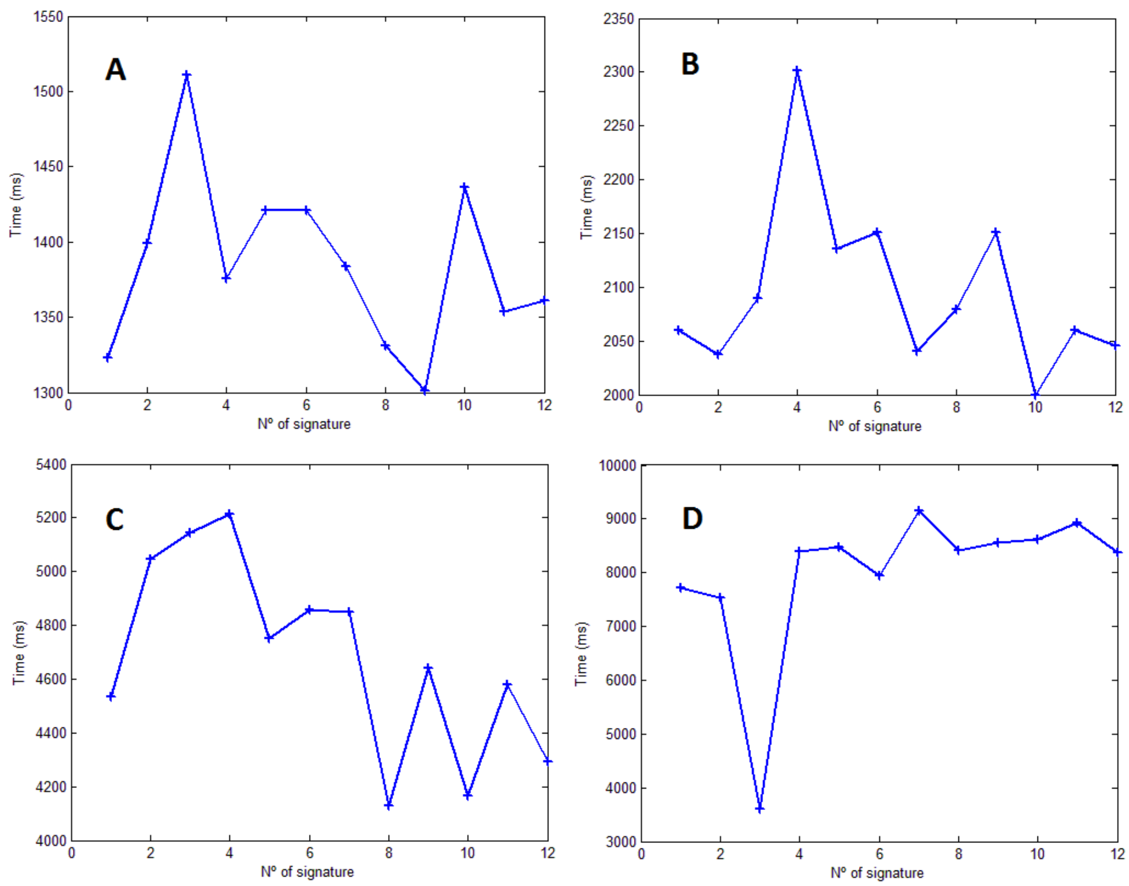

Figure 45. Time from the first to the last signature in users of the four categories randomly chosen

5.6 EXPERIMENT 5: User Acceptance in biometric recognition

This experiment is focused on the user acceptance of biometric recognition, specifically on fingerprint. User acceptance, which is an important part in the usability analysis is a key factor when designing biometric systems. The scarce number of works in this line encouraged us to carry on this study. In this experiment, 589 users have used 3 planar semiconductor fingerprint sensors and fulfilled 2 surveys (previously and after) where they were asked about several aspects regarding biometrics in general and fingerprint recognition in particular. Our results include how user's opinion changes when interacting with biometrics and what are the ergonomic preferences, among other findings. Performance aspects have not been overcome in this experiment, but they are already published [88].

Motivation of the experiment

Our main goal in this experiment was to test the user acceptance of biometric recognition within a relevant enough user's number. To do this, we designed pre and post surveys when using various fingerprint sensors. The questions made where related to security, biometrics in general and fingerprint recognition. The reason for moving from DSV is that fingerprint recognition is widely better known and it is more integrated in the society: smartphones, entering buildings, ABC (Automated Border Control) systems, ID cards, etc. Then, we wanted to test user acceptance in the most representative biometric modality.

Influence on the H-B interaction

Pre and post surveys are common usability techniques to test the user interaction and not present in the methodology. Further factors to analyse are those related to the devices ergonomics. The interface, which is in this experiment handle by operators shows also information to users and therefore must be clear and easy to follow.

5.6.1 Evaluation set-up

Users

There were no restrictions to participate in the experiment (apart from having the middle, index and thumb fingers and being over 18 years old). Then, users are people of all ages, genders and laterality. Moreover, they have different studies degrees and technology knowledges. 589 users started the first session and 581 of them have completed the whole experiment (8 users did not attend to the second session). Finally, 515 users completed correctly the surveys and therefore their data is valid for this experiment. In Figure 46 are the graphics which include user's personal data (gender, age, studies degree and laterality) and knowledge about computers, cameras, mobile devices and biometrics.

Devices

Users interacted with 3 different planar semiconductor fingerprint sensors (Figure 47):

- NEXT NB-3010-U. This fingerprint sensor is thermal, its resolution is 385 ppi in 256 gray scale levels and its dimensions are 11,9 (width) x 16,9 (height) mm of active sensing area.
- FINGERPRINTS FPC1011F3. This sensor is active capacitive, its resolution is 363 ppi in 256 gray scale levels and its dimensions are 10,6 (width) x 14 (height) mm of active sensing area.
- UPEK EikonTouch 510. This fingerprint sensor is thermal, its resolution is 508 ppi in 256 gray scale levels and its dimensions are 12,8 (width) x 18 (height) mm of active sensing area.

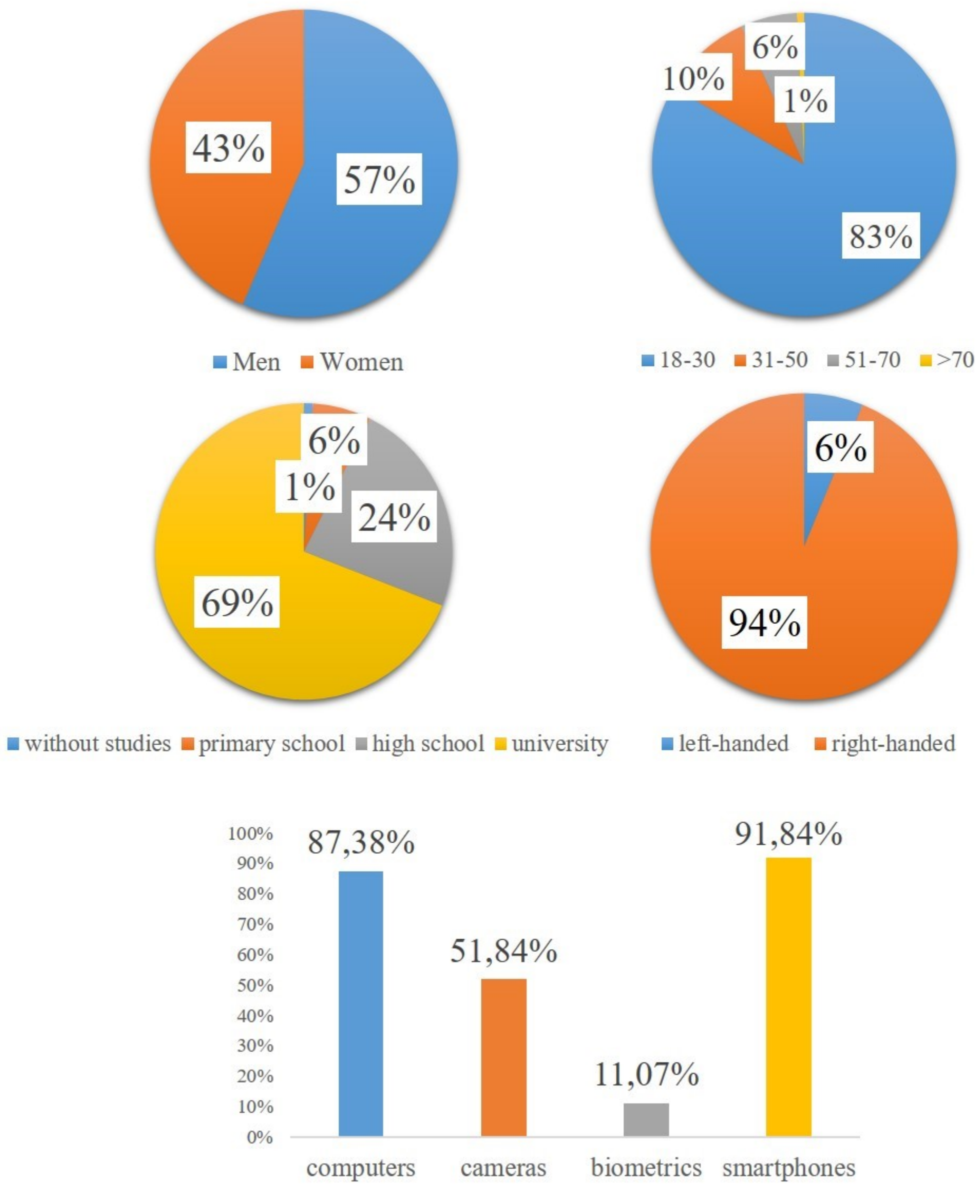

Figure 46. General characteristics of the experiment participants

Figure 47. Fingerprint sensors used in the experiment. From bottom to top: NEXT, Fingerprints and Upek

During the enrolment and verifications we applied fingerprint quality and recognition algorithms in order to:

Discard incorrect presentations. If the NFIQ [89] result was higher than 3 (in level 4 the quality is considered only “fair” and in level 5 is considered “poor”), the image was discarded automatically by the application and a new attempt was required.

Obtain verification results on the spot. When an image is equal to or below 3 in the NFIQ test (in level 2 the quality is considered “very good” and in level 1 “excellent”), then NIST NBIS [90] is applied. If this value is below a threshold (fixed in 20) then the verification is unsuccessful.

5.6.2 Study of the user acceptance

The user acceptance results are extracted mainly from the surveys which were intended to be totally completed by users (without external help). In order to not bias the user’s responses, the surveys were printed to be easier to complete autonomously (especially for elderly users who may find online surveys cumbersome).

Figure 48. App Interface of the enrolment. There are several elements intended to help the user and the evaluator (finger to present, device, attempt, sample, fingerprint quality (NFIQ), image and user ID)

Pre-experiment survey

In the S1 users completed the personal data and 6 questions regarding security and fingerprints:

- *Do you think fingerprint recognition is FASTER than PIN? Yes/No*
- *Do you think fingerprint recognition is more COMFORTABLE than PIN? Yes/No*
- *Do you think fingerprint recognition is more SECURE than PIN? Yes/no*
- *Generally, what do you prefer? Fingerprint/PIN*
- *Would you use your fingerprint to unlock your smartphone or laptop? (Select at least one): Yes / No – it is slow / No – it is uncomfortable / No – it is unsecure / No - other option.*
- *Would you use your fingerprint to make payments or get money from the ATM? (Select at least one): Yes / No – it is slow / No – it is uncomfortable / No – it is unsecure / No - other option.*

Post-experiment survey

There are the same 6 questions in the post-experiment survey in order to check the evolution of the user's opinion. In addition, there are other 5 questions regarding usability and ergonomics:

- *How was the use of the fingerprint? (Mark from 0-very uncomfortable to 6-very comfortable).*
- *The time per presentation was… (Mark from 0-very short to 5-very long).*
- *Mark from 1-very bad to 5-very good the convenience and speed of the 3 fingerprint devices.*
- *Were the received instructions clear and enough? Yes/no*
- *Did you have any inconvenience to complete the experiment? No / Yes - uncomfortable devices / Yes - Too many attempts needed / Yes - It is slow.*

5.6.3 Results

The results obtained are shown in figures to make them easier and faster to understand. In Figures from 49 to 52 are the results of the S1 and S2. In all the figures FP is fingerprint. We have compared fingerprint versus PIN due to PIN is the most usually applied security tool in those scenarios where fingerprint could be also used and people is generally familiarized with it.

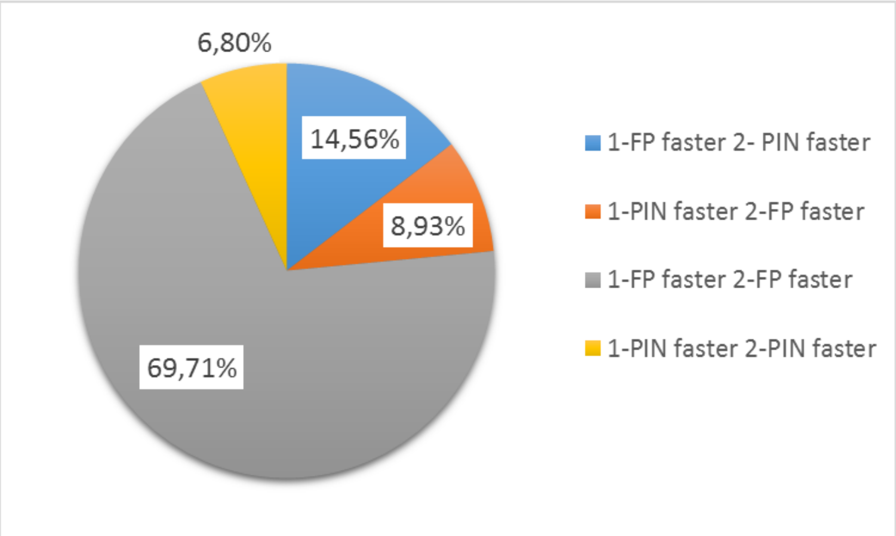

Figure 49. User's preferences regarding speed

Most of the users (69,71%) chose the fingerprint in both surveys as the faster solution (Figure 49), the most comfortable (80,19% as shown in Figure 50) and the most secure (78,64%, Figure 51). The majority of users (67% as shown in Figure 52) preferred to use fingerprint than

PIN and maintained their opinion at the end of the experiment. It is also significant that 17% of users preferred fingerprint at the beginning and changed their mind once they have completed the experiment.

Regarding the questions about preferences for locking/unlocking and make payments, most of the users kept their willingness of using fingerprint instead of PIN as it is shown in Figure 53 and Figure 54. In Figure 55 are the user's scores of the device's comfort and speed. There are not big differences between them (except for the device made by Fingerprints, which scored slightly lower than the other two).

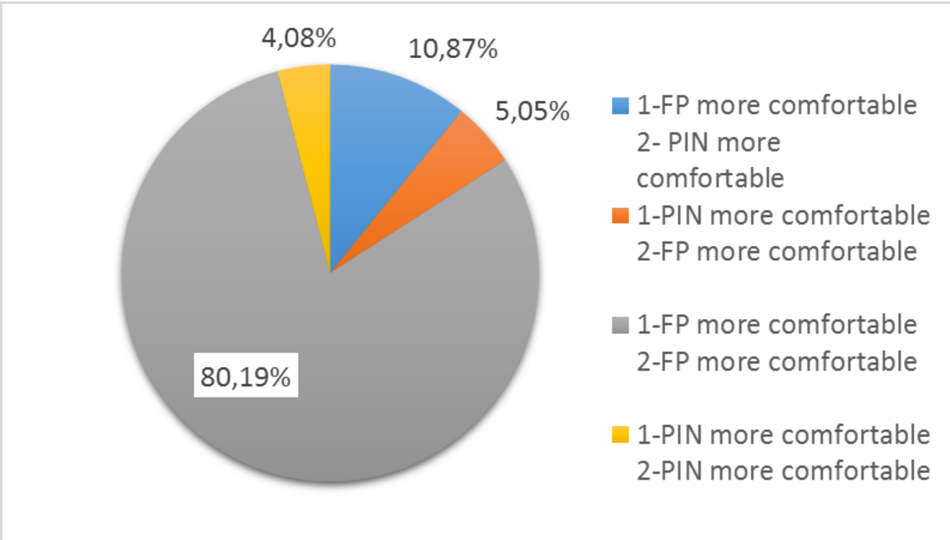

Figure 50. User's preferences regarding comfort through S1 and S2

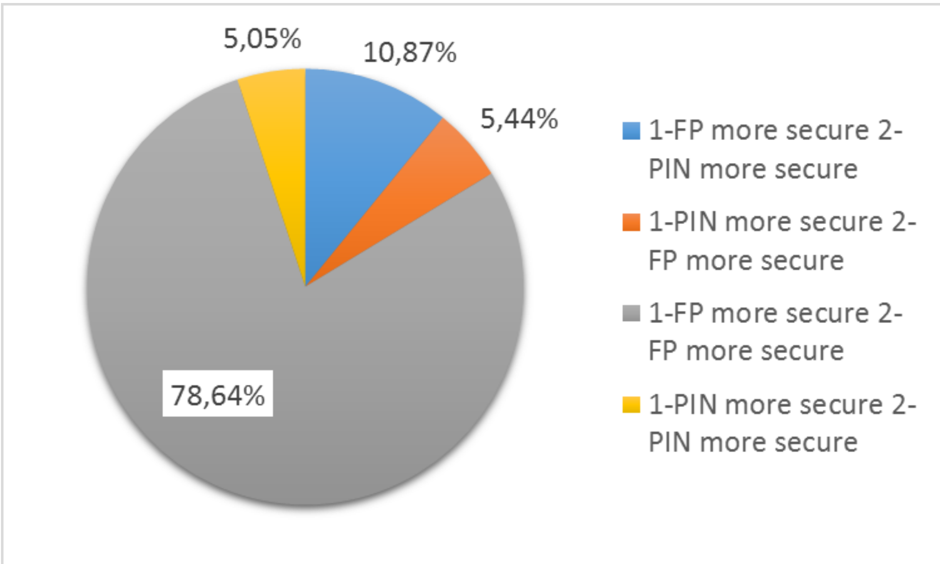

Figure 51. User's preferences regarding security through S1 and S2

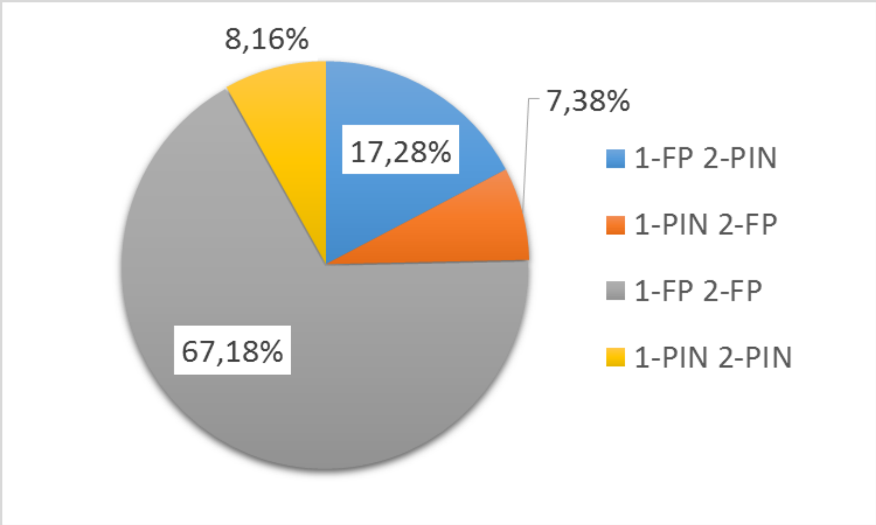

Figure 52. User's preferences regarding the use of Fingerprint or PIN through S1 and S2

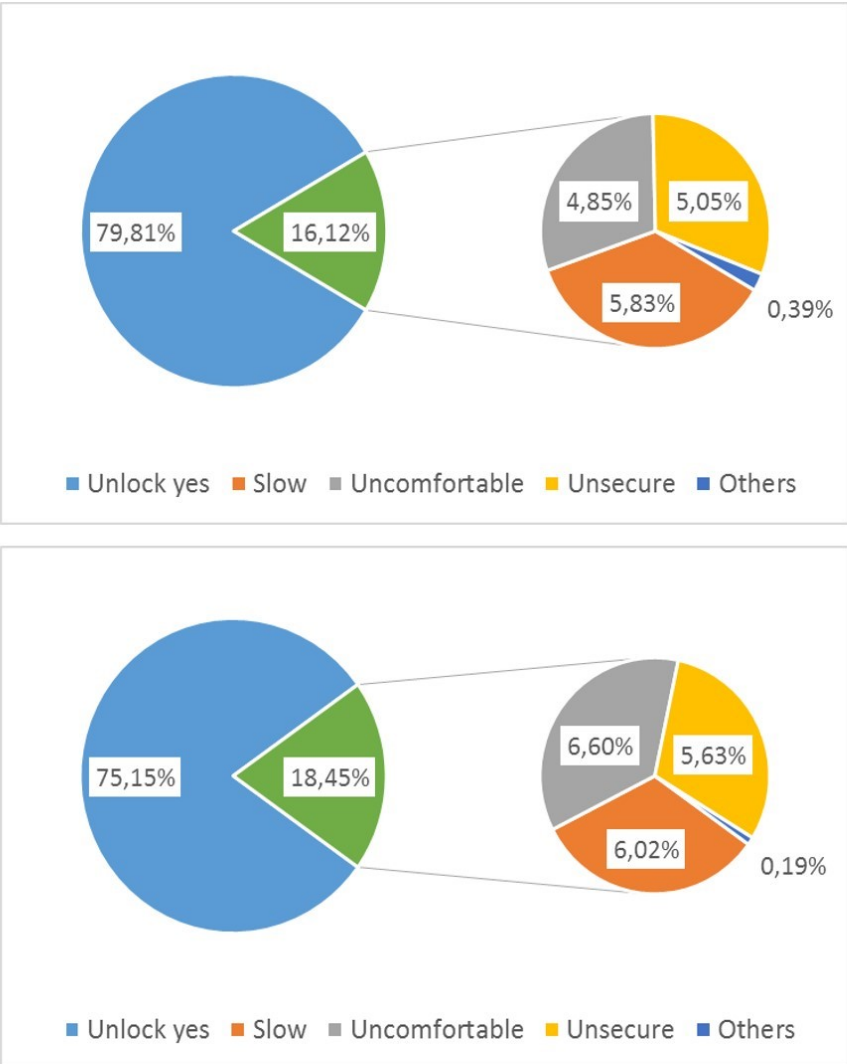

Figure 53. Percentages of users who would use fingerprint for locking/unlocking devices and those who not (right pies indicate why not). In the upper part of the figure is the S1 results and in the below part is the S2 results

5.6.4 Conclusions obtained for future improvements

Several conclusions can be extracted from the results obtained. In this section those conclusions which authors considered as the most relevant are discussed by categories:

Users. The majority of the participants (83%) are between 18 and 35 years old and have university degrees (69%). People from this age range are the current or potential biometrics consumers/users. Therefore, this results may be a good indicative of the biometrics future user's preferences.

Devices. Some users complained about the capture area of the fingerprint devices. Especially those participants who have long nails, who experimented problems when placing the fingerprint on the sensor. This is an ergonomic concern that should be overcame. Also users with dry hands had problems to be detected by the sensors. On the other hand, according to their size, these sensors could be embedded in a mobile device. Then, authors suggest that those users who did not find inconveniences when interacting with the devices would not experience problems using fingerprint in mobile devices.

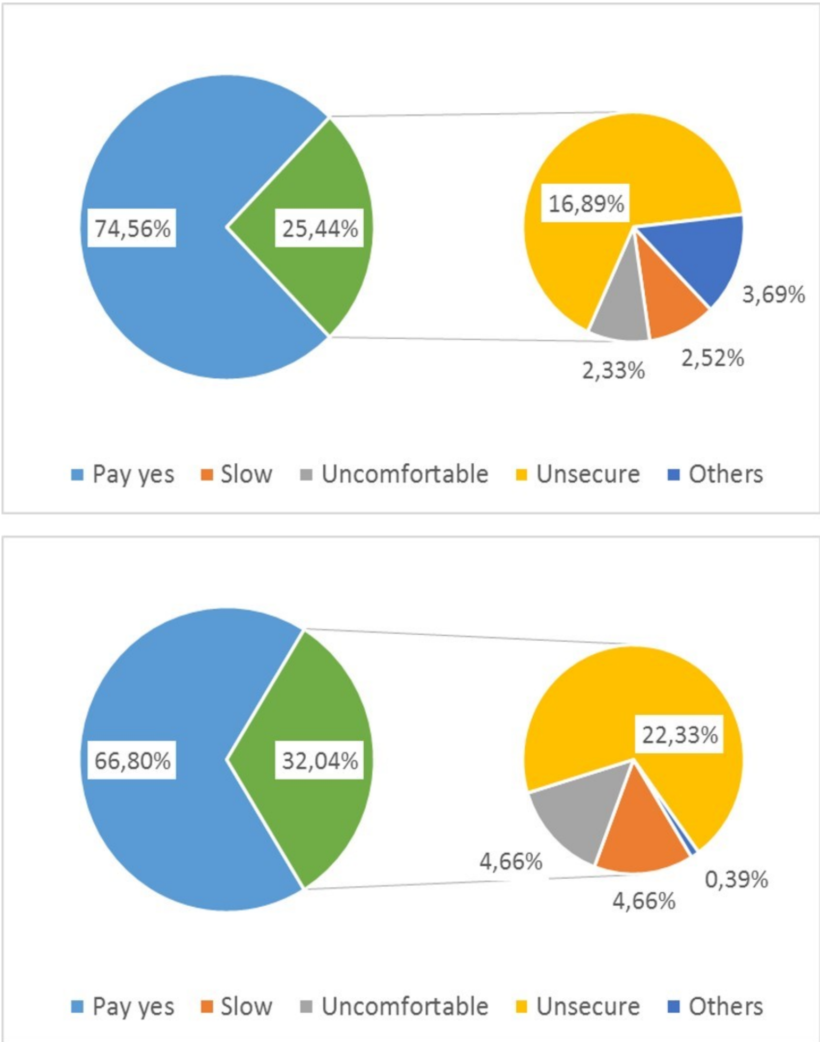

Figure 54. Percentages of users who would use fingerprint for making payments and those who not (right pies indicate why not). In the upper part of the figure is the S1 results and in the below part is the S2 results

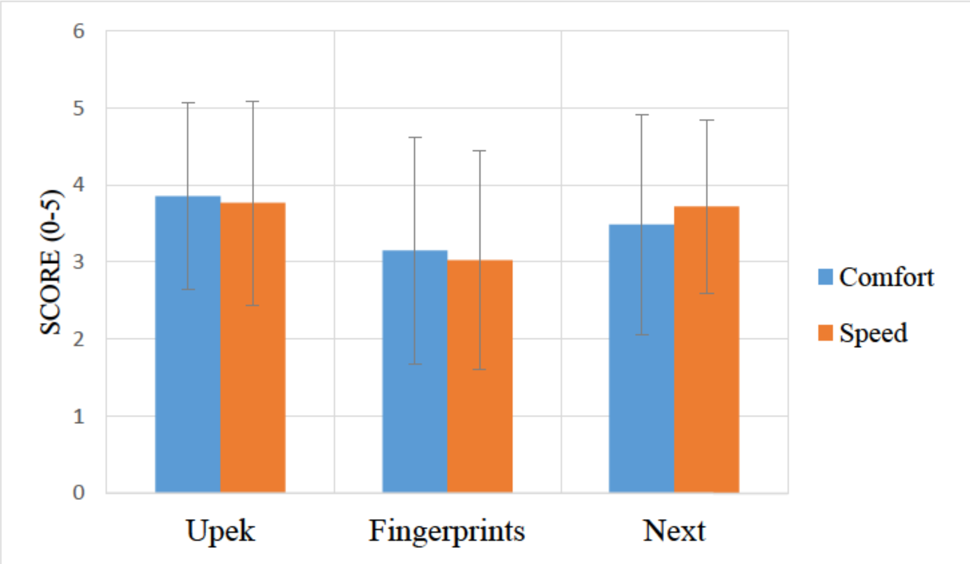

Figure 55. Average user's scores regarding device's comfort (left columns) and speed (right columns). These values were extracted from S2. Segments are the variances

The most preferred devices regarding ergonomics are the NEXT and the UPEK, which are the lighter ones. This could happen due to several users picked up the devices when placing thumbs to avoid the wrist twisting. Therefore, the lighter the device, the more comfortable. Regarding the speed, the scores are similar, which could be motivated also for the wrist twisting. One of the possible reasons why the devices did not have a higher score is the high number of times that users had to present their fingers. Then the tiredness could bias the user's opinion.

<u>Interfaces</u>

Most users remarked the utility of the interfaces and proceeded several times without the necessity of guidance. Then we conclude that showing to users which combination of sensor-device they have to use, they feel more comfortable and complete processes making less mistakes. Also when users see the fingerprint image they quickly check if the finger was correctly presented to the device or not (and why not in that case).

<u>Fingerprint vs. PIN</u>

Three factors (those who authors considered the most relevant when deciding which technology to use) were compared: comfort, speed and security. Finally, authors asked whether the users prefer fingerprint or PIN. For all of these questions the majority of users prefer fingerprint in all cases (in both surveys). Nevertheless, there were more users who changed their opinion from fingerprint to PIN than from PIN to fingerprint also in all cases. This fact could be motivated by the time spent in the experiment and the ergonomic issues above mentioned (long nails, dry fingers, etc.). Almost 70% of the participants preferred to use fingerprint over PIN before and after the experiment, but again a significant percentage of them (17%) changed their opinion after the experiment.

Uses of fingerprint recognition

Users were asked about the use of fingerprint in two different situations covering two degrees of security and usage: lock/unlock an electronic device (low security and high frequency of usage) and using banking services (high security and not high frequency of usage). In both cases, the percentage of users who would use fingerprint decrease almost 10% after the experiment. Regarding the devices locking/unlocking, users would generally use fingerprint recognition (75%-79%). User who would not use it are equally distributed among those who argue comfort, security and speed concerns (in both surveys). In the banking scenario, still the majority would use biometrics but the percentage is lower (74%-66%). In this case, those users who would not use fingerprint argued mainly security reasons (16%-22%). Therefore, authors suggest that users feel comfortable with the use of fingerprint recognition but they do not feel totally confidence relying medium-high security issues on it.

5.6.5 Outcomes to improve the H-B interaction

Outcomes obtained after this experiment suggest the following improvements to the H-B interaction:

Usability evaluation methodologies. Proper prior evaluations, including user's opinions since early prototypes, may improve the devices design and save time and money. Those tests, in UCD mode, should be included in the H-B interaction

Factors depending on the user. Long nails, injuries and other physical factors influence the interaction. A proper complete classification of user's characteristics must be included in the H-B interaction, covering the main biometric modalities.

Factors depending on the system. Accessible parameters regarding physical barriers (included in the EN 301 549) must be adapted to the H-B interaction. Also the device-system connection characteristics is important: wires and other connection characteristics may complicate the interaction.

Factors depending on the user-system interaction. The presence of operators or other people may help users or disturb them. In any case, this factor must be added as part of the environment.

Chapter 6

Enhancements to the H-B interaction methodology

Once the experiments have been explained and we have highlighted the preliminary improvements, it is time to suggest modifications and further improvements. Furthermore, several conclusions were extracted during the experiment's analysis and are summarized in this section. Next, we highlight the methodology's parts we consider should be modified in accordance with our experiments outcomes.

According to the H-B interaction structure, the modifications would not involve changes neither in the general model nor the evaluation model (ISO/IEC 19795 based), which are the base of the methodology. Reasons to not suggest changes in those parts are:

- The ISO/IEC 19795 is a mature standard, widely used and continuously validated. Even when it is performance-specific, provides means for accurately test several joint factors with usability assessment.

- The H-B interaction, based on the ISO/IEC 19795, was also validated. The general model is based on Common Criteria and therefore allows a proper analysis of targeted factors (evaluation condition specification –REC/TEC-).

Finally, the suggested modifications would affect these parts:

- *H-B interaction factors*
- *H-B interaction metrics*
- *H-B interaction fundamental requirements for executing*

6.1 H-B interaction factors

Modifications in the interaction factors are motivated by several reasons. First, the necessity of extend the methodology to the rest of biometric modalities involves the inclusion of new factors covering the new variables. Changes in the biometric uses motivate also new factors. Moreover, some of the common usability factors are not currently included in the H-B interaction. Then, next is a new factors classification proposal. The categorization remains the same as in the base methodology, split in 3, namely: User, System and User-System Interaction. To better understand this new classification, we separate the factors also in 3 different tables. In order to be backwards compatible, we keep the H-B interaction's table format.

6.1.1 H-B interaction factors depending on the user.

The new user-dependant factor's categorization is in Table 9. Coloured in blue are those modified or added factors. Changes on these factors are mostly related to behaviour and accessibility.

Behavioural

New behavioural aspects are mostly motivated by the experiments 2-4, where user's condition influenced the results. Stress, impatience, motivation and distrust are some of them. Also during the stress experiment, user's movement's variability (derived from the audio-visual signals) affect final results. We added also factors related to speech in order to cover also voice recognition, where the same variability in the trait presentation may be present.

Physical

We have slightly modified the classification of the physical factors, ordered now by fix/non-fix and artificial/natural for better understanding. The inclusion of accessibility concerns classification is motivated by all the experiments, especially by those with several interfaces or phases to complete. As we had users from the elderly group, we realise these users find more concerns when using biometrics (especially technology issues).

As long as new modalities are arising continuously, further factors could be added, so this table may be updated regularly.

6.1.2 H-B interaction factors depending on the system

Several changes are suggested in the system-dependant factors (Table 10). Coloured in blue are those modified or added factors. The most important are those related to interfaces.

Software

Factors related to accessibility and usability were added to the table. These factors are especially relevant in apps where the user interacts with the system autonomously. Accessibility concerns may be covered by following the current EN 301 549 standard, which is based on the WCAG 2.0 for assessing interfaces. Factors related to usability were defined by Norman Nielsen [57] and are directed to cover the interface's main flows in usability.

We noticed in most of the experiments (especially in the Experiment 4), mistakes when designing the interfaces in terms of fonts sizes, guidance within the screens or screen contrast among others. This should be overcame in future designs. The kind of process (local or remote) and the transparency to users (observed in the experiments where the samples were sent to servers and when users received feedback), influence also the interaction.

Hardware

As we carried out several experiments with ergonomics, we noted multiple hardware factors influencing the interaction: feedback in all the experiments, the elements disposition during the experiments 2 and 3, separable parts when using styluses and the devices wired connection to a computer during the Experiment 5. Hardware accessibility concerns are covered by the EN 301 549.

6.1.3 H-B interaction factors depending on the user-system interaction

This part was not too much modified and the changes made were mostly regarding the biometric trait presentation, which relays on the modality (Table 11).

Human-biometric capture device

Experiment 4 motivated many different user's reactions when interacting with the app during the stress phase. Then, variations when presenting the biometric trait have been added to the influential factors. Also the experiment's phase was demonstrated to be relevant during the experiments were enrolment and verification were split. System errors occur during all the evaluations (human or machine errors), then we added them to the Table 11.

Environment

Apart from the environmental conditions, the presence of other people during the evaluation motivates different reactions. We can observe this variation in most of the video recordings made during the first experiments. This factor was added to Table 11 also.

6.2 H-B interaction metrics

The set of metrics included in the H-B interaction methodology are the traditional usability measurements (ISO 9241) and various "ergonomic metrics". Modifications to the methodology were made in the way to better classify these metrics (according to physical or cognitive features), add the user acceptance (very relevant, as shown in Experiment 5) and classify some of the metrics as accessibility metrics:

- **Physical**

Percentage of test subjects that can use the biometric capture device.

- **Cognitive**

Percentage of test subjects that know how to use the biometric capture device.

Percentage of test subjects that learn how to use the biometric capture device.

Percentage of test subjects that remember how to use the biometric capture device.

The last two are indeed well-known usability measurements: learnability and memorability. The other two, may be classified as accessibility measurements: one related to physical accessibility concerns ("*subjects that can use biometrics*") and the other related to cognitive accessibility concerns ("*subjects that know how to use biometrics*"). User acceptance, which is very relevant in usability design and testing was also added to the metrics. Furthermore, as in the HBSI model, signal processing metrics are used (provided by ISO/IEC 19795). Though signal processing is not a usability measurement itself, it is a possible sign of usability flaws (e.g. poor performance may be motivated by wrong interactions). A new classification including all these modifications is in Table 12.

Further modifications to the HBSI are out of the scope of this Thesis, but may be cover when using the model in other modalities different from fingerprint. We demonstrated in Experiment 4 the necessity of modifying the model to be adapted to each evaluation specific features.

6.3 Fundamental requirements for executing

H-B interaction methodology specifies in detail several steps for executing a usability evaluation of biometrics. The only improvements to be done are related to previous evaluations that should be carried out prior to an evaluation with final products. This previous usability testing does not need a formal lab or hundreds of users. Nevertheless, it is necessary to have groups of potential users testing designs, so then, they require expending time and money. At these stages, usability testing consists of:

- *Users:* express opinions, participate actively and stay motivated
- *Designers:* take notes, record videos, suggest modifications, etc.

The most reliable usability evaluation methods [56] recommend to perform this prior techniques and they may be carried out during several phases of the biometric system design, though it is advisable to start since the very early stages. The better known are the following:

- *Walkthroughs of design concepts.* It is a meeting where the designers and the users discuss about the design concept and evaluate it with actual tasks. They are recommended to be carried out after each meaningful advance of the design.
- *Usability testing on low-fidelity prototypes.* Low fidelity prototypes are usually those made with paper and pencil. Then, many options may be analysed without investments. Users are encouraged to participate actively in this previous evaluation.
- *Usability testing cyclically as part of the UCD.* Results obtained from those test are the feedback for new improved designs. This idea has been explained at the beginning of the Thesis and applied for updating the H-B interaction.
- *Pre and post surveys.* User's opinions may change as a consequence of the user-biometric interaction. Surveys are a well-known way to monitor these changes. Those surveys are though to measure user acceptance and satisfaction. It is important to watch out for user's opinions/feelings before and after the assessments in order to check the impact of the system's usability. We have used both kinds of surveys to complete the Experiment 5.

Table 9. H-B interaction factors depending on the user. Factors in blue have been added or modified from the original version.

Type	Possible variations			Examples
Behavioural	Speech	Language		Dialect, accent
		Voice		Expression, intonation, volume, pitch
	Condition	Temporal	Emotions	Stress, tension, mood, fear, euphoria, motivation, social influence, acceptation, trust
		Permanent		Patience, intelligence
	Movements or changes	Voice		Misspelling, talk loud
		Physical		Facial expressions, poses, positioning
	Experience	Habituated		Use of biometrics
		Non-Habituated	With technical knowledge	Engineers, technical experts
			Without technical knowledge	Non technology users
Physical	Anthropometric data	Body dimensions		Tall, thin
		Physical features		Eyes color, hair color, language accent, human laterality
	Demographics	Age		Children, seniors
		Gender		Men, women
		Ethnic origins		Caucasian, Afro-Americans, mongoloid
		Occupation		Hand-works, under stress
	Interaction-influential	Fixed	Natural	Hair (beard, head, eyebrows, moustache), nails, birth marks, loss of voice, bruises, sties, allergies, reflex
			Artificial	Tattoos, piercings, clothes, contact lens, surgeries
		Non-fixed	Artificial	Piercings, glasses, handkerchiefs
			Natural	Sweeting
	Accessibility concerns	Temporal	Long term	Pregnancy, ictus
			Short term	Dizziness, vertigo, tiredness
		Permanent	Visual	Blindness, myopia
			Auditory	Deafness
			Physical	Limp, Tetraplegia
			Cognitive	Dyslexia, learning difficulties

Table 10. H-B interaction factors depending on the system. Factors in blue have been added or modified from the original version

<table>
<tr><th>Type</th><th colspan="3">Possible variations</th><th>Examples</th></tr>
<tr><td rowspan="18">Software</td><td>Accessible</td><td>EN 301 549 (WCAG 2.0)</td><td>Yes/No</td><td>Good/Bad contrast</td></tr>
<tr><td rowspan="11">Interface</td><td>Visibility of system status</td><td>Yes/No</td><td>Users know what happens next</td></tr>
<tr><td>Match between system and the real world</td><td>Yes/No</td><td>Familiar expressions</td></tr>
<tr><td>User control and freedom</td><td>Yes/No</td><td>Easily exit of unwanted states</td></tr>
<tr><td>Consistency and standards</td><td>Yes/No</td><td>Follow platform conventions</td></tr>
<tr><td>Error prevention</td><td>Yes/No</td><td>Confirmation options</td></tr>
<tr><td rowspan="2">Recognition rather than recall</td><td>Yes/No</td><td rowspan="2">Not remembering information</td></tr>
<tr><td>Yes/No</td></tr>
<tr><td>Flexibility and efficiency of use</td><td>Yes/No</td><td>Use of accelerators</td></tr>
<tr><td>Aesthetic and minimalist design</td><td>Yes/No</td><td>Not irrelevant information</td></tr>
<tr><td>Help users recognize, diagnose and recover from errors</td><td>Yes/No</td><td>Clear error messages</td></tr>
<tr><td>Help and documentation</td><td>Yes/No</td><td>Not long text</td></tr>
<tr><td rowspan="2">Process</td><td>Local</td><td></td><td>Face unlocking</td></tr>
<tr><td>Remote</td><td></td><td>Banking App</td></tr>
<tr><td rowspan="2">Transparency to user</td><td>Transparent</td><td></td><td>Airport cameras</td></tr>
<tr><td>Non-Transparent</td><td></td><td>Fingerprint in ABC</td></tr>
<tr><td colspan="4" style="display:none"></td></tr>
<tr><td colspan="4" style="display:none"></td></tr>
<tr><td rowspan="19">Hardware</td><td>Ergonomic design</td><td></td><td></td><td>Devices surface</td></tr>
<tr><td rowspan="3">Signals (feedback)</td><td>Visual</td><td></td><td>Lights, images, text, video</td></tr>
<tr><td>Acoustic</td><td></td><td>Beeps</td></tr>
<tr><td>Movements</td><td></td><td>Vibration</td></tr>
<tr><td rowspan="6">Elements disposition</td><td>Height</td><td></td><td>Wall, kiosk, turnstile</td></tr>
<tr><td rowspan="2">Orientation</td><td>Rotation</td><td>Fingerprint swipe sensor vertically or horizontally</td></tr>
<tr><td>Inclination</td><td>Different angles: wall or table</td></tr>
<tr><td rowspan="3">Location</td><td>Fixed</td><td>Palmprint sensor on a wall</td></tr>
<tr><td>Movable</td><td>Fingerprint traditional sensor</td></tr>
<tr><td>Embedded</td><td>Iris recognition in mobile devices</td></tr>
<tr><td rowspan="2">Sensor type</td><td>Biometric</td><td></td><td>Fingerprint sensor</td></tr>
<tr><td>Non-biometric</td><td></td><td>Camera, microphone</td></tr>
<tr><td rowspan="2">Separable parts</td><td>With separable parts</td><td></td><td>Digitizer with a stylus</td></tr>
<tr><td>Without separable parts</td><td></td><td>Face recognition camera</td></tr>
<tr><td>Condition</td><td></td><td></td><td>Brand new, used, dirty</td></tr>
<tr><td rowspan="2">Connection</td><td>Wired</td><td></td><td>Traditional fingerprint sensor</td></tr>
<tr><td>Wireless</td><td></td><td>Wi-Fi cameras</td></tr>
<tr><td>Accessible</td><td>EN 301 549</td><td></td><td>Easy to grasp devices</td></tr>
</table>

Table 11. H-B interaction factors depending on the user-system interaction. Factors in blue have been added or modified from the original version

<table>
<tr><th>Type</th><th colspan="3">Possible variations</th><th>Examples</th></tr>
<tr><td rowspan="6">Human-biometric capture device interaction</td><td rowspan="6">Presentation of the biometric characteristic</td><td>Translations</td><td></td><td>Palm not centred on the sensor</td></tr>
<tr><td>Rotations</td><td></td><td>Fingerprint rolling</td></tr>
<tr><td>Voice intensity</td><td></td><td>Loud voices</td></tr>
<tr><td>Blinking</td><td></td><td>Iris recognition</td></tr>
<tr><td>Shift the gaze</td><td></td><td>Face recognition</td></tr>
<tr><td>Displacements</td><td></td><td>Gait recognition long step</td></tr>
<tr><td rowspan="14">Human-biometric system interaction</td><td rowspan="5">Guidance</td><td>Without guidance</td><td></td><td>Non-guidance</td></tr>
<tr><td rowspan="3">Non attended guidance</td><td>Visual guidance</td><td>Poster, pictograms</td></tr>
<tr><td>Audio guidance</td><td>Sounds</td></tr>
<tr><td>Audio visual guidance</td><td>Video</td></tr>
<tr><td>Attended guidance</td><td></td><td>With attendance</td></tr>
<tr><td rowspan="2">Training</td><td>With training</td><td></td><td>Users receive instructions about the use of biometric system</td></tr>
<tr><td>Without training</td><td></td><td>Users do not receive instructions about the use of biometric system</td></tr>
<tr><td rowspan="4">Feedback</td><td>Without feedback</td><td></td><td>Biometric system without display, lights</td></tr>
<tr><td rowspan="3">With feedback</td><td>During the process</td><td>The system indicates to the user to move forward for presenting the biometric characteristic</td></tr>
<tr><td>At the end of the process</td><td>The system provides guidance after it is not able to capture the biometric sample</td></tr>
<tr><td>Both</td><td>A system that includes both types of guides</td></tr>
<tr><td>Phase</td><td></td><td></td><td>Training, enrolment, identification</td></tr>
<tr><td>System errors</td><td></td><td></td><td>Capture, algorithms, app, devices, operator, user</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td rowspan="4">Environment</td><td>Environmental conditions</td><td></td><td></td><td>Temperature, humidity, illumination, noise</td></tr>
<tr><td rowspan="3">Population</td><td>User alone</td><td></td><td>Face unlocking smartphone</td></tr>
<tr><td rowspan="2">User Accompanied</td><td>Operator</td><td>Enrolling in a system</td></tr>
<tr><td>External</td><td>Fingerprint in the gym</td></tr>
</table>

Table 12. Suggested new distribution of H-B interaction metrics. Metrics highlighted in blue have been added or modified.

Type	Variables	Definition
Usability	Effectiveness	Errors, assistance actions, tasks completion
	Efficiency	Time spent in the processes
	Satisfaction	Degree of users satisfaction
	Learnability	User who learnt how to use the system
	Memorability	User who remember how to use the system
	User Acceptance	Willingness to use the system
Accessibility	Physical	Subjects that can use the system
	Cognitive	Subjects that know/learn/remember how to use the system
Signal Processing	Biometric sample	Quality metrics
		Time to capture
	Processing capability	Number of segmentation errors
		Number of features extraction errors
		Segmentation time
		Features extraction time

6.4 Conclusions

Further validation of the H-B interaction has brought changes and modifications in the methodology. Each experiment has motivated improvements, which encourage us to plan other evaluations. Future experiments could include prior usability testing to better analyse the impact of those tests in biometric systems. Moreover, experiments using different modalities could motivate new changes in the methodology.

H-B interaction, as well as HBSI, give more priority to performance than to final users. This could lead to reliable systems which people do not like to use. Then, usability analyses should be focused on users, who will use systems in the end. In our experiments, we have put more attention on user's acceptance and satisfaction, gathering several videos, interviews and notes. This has given us a new perspective of usability in biometrics, very useful for future designs.

Regarding the outcomes obtained from the experiments, maybe the most valuable (and the most difficult to measure) are the user's opinions and reactions. As long as usability testing involves subjective measures, usability evaluations tend to be challenging but rewarding: improvements in usability have effects on final performance and benefits users and designers.

Part IV: Accessibility in Biometrics

Chapter 7

Improvements of citizen accessibility by the use of biometrics

We strongly believe that biometric recognition could help people with accessibility concerns to perform common tasks requiring authentication (e.g. ATM transactions, border crossing, etc.). Having this in mind, we have decided to deeply analyse accessibility in biometric recognition systems and therefore, in the H-B interaction.

In this Thesis we have improved the H-B interaction including accessibility concepts at user and system levels. Once this was accomplished in Chapter 6, the accessibility is also guaranteed during the user-system interaction. The modifications made have been the inclusion of the EN 301 549 and implicitly, the inclusion of the WCAG 2.0. Further modifications are the inclusion of accessibility metrics and the re-adjustment of various traditional usability terms.

The study of accessibility in biometric recognition systems is motivated by our ongoing research about usability in biometrics in the GUTI (Grupo Universitario de Technologías de Identificación – University Group for Indentification Technologies). First ideas came when we contacted groups of people with accessibility concerns: the IMSERSO (Instituto de Mayores y Servicios Sociales - Institute for Older Persons and Social Services) and the CRMF (Centro de Recuperación de Minusválidos Físicos - Recovery Center for Physically Disabled People). We have focused on making improvements in their daily common tasks. We also noticed that not too much effort have been carried out regarding inclusive designs in biometric recognition. Other important point is that the suggested updating of the H-B interaction requires the validation of the new measurements and testing methods. Then, we apply the new updates (EN 301 549 and the well-known usability metrics) to evaluate accessibility in further experiments.

During the next sections, the steps followed to reach a final accessible design of a biometric system are explained. Afterwards, two experiments validating the EN 301 549 and the well-known usability metrics in such design are shown. The methodology to evaluate the accessibility is again the H-B interaction and the notes about the experiments cited at the beginning of Chapter 5 apply also to these experiments. Then, we continue with the experiments numeration (i.e. next experiment is Experiment 6).

7.1 Practical improvements in daily tasks and EN 301 549

This research line is motivated by the aim of ease the life of people with accessibility concerns via biometrics. Supported by the CRMF, the IMSERSO and the Spanish national project URBE, we have been developing biometric systems specially designed to be accessible and evaluating them with real users. First of all, we have designed a prototype for accessing ATMs and making payments. Next, we explain the technological approach, the requirements and design specifications. Finally, we show the experiments carried out to evaluate our design.

This kind of tasks (making payments) are common and easy for most of the population, but brings enormous challenges to citizens presenting certain kinds of accessibility concerns. These challenges provide not only inconveniences for the person, but also, in many cases, a complete loss of the security. For example, when a person has limited mobility and has to use a wheelchair, he is not able to access the ATM, and has to ask somebody to insert the card for him, and also to type the PIN code. Furthermore, that person has to take the money, and return everything to the user with accessibility concerns. If that person is not faithful, it is clear that not only the money, but also the whole access to the banking account is compromised.

The suggested solution comes to provide adaptable solutions for the needs of disabled people to be integrated into current society, with the same level of security, both for the citizen and for the system, than with non-disabled citizens. Current solutions usually provides a parallel, more unsecure way, to access the same services for disabled people, which usually represent a security back door for the whole system. In some cases, the potential existence of such a back door lead system designers to deny the use to disabled people. The proposed scenario is based on the authentication of the user's identity by means of a portable device (Smartphone, PDA, tablet, etc.) jointly with the security of smartcards and contactless communications. The mobile device is used as a personal device for capturing and processing the biometric data, whereas the smartcard carries out the comparison process against the biometric template within a personal context, secure and reliable.

7.1.1 Technological approach

In order to provide a solution for improving the universality of authentication systems, the combination of several technologies is considered. This section briefly introduces each of them, and also comment their potential applicability.

7.1.1.1 Biometrics

From the most known biometric modalities, this research line is focused on fingerprints, face, iris and handwritten signature. This is motivated by their easiness of use, their transparency (specially face and fingerprint) and the familiarity for users (handwritten signature). Fingerprints present the advantage of being one of the most studied and tested modalities, with the availability of a wide variety of sensors and being well accepted. It also requires a medium-low level of interaction, as only positioning of the finger on the sensor is demanded. In addition, the user present as many identities as fingers are available, so he/she can choose different fingers for different services. From the disadvantages, it can be highlighted the need of interaction and the size of the sensor when high quality ones are required to avoid problems with different skin conditions (i.e. using optical sensors instead of semiconductor ones).

Face recognition is quite simple, require little interaction and can be acquired at a reasonable distance. Therefore its potentiality to be used for gaining accessibility is extremely high. Even sensors can be really cheap and small, so their integration is viable in large variety of solutions. Unfortunately there are some inconveniences. This biometric modality present error rates higher than the ones traditionally demanded by certain services. In addition, facial recognition suffers from lack of robustness against certain environmental conditions, particularly with background illumination. Therefore, it is used in open areas may rise the error rates to a non-admissible level.

Iris recognition present excellent low error rates and it is very robust against fraudulent attacks. The last generation of sensors are even capable to acquire samples at reasonable distances, and the user present up to two different identities, one for each available eye. But iris sensors are expensive and high enough as to not being viable to consider their installation at the hand of the disabled person. Therefore the sensor should be installed at the point of service (e.g. the ATM) and a secure communication shall be used between the sensor and the device belonging to the disabled citizen.

Finally, handwritten signature is a way of authenticating and acknowledging contracts and financial transactions, which is common to most of the population, as it has been used for decades. The sensor can even be the touch screen of a smartphone and the error rates achieved are within reasonable levels for most applications. But the level of interaction may be excessive for those users with reduced mobility on the upper limbs.

7.1.1.2 Smartcards

Smartcards are identification cards that embed an integrated circuit (i.e. a microcontroller) that manages the information stored in its memory and several robust security mechanisms as to control the access to that information [91]. The microcontroller executes all its functionality based on the Smart Card Operating System (SCOS) installed in its non-volatile memory. The security in its manufacturing, plus the security mechanisms included in the SCOS allows naming smartcards as tamper-proof devices. They are even used for executing the security protocols in the communication between point of services and financial institutions. Therefore the smartcard provide a way to ensure the privacy of the personal data of the citizen, and with a reasonable implementation of the system, guaranteeing the security of the authentication system. It is also possible that in addition to the internal verification of PIN codes, they can provide the internal comparison of the biometric sample sent to the card, with the biometric reference previously stored securely in the smartcard. This kind of solution is called On-Card Biometric Comparison, or also known as Match-on-Card [92].

7.1.1.3 Short-range wireless communications

One of the first problems that can be addressed is to provide a solution to those citizens that cannot physically access the deployed system (e.g. the ATM). A way to solve this is to motorize the whole system so that it can be moved and adapted to the need of the user. That kind of solution may be too expensive, but not only in money, but also in the area needed. For example an ATM could be developed so that it can use the whole height of wall where it is installed, and provide embed in the wall the whole set of mechanisms needed for the adaptation. But in a shop, the point of service should be as small as possible and the shop assistant should not be bothered with an excessive interaction with the system. Therefore in such a case a robotized solution is not viable.

Another approach is that the whole system is split in two parts, the current one (e.g. the ATM or the point of service), and a secure personal device that is in the range of the user (e.g. an embedded system installed in the wheelchair of the citizen). Then there is the requirement of having a secure wireless data link between both parts, and such data link should also be in a reasonable short range, so that man in the middle attacks could be minimized.

Currently there are a lot of short-range wireless communication solutions that can be used, having each of them particular characteristics. WiFi is a connection that could not be considered as short-range, but better as mid-range. Also, the communication is open to everybody, although data can be ciphered. Bandwidth is enough for most of the applications, but the time needed to establish a connection may not be as fast as required in some applications. Other well extended short-range communication protocol is Bluetooth, which is already included in all smartphones and most of the tablets. Both, the bandwidth and the security of the communication are not the best in the market, but the traditional range of 10 meters are more than enough for most of the applications related with this work. All these communication protocols, when using a smartphone, will require that the information is managed by the smartphone operating system, which may be infected by malware.

But in the last years, the extension of RFID to provide solutions for payments, have created the specification of Near Field Communication, mostly known as NFC [93]. If this communication has already been approved by financial institutions, then it could mean the perfect solution to provide a secure data link between the system and the citizen device. One of the most powerful options of NFC in mobile devices, is that the smartcard can be physically linked to the NFC antenna of the device, and therefore the NFC communication can be done directly with the smartcard, avoiding security leaks from the mobile device operating system. The major drawback is that current implementations of NFC require a very short range communication, even requiring that both devices touch themselves, which would reduce the applicability of this technology for improving accessibility.

7.1.2 Specification of system requirements

With all the information stated in the previous sections, a set of requirements for accessible authenticated systems can be specified. There are a set of requirements that can be considered as generic (i.e. not application specific). These generic requirements are:

• Systems shall not be defined with a single user interface, neither allowing only one single way for authenticating, as different people may present different requirements.

• Systems shall provide user with feedback in a variety of ways, i.e. not only considering visual feedback, but also auditory signals.

• A system that require authenticate users, will, in some moment in time, need to know the claimed identity of the person. At that same moment, if not earlier, the disabilities of that user could be communicated to the system, as to allow the system to better adapt to the needs of the user, and not providing information in a way that the user would not benefit from it, and that could be used by others for fraudulent use.

• For assuring adaptability to all potential cases, it may be a better approach to interact with an interface device belonging to the user, and which is fully adapted to the needs of that user. When such personal interface is in use, the generic interface shall be stopped as to not provide personal information to other people.

• The Personal Interface Device (PID) shall allow its discovery by the system as soon as possible, and in a way that it does not limit the interaction with the user. In the ideal case, this shall happen at the time the user approaches the system.

• All data and interface communication shall be strongly secured, to assure the privacy of the citizen.

When applying this generic requirements to a specific application, some further requirements appear. In order to cover them, it is important to first define the application targeted, and then specify the requirements.

7.1.2.1 Target application

One of the most sensitive situations that disabled people have to face nowadays are the handling of financial products such as banking cards. With these cards, they could be able to withdraw money from their accounts and also perform the required payments. In this research two common situations are considered. The first one is interacting with an ATM, while the second one will be using a Point of Service (PoS) at a shop or restaurant, to pay the bill.

It is clear that there will be additional considerations to be done in certain aspects, such as how to handle the withdrawn money to a mobility disabled person. Those cases are considered as future work lines.

7.1.2.2 Specifications for ATM operation

There are several characteristics from ATMs that impact the definition of specifications, some in a very positive way, while others in providing serious constraints. For example, the ATM is a big and expensive piece of equipment. This allows the integration of all kind of authentication methods, as cost and volume are not major issues. An ATM is also a powerful computational machine, and both strong security mechanisms and processing capabilities are available. Finally, an ATM is installed with a fixed and reliable connection to the service provider and electric power is all-time guaranteed.

But on the other hand, the ATM is installed in a fixed position, not allowing its mobility. Also, the ATM usually handles more sensitive operations than a PoS (i.e. while the PoS only perform payments that can always be audited, the ATM can withdraw money for later non-auditable payments, or even manage account information and operations). Finally, an ATM is usually a non-attended device, which does not allow that an operation could assist the disabled people. Therefore the following specifications as stated:

• The ATM shall be installed in a place that do not limit the access to all different kind of users, including those requiring wheelchairs and allowing interacting with the system with any of the two parts of the body (i.e. not designing it to be only usable for right handed people).

• The ATM shall be continuously polling the environment to detect if a PID is approaching. The PID should be detected when the user is up to 2 meters from the ATM.

• As soon as a PID is detected, the ATM shall communicate with it as to obtain the interface requirements of the user, adapting its functionality to such interface device.

• Biometric authentication shall be requested, considering several modalities and different locations of the sensor. As many sensitive operations can be performed in an ATM, strong biometric authentication is required. For example, iris recognition can be considered for many users, with iris detection at a distance and in movement. Also facial recognition shall be included, both, for authentication, and for fraud resistance. The use of other biometric modalities shall be referred to the PID.

• The ATM shall ensure that all communications are performed in a secure way, after performing internal and external authentication with the PID.

• Multiple angle video recording shall be installed for auditing purposes, particularly if there are operations where additional people could be involved (e.g. physically getting the withdrawn money to the disabled person).

7.1.2.3 Specifications for PoS operation

In the case of a PoS the specifications changes considerably. A PoS is being operated in a shop or a restaurant, so its size shall be small, and in most of the cases it should be operated in several locations (e.g. at the dining table of the user). Therefore a PoS is usually attended by an operator, present mobility, may face power consumption limitations, and may not be able to perform complex processing. Also, PoS may be limited in the devices that they can integrate (e.g. biometric sensors). Obviously PoS are by far cheaper than ATMs, so the addition of more devices may impact seriously in the overall price. And in a PoS the security is usually managed by an internal smartcard.

These characteristics have a direct impact in the specifications of PoS operation. Most of the disabilities can be handled by current PoS, particularly if they are mobile devices. The operator (e.g. shop attendant) can handle the PoS to the location the user finds more suitable, and can assist him/her in the operation. But there are still several cases that shall be addressed:

• Feedback from the PoS is usually given only in as visible information, so the sight-related disabled people has to rely on what the operator is telling them. Auditory feedback, if done openly, may also affect the privacy of the user. Therefore, either the PoS provides a headphone jack, or it connects to the PID to handle the feedback.

When using a PID, as PoS might be mobile, the distance between the PoS and the place where the antenna of the PID is placed, can be minimized, even allowing that they touch themselves.

• It is also important that the PoS provide feedback for the shop attendant, as he/she will have to validate the whole process.

• In a PoS, the authentication requirements can be a little bit lowered to those in an ATM. In this case, it is more important that the authentication mechanisms are able to adapt, not only to the user needs, but also to the retailer ones. Due to size and cost, iris recognition is not considered as viable in a PoS, but other biometric modalities, such as fingerprint or handwritten signature, can be used by a big portion of disabled people. This is especially important if the biometric authentication, and/or the PIN-code typing is done at the PID.

• But as in ATM, the whole communication shall be secured to ensure the citizens privacy and the authenticity and integrity of the transaction.

7.1.3 Characteristics of a proper design

After the specifications have been defined, the design of a prototype to set a first proof of concept began. Two system prototypes have been designed, one for the ATM and another for the PoS. Both have used the same set of PIDs, as to analyse the interoperability among them and also the feedback received from users suffering different kinds and levels of disabilities. When starting the design, some decision were taken for this first approach:

• The PID will be an Android smartphone. Initially it has been considered the Samsung Galaxy Note 2, due to its computational power and the number of peripheral and services available.

• Security will be handled by the SIM card of the smartphone, which has been substituted from the original operation SIM card to a JavaCard powered one with NFC support.

• Exchange of information will be done using NFC, although this will mean that a future work will consist on improving communication distance for the ATM case.

• Biometric sample acquisition has been designed to be performed by the ATM in the case of iris and face, with the smartphone touchscreen in the case of signature, and with an optical USB sensor connected to the smartphone via OTG, for fingerprint.

• In all cases, biometric processing is done at the smartphone and biometric comparison is done inside the JavaCard.

• To emulate the ATM, a personal computer has been used with a program programmed in C# and .NET. The PoS was designed and developed using an Android tablet. The PIDs designed include:

• A PID with authentication via iris and fingerprint (one or the other depending on the user preferences) for ATMs and the authentication using fingerprint and/or handwritten signature for PoS. This device is targeting those users with may experience any kind of disability, but are still able to move the upper limbs (and fingers). This device has been developed with both, visual feedback, but also acoustic feedback via the headphones for the visually impaired users. It is important to notice that the acoustic feedback covers the aid in the alignment with the ATM, and he aid in the signing process as not to allow them to sign out of bounds.

• A PID with authentication via iris and/or face for ATMs and fingerprint and/or PIN for PoS. This is intended for those users not able to move their upper limbs, either totally or partially. In the PoS, the operation can help in the positioning of the finger on the sensor, and the introduction of the PIN-code can be done by a variety of means already designed for this kind of disabilities. PIDs have been designed in a way to cover multi-lingual and multi-cultural differences, and where users with cognitive disabilities are expected to use them, although this will only be checked once the system in placed under text with real users.

7.1.4 EXPERIMENT 6: Accessibility Evaluation in PoS for Elderly – EN 301 549

This experiment [94] is the first application of the suggested design with real users under a scenario evaluation. Next, it is described the design, development and evaluation (both accessibility and performance) of a mobile app for making payments in PoS using fingerprint and handwritten signature recognition. All the directives of design have been taken from the previous section.

The app has been developed using a UCD approach, with the aim of adapting a biometric recognition solution to those users with accessibility issues, more specifically elderly persons. We have prepared a scenario for making fictitious payments in a PoS with a smartphone using biometric recognition and NFC to authenticate users. Users place the smartphone close to the PoS (a Mifare smartcard) to pay and then the authentication is made in the smartphone through fingerprint or handwritten signature recognition. Finally, users end the process placing the smartphone close to the PoS again. An example using fingerprint recognition is shown in Figure 56.

Motivation of the experiment

This experiment is the first application of the EN 301 459 to a biometric recognition system. Our plan was to design a realistic product to be used in a realistic context, following the idea of improving the life of people with accessibility concerns. We focused on elderly people because they are the group who found more issues when interacting with biometrics during the previous experiments. The context is in the line of the other experiments: mobile devices and DSV. Furthermore we added fingerprint to increase the number of users who could use the system. Then we have one static and one dynamic modality. By using two modalities we expect to increase the range of users who can properly use the app. For instance, users with Parkinson disease cannot generally use handwritten recognition properly, but they could probably use their fingerprint. In the same way, users without fingerprints could sign. Moreover, the EN 301 549 requires the use of at least 2 different enough biometric modalities.

Influence on the H-B interaction

With this experiment, we validate the main accessibility updating made to the H-B interaction, the EN 301 549. All the standard clauses are defined and then, all the requirements which apply studied for this specific case. Therefore, this an example of the standard application from the beginning to the end, where we show tables including both the satisfied and the unsatisfied requirements.

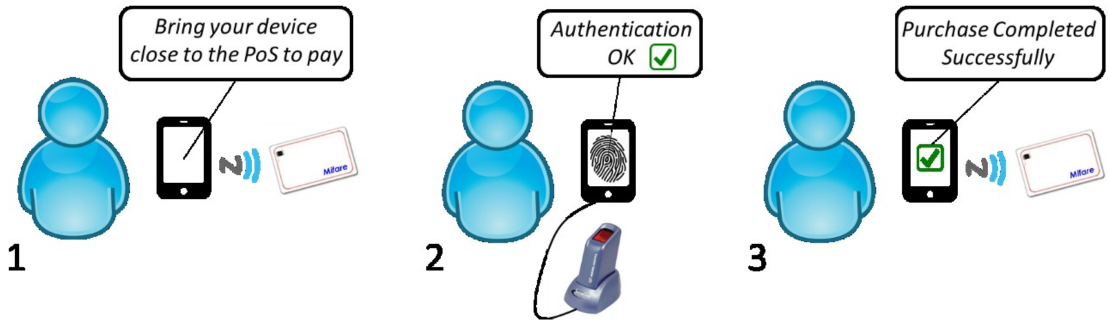

Figure 56. Example of app's use. The communication starts in 1, then in 2 is the biometric recognition. In 3 the fictitious payment is made. The smartcard acts as the PoS and communicates with the smartphone through NFC

Once we had developed the system we have made a checklist-based accessibility evaluation (following EN 301 549) and a performance testing with elderly users. Our main intention is to show up how a common biometric recognition product fails when providing accessibility and how to fix these concerns with the application of the standards.

7.1.4.1 Evaluation set-up

This section describes the evaluation protocol, the participating users, the hardware and software used for evaluation, and the functionality of the system.

Evaluation protocol

The evaluation process was divided into 2 sessions, 2 weeks apart, so that users would not feel tired or uncomfortable. An operator was always present to assist users and complete their personal data in case they did not know how to proceed. Once the personal data acquisition is completed the operator would only intervene in case of doubts or inconveniences.

Users

The usability testing has been focused on elderly users. We have chosen this specific group of people because biometric recognition is mainly used to make procedures easy (guaranteeing security), so it could be highly useful for them who may find even more cumbersome to remember passwords or carry tokens, smart cards, etc.

A total of 85 elderly people went to a previous explanation meeting and 40 of them initially agreed to participate. Before starting the evaluation 2 users left due to distrust so 38 users started the 1st session. During the experiment 5 users did not attend to the 2nd session and 5 other could not complete the process because their fingerprints were impossible for the system to read. Finally 28 users completed the evaluation. Half of them are between 60 and 70 years old, 42,8% between 70 and 80 and 7,2% are over 80. 75% are women and 25% are men. Regarding their jobs (before retirement) 16 of them worked in an office and 12 worked with their hands.

Hardware

The participants interacted with an app installed in a Samsung Galaxy Note II GT-N7100 which incorporates a proprietary stylus. The fingerprint sensor used is a capacitive SecuGen Hamster Plus with a resolution of 500 dpi. A Mifare Ultralight, which allows a contactless data transmission up to 10cm, was used as a PoS to interact with the smartphone through NFC. The real scenario is shown in Figure 57.

Software

The evaluated application was developed under Android 4.0.3 and the accessible characteristics have been designed in accordance with the “Green Book of Accessibility” (originally "Libro verde de la accesibilidad") [95] which deals with the dissemination and analysis of the accessibility and also provides best practices.

Figure 57. The scenario includes the PoS, the smartphone with the fingerprint sensor connected and the stylus for signing.

Some of the accessibility-related features of the application are: low number of colours, high contrast between foreground and background colours, big font sizes, simple interfaces, easy app flow, audio and text relay. As described above, the system was used in two sessions. During the first session the user accepts the experiment conditions, provides personal data, choses preferences for the user interface, completes the training and the enrolment, provides the biometric samples and makes the purchases. In the second session the user completes the training if needed, provides biometric samples, makes the purchases and finally fulfils a satisfaction form. In summary, there are four main tasks in the user interface: training, enrolment, data collection and purchases. These tasks are described below. The evaluation roadmap is in Figure 58.

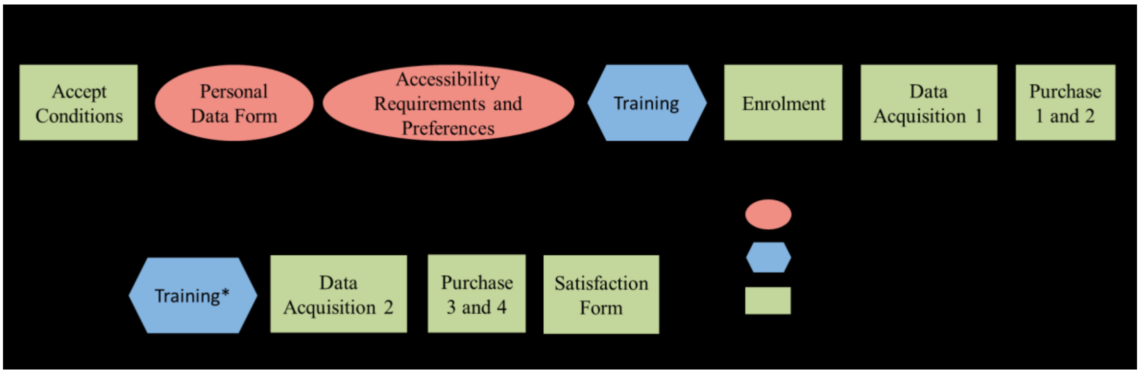

Figure 58. Roadmap of the evaluation

<u>Training</u>

The training was included to help users in the first steps of interaction with biometrics and it was especially programed for those who have not skills interacting with ICT products. During the training users are intended to gain skills signing and using the fingerprint sensor (training can be repeated as many times as desired). An example of the training interface is shown in Figure 59.

<u>Enrolment</u>

After the training the user is required to enrol his fingerprints and signatures. The samples acquired during enrolment are the user's templates for later comparisons. The app requires 2 fingerprint presentations of the same finger to be similar (over a prefixed threshold explained afterwards) to complete the fingerprint enrolment. If the 2nd fingerprint is not similar enough to the 1st the user has to repeat the presentation. Index and middle fingers of each hand are available to be selected or deselected (just in case the user is not able to use some of them).

The handwritten signature enrolment starts after the fingerprint enrolment. The user is encouraged to sign at least 5 times in a blank square with the possibility to delete a signature anytime if he/she is not happy with it. If the signature is made totally or partially out of bounds a message box with a warning is launched and the user has to repeat that signature. To obtain the required 5 signatures the algorithm compares each one with all of those previously acquired, starting with the second one (2nd vs 1st; 3rd vs 2nd and 1st; 4th vs 3rd, 2nd and 1st; and 5th vs 4th, 3rd, 2nd and 1st) and returns the best result. This matching is performed in order to check if the user is making a different signature or the acquisition process has not captured a significant number of sample points. In accordance, the application shows a message with the result: "*Correct signature*" or "*Wrong signature, please repeat it*".

<u>Data collection</u>

In this part users have to provide 5 correct fingerprints of each finger and 10 correct signatures. These are the samples to be matched with the templates (obtained in the enrolment) to obtain performance results. This process is included in session 1 and session 2, therefore each user has to complete 10 fingerprints of each finger and 20 signatures by the end of the evaluation. In this phase the app gives visual and audio feedback according to the algorithm results: a green tick if the matching against the template is over a threshold and a red cross if the matching is below that threshold. This feedback is the same for both modalities.

Purchases

This phase is the last in each session and consists of making 2 fictitious payments (4 in total) on a PoS using the fingerprint or the signature (user's decision) and NFC for the information transmission. The app asks the user to put the smartphone close to the PoS, then the user is required to present his biometric characteristic first and finally to put the mobile again close to the PoS in order to complete the purchase. The app reports the result of the matching with visual and audio feedback (the same as in the data collection phase) and the purchase is not completed if the matching is not successful.

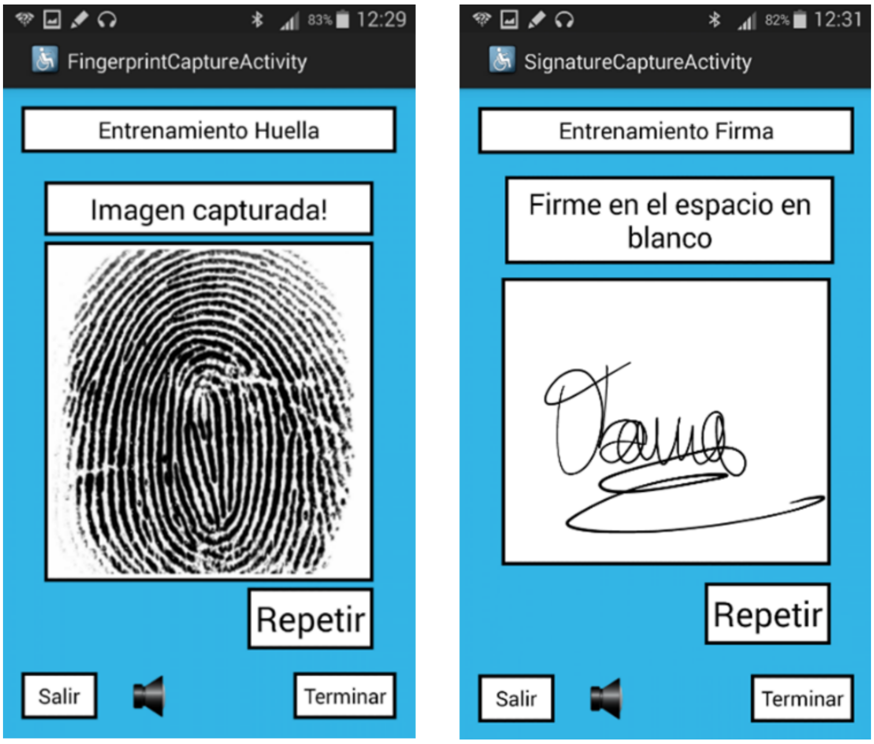

Figure 59. Example of the training interfaces. The other interfaces have the same design but different text messages. Text translation (Spanish), from up to down and from left to right: First image: Fingerprint Training / Image captured! / Repeat / Exit / Finish. Second Image: Signature Training / Sign in the blank space / Repeat / Exit / Finish.

User interface

The app user interface was designed with the aim to provide accessibility to elderly people, a group of people who typically have several combined impairments: low vision, hearing problems, shivering, memory loss, etc. Therefore, the user interface is configurable and users can select the screen brightness level (low, medium or high) and the volume level of guidance messages (low, medium or high). The app remembers the settings during the evaluation, though users can modify them at any time.

The interface itself is austere showing just the indispensable (short text messages and buttons) and the colours have been chosen according to the Green Book of Accessibility guidelines. Users are able to select between 3 colour combinations: (1) blue background screen with white background buttons, (2) white background screen with blue background buttons and (3) black background screen with white background buttons. Figure 60 shows these three combinations.

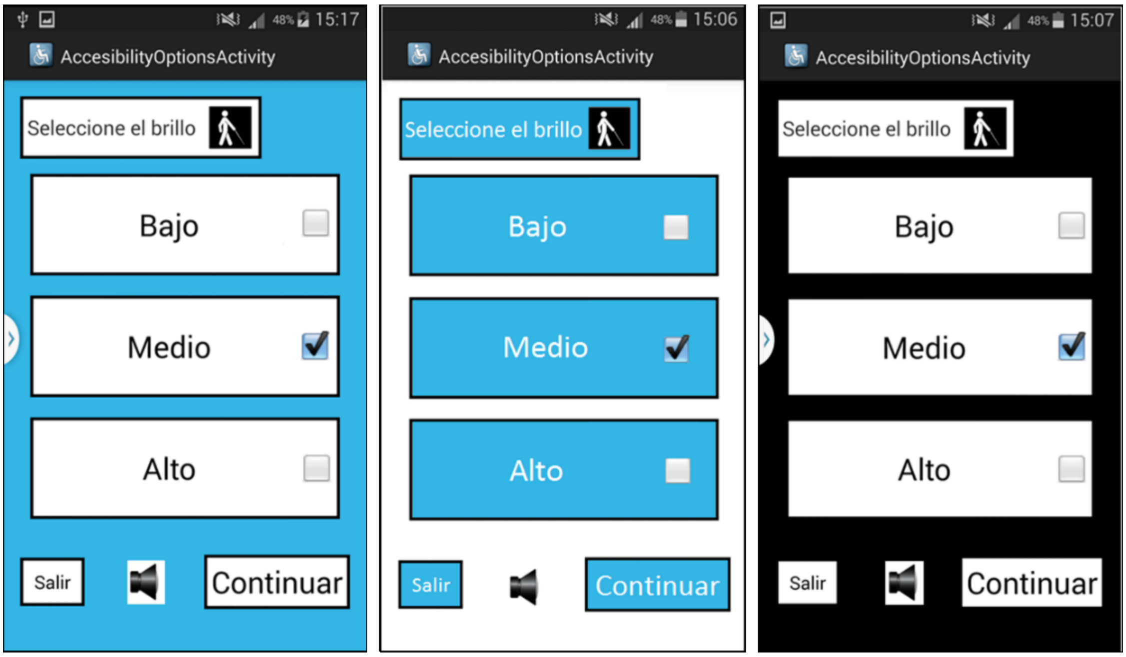

Figure 60. App settings interfaces. The three colour combinations in the brightness selection screen. Text translation: From up to bottom and from left to right: Select the brightness / Low, Medium, High / Exit / Continue).

<u>Algorithms</u>

We have used biometric recognition algorithms during the evaluation to provide feedback to users, make the payments and to guarantee high confidence between samples. They have been also applied to obtain performance results once the capture process had finished.

Fingerprint recognition

We used the NIST MINDTCT (template extractor) and Bozorth3 (matcher) [90] for fingerprint recognition. The implementation we used returns a matching result between 1 (worst) and 481 (best), and the threshold used is 30 (i.e. a result below 30 is considered an impostor and 30 or above is considered genuine). The threshold is quite low in order to avoid too many rejections during the evaluation and therefore a user's bias: when a user receives several rejects, he or she may feel disappointed or frustrated. Users tend to proceed more fluently when they receive positive feedback [96].

Handwritten signature

A DTW-based algorithm was used for handwritten signature recognition (same as in previous experiments) and the X and Y time series coordinates were used as inputs. The implementation developed returns a matching result between 0,00 (the worst) and 5,00 (the best). In this case, the threshold was fixed in 4,00 due to this algorithm has returned very good results in previous works [48].

7.1.4.2 Accessibility evaluation

Authors believe that accessibility is an important factor to enable the widest adoption of biometric recognition systems by providing adequate user experiences for most users.

The app described here has been designed and developed following accessibility guidance. In this section the accessibility evaluation of the resulting app is shown. Unfortunately, there are not specific accessibility standards for biometric systems published by ISO/IEC/JTC1/SC37 Biometric Recognition. However, as our app is an ICT system, the recently published European Standard EN 301 549 can be applied to determine its degree of accessibility.

<u>EN 301 549 Application</u>

This section summarizes the parts of the EN 301 549 in order to better understand which of them may apply to our app. As a definition of the EN 301 549 is in Section 2.3., next all its clauses, starting from clause 4 are described in detail (clauses 1 to 3 are the scope, definitions and abbreviations):

Clause 4: This clause describes the needs of persons with disabilities, in the form of functional performance statements. These statements are intended to describe the functional performance of ICT enabling people to locate, identify, and operate ICT functions, and to access the information provided, regardless of physical, cognitive or sensory abilities. The requirements and recommendations of the EN are defined in clauses 5 to 13.

Clause 5: In this clause are the generic requirements, which are generally applicable to all ICT products (e.g. non visual access, auditory output, speaker volume, key repeat, etc.). Particularly interesting for this Thesis is the clause 5.3 Biometrics:

"*Where ICT uses biological characteristics, it shall not rely on the use of a particular biological characteristic as the only means of user identification or for control of ICT. Biometric methods based on dissimilar biological characteristics increase the likelihood that individuals with* *disabilities possess at least one of the specified biological characteristics. Examples of dissimilar biological characteristics are fingerprints, eye retinal patterns, voice, and face.*"

Therefore, the only requirement for biometric recognition systems is to offer different enough modalities covering dissimilar biometric characteristics.

Clause 6: This clause deals with ICT products including two-way voice communication (e.g. including voice calls).

Clause 7: Requirements related to ICT with video capabilities are in this clause.

Clause 8: This clause is related to hardware, including hardware with speech output, physical accesses, mechanically operable parts, etc.

Clause 9: In this clause are Web requirements (e.g. captions, colours, text, etc.) and most of them shall satisfy WCAG 2.0.

Clause 10: Refers to non-web documents, including those which: are not web pages, are not embedded in web pages and are embedded in web pages but not used in the rendering or that are intended to be rendered together with the web page in which they are embedded.

Clause 11: Refers to software, including platform software, authoring tools, software that operates as assistive technology and software that provides a user interface including content that is in the software. Many of these requirements point to others in Clause 9, so then, may satisfy WCAG 2.0 also.

Clause 12: This part is related to documentation and support services.

Clause 13: The last clause is about ICT providing relay or emergency service access (e.g. text relay services, sign relay services, etc.).

The EN includes 11 user's needs (functional performance statements), 213 requirements and 26 recommendations. Given the feature-based structure of the EN, the text of each requirement and recommendation starts with a pre-condition that states when the requirement or recommendation applies. Even so, the large amount of clauses makes non-trivial the task of determining which requirements of the EN may apply to any given ICT product. To this purpose, a decision tree-based approach has recently being defined [97] [70] to simplify that task.

Applicability of accessibility requirements to the biometric system

The decision tree consists of a set of questions that are used to get information about the features provided by an ICT system and to determine which sets of requirements are applicable in principle to that system. In the case of this biometric system, the following main features were found:

- *The system has operable parts (hardware and software-based). That means that clauses 5.5 (operable parts) and 5.6 (locking or toggle controls) of the EN apply.*
- *The system includes hardware components (the fingerprint sensor), but the hardware does not provide speech output and the hardware does not have physical dimensions that are integral to the ICT and that may restrict physical access. That means a subset of clause 8 (hardware) applies.*
- *The system has a software component that provides a user interface, so most content of clause 11 (software applies).*

In addition, there are generic requirements under clause 5 that apply to most systems. The resulting list of clauses that contain requirements applicable to the biometric system is shown in Table 13. Each of the chapters listed in Table 13 contains several requirements and recommendations, which have been analysed to further specify their applicability.

The use of the question-answer software has decreased significantly the time consumed and has ensured a correct application of the standard.

Table 13. Requirements groups of the EN 301 549. The first column is the chapter in the standard. Clause in green apply to this system.

No.	Clause title	Applies
5	Generic requirements	Yes
6	ICT with 2-way voice communication	No
7	ICT with video capabilities	No
8	Hardware	Yes
9	Web	No
10	Non-web documents	No
11	Software	Yes
12	Documentation and support services	No
13	ICT providing relay or emergency service access	No

Table 14. Conformity of the system with clause 5 – Generic requirements. Requirements in red were not satisfied.

Clause	Name	Accomplished
5.2	Activation of accessibility features	Yes
5.3	Biometrics	Yes
5.5.1	Means of operation	Yes
5.5.2	Operable parts discernibility	Yes
5.6.1	Tactile or auditory status	Yes
5.6.2	Visual status	Yes
5.7	Key repeat	No
5.8	Double-strike key acceptance	No
5.9	Simultaneous user actions	Yes

Table 15. Conformity of the system with clause 8 – Hardware requirements. Requirements in red were not satisfied.

Clause	Name	Accomplished
8.2.1.1	Speech volume range	Yes

Evaluation of conformity with applicable requirements

Once the applicable requirements and recommendations were identified, we have inspected our system to evaluate which of the requirements are satisfied. The results are shown in tables 14 (generic requirements – clause 5), 15 (hardware – clause 8) and 16 (software - clause 11).

The evaluation results (Tables 14, 15 and 16) show that, on the one hand, most of the applicable requirements in clauses 5 (Generic requirements) and 8 (Hardware requirements) have been satisfied by our system. The only non-satisfied requirements of those parts are the related to the key repetition, because that functionality was not considered in the system.

On the other hand, 15 of the 51 applicable clauses of chapter 11 (software) are not satisfied. This is mainly due to the fact that the developed prototype implementation did not use the platform accessibility services provided by Android. Because of that, the application fails to satisfy several requirements under clause 11.3. In addition, some of the requirements in section 11.2, which are related to WCAG 2.0 are also not satisfied: provide enough colour contrast, enable text resizing, avoid images of text and provide programmatic access the language of the user interface.

7.1.4.3 Performance testing

The performance analysis is made to obtain the error rates for both biometric modalities. Then, further relationships between accessibility concerns and performance errors can be stablished. For this specific evaluation set-up we expect worse performance results than the current state of the art: usually, elderly people are not familiarise with biometrics and their biometric traits are probably damaged by ageing [98].

We have divided the performance results by sessions in order to analyse the user's evolution and by operations (from now on DC – Data Collection and P – Purchase) to analyse the differences under normal conditions and when making payments (where factors like stress influences the user's behaviour [99]). For example DC2 is the performance result of the data collection in the second session. We have used FRR and FAR to obtain the performance results. Having both FRR and FAR, we represent the results with the ERR.

The fingerprint FRR for DC1 and DC2 is obtained comparing each of the 5 genuine samples with the corresponding template, and doing this with each of the users enrolled in the system.

The FAR is obtained by comparing the template from each user with the whole set of 5 fingerprints of the other users, which are considered here as impostors.

Table 16. Conformity of the system with clause 11 – Software requirements. Those in red were not satisfied

Clause	Name	Accomplished
11.2.1.1	Non-text content (screen reading supported)	Yes
11.2.1.2	Audio-only and video-only (pre-recorded	Yes
11.2.1.3	Captions (pre-recorded)	Yes
11.2.1.4	Audio description or media alternative (pre-recorded)	Yes
11.2.1.5	Captions (live)	Yes
11.2.1.7	Info and relationships	No
11.2.1.8	Meaningful sequence	Yes
11.2.1.9	Sensory characteristics	Yes
11.2.1.10	Use of colour	Yes
11.2.1.12	Contrast (minimum)	No
11.2.1.13	Resize text	No
11.2.1.14	Images of text	No
11.2.1.15	Keyboard	Yes
11.2.1.16	No keyboard trap	Yes
11.2.1.22	Focus order	Yes
11.2.1.25	Headings and labels	Yes
11.2.1.26	Focus visible	Yes
11.2.1.27	Language of software	No
11.2.1.29	On focus	Yes
11.2.1.30	On input	Yes
11.2.1.33	Error identification	Yes
11.2.1.34	Labels or instructions	Yes
11.2.1.35	Error suggestion	Yes
11.2.1.36	Error prevention (legal, financial, data)	Yes
11.2.1.38	Name, role, value	Yes
11.3.2.3	Use of accessibility services	No
11.3.2.4	Assistive technology	No
11.3.2.5	Object information	No
11.3.2.7	Values	No
11.3.2.8	Label relationships	No
11.3.2.9	Parent-child relationships	No
11.3.2.10	Text	No
11.3.2.11	List of available actions	No
11.3.2.12	Execution of available actions	No
11.3.2.13	Tracking of focus and selection attributes	No
11.3.2.14	Modification of focus and selection attributes	No
11.3.2.15	Change notification	No
11.3.2.16	Modifications of states and properties	No
11.3.2.17	Modifications of values and text	No
11.4.2	No disruption of accessibility features	Yes
11.5	User preferences	No

The handwritten signature FRR for DC1 and DC2 is obtained comparing each of the 10 genuine samples with the corresponding template, and doing this with each of the users enrolled in the system. The FAR is obtained by comparing the template from each user with the whole set of 10 signatures of the other users, which are considered here as impostors.

Users decide whether to use fingerprint or handwritten signature for the purchases. The performance result in that case is given by the percentage of unsuccessful purchases (i.e. under the different thresholds of fingerprint or handwritten signature).

7.1.4.4 Performance Results

There were 108 purchases in total, 66 with fingerprints and 42 with signatures. Only 13 purchases were unsuccessful (all of them with fingerprints). All the results are shown in Table 17, including DCs and Ps for both modalities. The differences in performance are noticeable between both modalities, being the handwritten signature more reliable in all cases.

Table 17. Performance results by modality. DC1 and DC2 are error rates with respect to biometric recognition in terms of EER. P is the percentage of unsuccessful purchases.

Modality	DC1_EER (%)	DC2_EER (%)	P_unsuccessful (%)
Fingerprint	30,41	35,84	19,70
Handwritten Signature	1,85	7,43	0

7.1.4.5 Discussion

Accessibility conclusions

The design of the app was made taking into account general guidance on accessibility, but not the European standard EN 301 549. Even so, the results of the accessibility evaluation show a good coverage of the requirements of the EN.

The biggest accessibility gap of the application lies in not using the accessibility services of the platform. That makes the app inaccessible for persons that require assistive technologies such as screen readers (for blind people), screen magnifiers (for low-vision users) or alternative inputs (for persons with motor or dexterity limitations).

In the future we plan to improve on the current situation, by making the app using the accessibility services provided by the Android platform. In most cases that just implies properly using the standard user interface components and providing text-based information associated with non-textual elements. At the end of the process we expect to have a high degree of conformance with the EN 301549.

Performance conclusions

The results obtained are not in accordance with the current state of the art, where fingerprint recognition is highly reliable [88] and handwritten signature recognition does not return as good results as in this experiment [100]. The most plausible reasons of these results are:

-Elderly people are used to sign documents and not to place the fingerprint on biometric devices, so that, the lack of practice can involve poor performance.

-It is noticeable that fingerprints erode with time, especially in people who work with their hands. Then, for elderly, fingerprint recognition is not probably the best biometric modality and bad performance should be expected.

-The initial inconveniences when signing on a mobile device with a stylus were overcome in the training phase. Thus, when users started the real evaluation they were already used to the device.

There is not an improvement between DC1 and DC2, pointing that users did not learn how to deliver the biometric traits properly (actually, the performance decreased in all cases). This suggests that elderly people need more training sessions to be used to biometrics. Anyway, there is not too much room for improvements in fingerprint recognition as the removed fingerprints cannot be rebuild.

Regarding the purchases, it was an initial distrust from several users who felt unsure about making payments through a smartphone. This involved that many users rejected to participate in the evaluation. There is a necessity of design interfaces to make users feel safer when making banking procedures.

Further conclusions

Our findings suggest that the accessibility concerns found in the app do not involve a performance decrease, which is due to the poor quality of the fingerprints. Nevertheless, an improvement into the app interface would lead to make the app usable for a wider percentage of people with accessibility problems. This work shows that accessibility can be reached by following the standards. Otherwise common mistakes may drive to designs that are not properly usable.

7.1.5 EXPERIMENT 7: Accessibility Evaluation through traditional usability metrics

Once all the non-satisfied requirements have been overcame, a new accessibility evaluation was suggested. In this case, the scenario is the same as in the previous experiments, as well as the app, devices and procedures. Nevertheless, the user's group is different (physical and mental accessibility concerns) and the accessibility is not evaluated through the EN 301 549 but by usability-related parameters described in Part III. Performance results are obtained in the same way as in the previous experiment. Next, we explain the factors which change with respect the previous experiment (users), the measurements, results and conclusions.

Motivation of the experiment

When measuring accessibility it is important to remark the wide range of existing disabilities. In the previous experiment, we invited elderly people (a representative group in the society), but it is necessary to increase the range of users to know with certainty if the design is widely accessible. Then, we invited users with other different accessibility concerns to participate in the experiment. Accessibility results obtained from this experiment complement the results obtained from the previous experiment. Moreover, the accessibility issues of this experiment are expected to be different from the previous one.

During this experiment, the metrics used are the percentage of users who complete the experiment, the time spent in the process and the user satisfaction. Time and satisfaction are not accessibility metrics but considered as relevant in this specific case: even when the accessibility is correctly addressed, if the time is long or users are not satisfied, the app does not make sense.

Influence on the H-B interaction

This experiment validates the well-known usability metrics added to the H-B interaction: users who can use the system and users who know how to use the system. Furthermore, we tested users with different accessibility concerns, covering many of the accessibility factors depending on the user.

7.1.5.1 Modified factors

Users

Testing accessibility in ICT products requires the participation of people with accessibility problems. Therefore, the invited users are patients of the CRMF. A total of 70 people went to a previous explanation meeting and 35 of them initially agreed to participate. Before starting the evaluation 10 users left due to distrust so 25 users started the 1st session.

During the experiment 4 users could not complete the process because they were not able to follow the experiment instructions or could not use the devices properly. Finally 21 users completed the evaluation. 10 of them are between 18 and 30 years old, 9 are between 31 and 50 and only 2 are between 51 and 70 years old. There were 10 users with different illnesses in their hands:

- *3 of them could not use the right arm*
- *2 users have trembling in their hands*
- *3 users have not enough strength in their arms*
- *1 user has a brachial plexus injury*
- *1 user has a stroke in his left arm*

The rest have other illnesses and/or combine more than one:

- *8 of them use a wheelchair*
- *1 uses a crutch to walk*
- *2 of them suffered an ictus and have trembling*

None of the users have vision problems, auditive problems or health issues derived from elderly. All of them had previous experiences with mobile devices and computers but not using biometrics.

7.1.5.2 Accessibility Evaluation Methodology

Because this app is EN 301 549 compliant, it is accessible a priori. Nevertheless, further measurements, through user's interaction, are required to corroborate that hypothesis. Therefore, usability related parameters have been applied to evaluate the system accessibility:

- *What is the % of users who completed the process?*
- *What is the % of users who completed the process with external help?*
- *Is the time spent in the process acceptable?*
- *Are the users satisfied with the system?*

7.1.5.3 Results

<u>Accessibility results</u>

Regarding the accessibility results, in terms of task completion, 12 users finished the experiment without problems, 9 users needed partial help and 1 user could not finish. Then 21 users **can use the system** and 12 user **know how to use the system**.

The reasons for asking help were mainly concerns with the fingerprint sensor (bad position, sweat or lack of fingerprint). The user who could not finish the process has very low mobility and requires assistance for most of the daily tasks. In Table 18 is the time spent in average for all the evaluation steps.

Table 18. Average time spent (in seconds) in the different experiment stages. The ± specifies the variances

Session	Training	Enrolment	DC	P
1	200,29 ± 32	131,19 ± 18	307,62 ± 48	79,76 ± 9
2	-	-	320,48 ± 47	71,43 ± 9

Table 19. Satisfaction results extracted from the surveys. The ± specifies the variances

Parameter	Question	Mark
Comfort	*Fingerprint sensor*	4,28 ± 1,33
	Stylus	4,05 ± 1,85
Time	*Fingerprint sensor*	2,05 ± 0,95
	Stylus	2,05 ± 087
Security	*Lock device*	Yes (17) No (4-Comfort)
	Money ATM	Yes (17) No (4-Security)

The satisfaction results gathered from the surveys are in Table 19. Those users who would not use biometrics for locking the device alleged comfort reasons and those who would not use biometrics for collect money from an ATM were not confident of biometrics security.

<u>Performance results</u>

The performance results obtained from the samples acquired in DC1 (session 1) and DC2 (session 2), given in terms of EER, are in Table 20. The low performance of fingerprint recognition (25~30%) is remarkable in comparison with the good performance of the handwritten signature (1~5%). Best results were not expected to be obtained in Session 1. The performance obtained through the purchases, in terms of percentages of successful attempts are in Table 21, divided by modality.

Table 20. Performance results by modality and session (S1 – Session 1, S2 – Session 2)

Modality	EER_S1	EER_S2
Fingerprint	27,53%	29,84%
Signature	1,36%	3,29%

Table 21. Purchases results by modalities

Successful 98,63%	Fingerprint	54,79%
	Signature	43,84%
Unsuccessful 1,37%	Fingerprint	1,37%
	Signature	0,00%

7.1.5.4 Conclusions obtained for future improvements

Accessibility

Once the experiment has been finished and all the users interviewed, we can conclude (based on user's opinions from informal chatting off record) that the accessibility level reached has been more than acceptable. This is even more encouraging when this app has not been specifically designed for this particular group of people (i.e. users with wheelchairs, arms illnesses, etc.). Moreover, going through the usability related measurements, only one user could not finish the experiment. Regarding the users who asked for help, more training could drive them to better locate the fingerprint on the device. Also other designs could solve the ergonomic concerns found.

The main drawback found has been the time spent in the process. This has not decreased in the second session with respect the first, which involves that users have not really got enough skills during the first session. Therefore, more training in required.

According to Table 19, the time spent is the worst scored feature in the satisfaction surveys. This could be motivated by the several repetitions of the process during the evaluation, where users were required to present their biometric traits several times. We believe that a common use of the app in a real scenario would not be annoying for users because the interaction time with the app is much less.

Regarding the comfort, users preferred generally the fingerprint sensor (scored with 4,28) because it is faster and requires less interaction than handwritten signature (scored with 4,05). Nevertheless, users expressed that they felt comfortable using both devices.

Users who would not use biometrics for locking their smartphone (medium risk action) claimed comfort reasons and in the case of the ATM (high risk action) claimed a lack of security. Though, most of the users (17) viewed favourably the use of biometrics for locking the smartphone or withdrawing money from an ATM. This encourage us to believe that most of them would use biometrics in their daily life.

Though this usability evaluation methodology could be appropriate to measure accessibility, proper evaluation measurements would be more convenient (apart from the EN 301 549, which is mainly focused in the app interfaces and is not specific for biometric recognition systems).

Performance

The results related to performance are not in the line of the current state of the art in biometric recognition, where fingerprint returns much better scores than dynamic handwritten signature recognition. In our case, the low fingerprint performance could be motivated by:

Ergonomics. This was the first time that users interacted with a fingerprint sensor. Furthermore, the sensor used is not embedded in the smartphone, it has a wire and it was not specifically designed for people with accessibility issues.

Fingerprint quality. Several users felt not really comfortable participating in the experiment, which motivated hands sweating in some cases due to stress or lack of confidence. As sweating decreases the quality of the fingerprint, the performance could decrease as well.

The results obtained in signature recognition were better than expected. Reasons for this performance could be that signing is a common procedure and the use of digitizers is becoming more usual nowadays (at the bank, postal services, etc.). Signing on a digitizer is similar than signing on a smartphone like the one used in the experiment (both actions are made with stylus over a capacitive screen and good signature feedback). In both modalities, the performance decreased in the second session pointing that users have not acquired enough skills and more training is needed.

Other conclusions

After this series of evaluations, we can conclude that develop universal accessible apps is nearly impossible due to the wide range of different accessibility issues that exist. In our experiment, even when several participants used the fingerprint sensor without concerns some others could not. This fact, joint to the poor performance results in fingerprint recognition, encourages us to suggest other biometric modalities such as face recognition (less invasive and requires less interaction). Nevertheless, other collectives could prefer fingerprint recognition (e.g. blind users). Therefore, our findings suggest that a convenient accessible design must rely on the subject characteristics.

Chapter 8

Evaluation of the interaction of disabled people with biometrics. A case study: blind subjects in facial recognition

The Blind Subjects Faces Database (BSF from now on) is a project developed between the GUTI and the University of Surrey. The project started with the intention of improving the accessibility in biometrics. Then, three experiments were performed from a collected database of blind and semi-blind users taking self-pictures (selfies) with a mock-up mobile device having different kinds of feedback. Next, we introduce the database, the characteristics of the evaluations carried out and the results and conclusions obtained.

Motivation of the experiment

This experiment is slightly far from the experiments line we followed regarding the validation of the H-B interaction. Once we have finished the planned validation, we continue with the work on accessibility in biometrics, but focusing on other aspects. As this work was carried out in collaboration with the University of Surrey, the methodology followed is the one which they follow. In the BSF, we focused on a specific case of accessibility: face recognition in blind users. The motivation of this work is to show up the main challenges when gathering a database of people with accessibility concerns and processing the results. Further conclusions regarding usability and accessibility are reached and explained. The experiments made are:

A. Applying alignment and alignment-free approaches to the BSF. This experiment analyses which kind of algorithm works better for this challenging database where subjects with visual impairments take selfies.
B. The FCD (Face Detection Confidence) in time. Here we analyse how is the evolution of the user's skills to take selfies in time according to the FDC. The different feedback modes have a noticeable repercussion on this factor.
C. Performance evolution in time. In this experiment several hypotheses regarding the performance obtained under the different feedback modes are contrasted.

Final conclusions of the three experiments were gathered and set together at the end of the experiments.

8.0.1 General Characteristics

During the data collection each subject was told to take self-videos of his/her face under different feedback modes. Each image is time-stamped so that subsequent studies can assess the efficiency of an interaction session. The camera used was a mock-up device which is very light and easy to handle and is connected to a PC. Four experiments per session were planned providing four different modes of feedback; thus allowing post-experimental analyses of usability, accessibility, and performance. The dataset collected includes videos of each user-session-experiment, personal data (gender, age, degree of blindness and opinions about the experiment) and more than 70 thousand face images. Each image was stored with the corresponding bounding boxes around a detected face (there may be several boxes per detected face), a time-stamp of the image, and the face detection confidence (FDC) which is the confidence of a face classifier that the region of interest is a face. The higher the FDC the higher the likelihood that the region of interest contains a face. Although there is a face detector used during the acquisition and interaction process in real time, a post-experimental evaluation may also include a more accurate face detection algorithm.

Our primary reason is to study if state-of-the-art face recognition is ready to be used by blind or partially blind subjects or not; and if they are not adequate, how alternative forms of feedback (other than visual), such as audio and tactile can be used to assist them so that the technology is more usable. Ultimately, our goal is to render the face recognition technology available for everyone to use. Other potential uses of the BSF database include:

- Usability and accessibility studies under four controlled feedback modes.
- Usability in terms of efficiency. The inclusion of timestamps allows further studies such as the study of performance or the user's behaviour that evolves dynamically with the time.
- Ergonomics. Is the camera used the most usable and comfortable? Has it any positive or negative effects over the results or the user's satisfaction?
- Benchmarking face detection algorithms. This database contains several types of realistic face images captured under very challenging conditions. Many images are not well-aligned, contain partial faces, or faces that are blurred due to swift

movements, or are out of focus. All these conditions make face detection and recognition very challenging.

- Benchmarking face recognition algorithms. Since many face images are difficult to be aligned, conventional face recognition systems that need initial alignment, such as eye landmarks, is likely to fail. One possible use of this database is, therefore, to compare the recognition performance between alignment and alignment-free algorithms.

8.0.1.1 Evaluation set-up

Users

There are 40 subjects in the database, consisting of 29 men and 11 women, covering a wide range of age groups, visual impairment levels, and the ability of taking selfies. None of the participants has ever used any biometrics device previously. Their age distribution is as follow: 45% are under 25 years old, 42% are between 25 and 50, and 13% are over 50. Regarding the level of visual impairment, 16 of them have low vision, 14 can distinguish light and darkness, whereas 10 are completely blind, out of which 5 are born blind. Almost half of the participants (15) claimed the ability of taking selfies and 16 claimed the ability of taking pictures of other people. The count of subjects' age, gender, and level of visual impairment is shown in Figure 61.

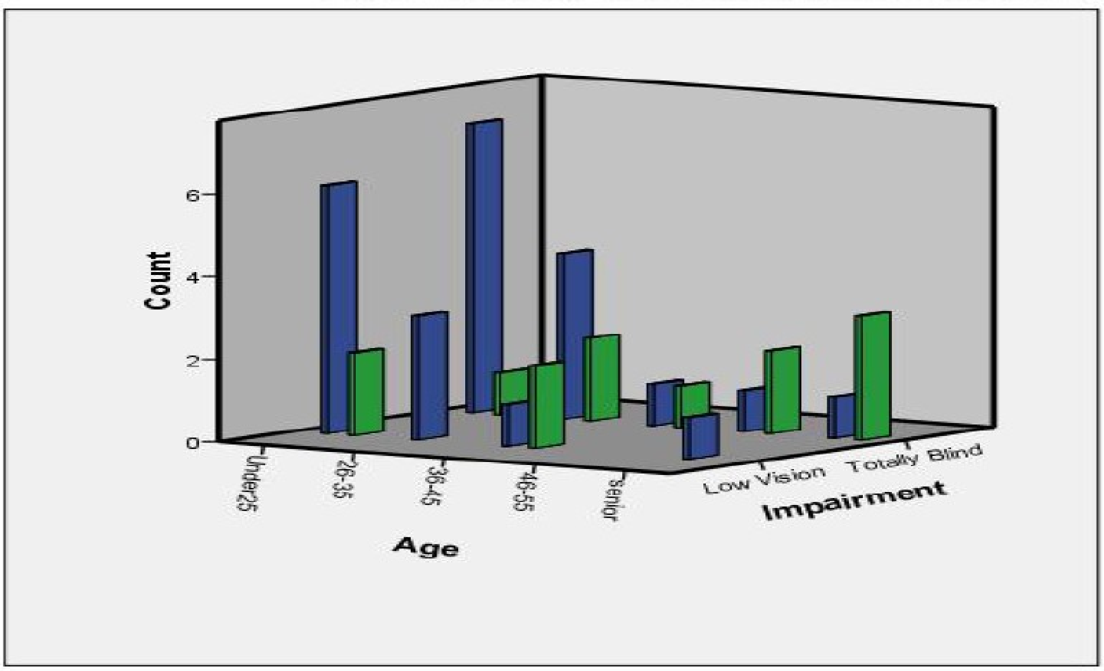

Figure 61. BSF participants by age, impairment and gender (blue-male, green-female)

Hardware and software

The camera used to capture the face images is an Advent Slim 300K web-cam, which is light and small. The device can be handled easily in a participant's hand, as shown in Figure 62. It is connected to a laptop where a desktop application controls the process flow. The application captures a video stream from the camera and stores images with an image resolution of 640-by-480 pixels at 30 frames per second. The user has a timeout of 45s in order to obtain a frontal face image. Then, the application processes each image and returns the audio feedback as described in E2.

8.0.1.1 Usability Experiments

The usability experiments have been carefully designed to consider both uncontrolled and controlled factors. Factors that affect the subjects such as moods and feelings are uncontrollable. Similarly, their degree of familiarity in taking selfies or pictures of others cannot be controlled but this information is recorded for subsequent analysis. Since our objective is to understand the how different feedback modes can improve the accessibility of the face recognition technology as well as its performance, we shall control the following three factors: habituation, instructions received, and audio feedback. This gives raise to four acquisition scenarios which are further enumerated below.

Despite the four scenarios, the task is the same in each setting, i.e., take a selfie emulating the scenario of unlocking a mobile device using face recognition. At the end of the experiment, we shall then measure the results of the acquisition process across the scenarios in terms of accessibility and recognition performance conditioned on the four feedback modes, as enumerated below:

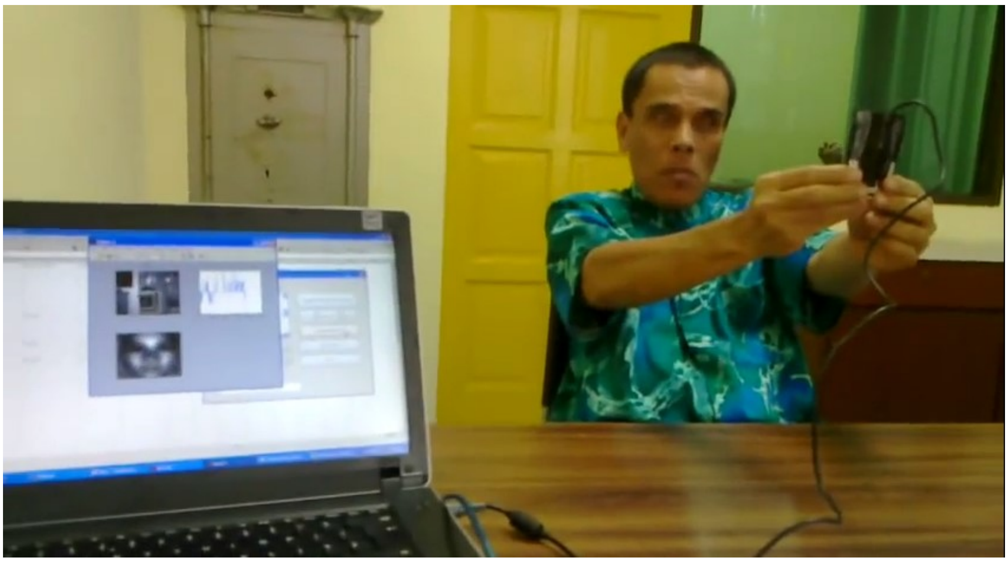

Figure 62. Example of the BSF acquisition, including the camera and the capture software

Experiment 1 (E1). The user receives no feedback or instructions when taking a selfie. This experiment is designed as the baseline. If a subject can take a selfie, he or she will perform better than another subject who does not have this ability. Without any instruction or training, a subject that is born blind (hence, classified as totally blind) naturally would perform worse than another subject who is partially blind. This experiment is expected to be the worst in terms of the number of detected faces as well as the facial recognition performance compared to the subsequent experiments to be described below because the user has not yet acquired the skill in taking selfies. Moreover, this experiment is the first one to be completed in the evaluation so that it does not bias the other experiments; since, by definition, the user has not accustomed to using the device.

Experiment 2 (E2). The user receives audio feedback just before taking his/her selfie. The audio feedback is set at 3 different frequency levels which depend on the FDC of the acquired image. The FDC is given by a Viola-Jones based face detector [71], which is able to detect more than one face per image. The provided frequency is low (1.5 KHz) if the face detector does not detect any face, medium (4.5 KHz) if it detects a non-frontal face and high (7.5 KHz) if a frontal face is detected. The definition of non-frontal versus frontal face image is distinguished through a systematic experiment carried out offline. The audio feedback is intended to help the user to better point the camera to capture a face as frontal as possible. This experimental setting is to be completed right after E1 so that the user is expected to acquire the skill of holding the camera by appropriately adjusting its (position and distance from his/her face during this experiment. A detailed description of the above audio feedback mechanism can be found in [72] and [71].

Experiment 3 (E3). In this experimental setting, audio feedback is not provided but instead, the user receives information about how to take the selfie before starting the experiment. This information consists of a supervisor who helps the user to adjust the distance between the camera and the face so that a proper selfie face image is taken. Although the intention is to isolate E2 and E3 (completed right after E2), this is not completely possible because the user has already acquired the skill during E2. Therefore, he or she would have known approximately how to grab the camera to obtain the best self-image.

Experiment 4 (E4). This experiment is a combination of E2 and E3. In this experiment, the user receives previous instructions on how to grab the camera and also audio feedback during the capture process. Therefore, the face detection and recognition performance results are expected to be the best in this experiment because, apart from the audio feedback and the instructions, the user would have also acquired the skill and habituation needed in using the camera from E1, E2 and E3 settings.

Figure 63. Images of the same user in the E1, E2, E3 and E4 from left to right. Images are expected to improve in the last experiments (i.e. in distance and focus).

Examples of images taken in each experiment are shown in Figure 63.

An ideal way to carry out the experiments is by using RCT (Randomized Controlled Trial), whereby a user is subject to one of the four modes of feedback. However, we were concerned with two issues with this approach:

The first is the fact that biometric performance is subject-dependent. Therefore, if the same experiment (with the same feedback mode) is conducted but on two different disjoint populations, two different results will be obtained. This is because the effect of the subject variability on the performance of face detection and recognition might be higher than the effect due to the mode of feedback. The impact of inter-subject variability on biometric performance is well known in the literature, and this phenomenon is referred to as Doddington's zoo or biometric menagerie [101].

The second concern is that the number of blind subjects is limited, which is about 40. For this reason, deploying RCT would mean that one has to divide the population into a smaller set with 10 subjects for each mode of feedback. This is arguably not the best use of limited samples available to us.

We have therefore, opted for subjecting every volunteer to all the four modes of feedback, but doing so carefully so that the effect of one feedback does not influence that of another. One way to achieve this is by exploiting the natural ordering of the modes of feedback. For example, the E1 setting does not have any feedback and so should be carried out first. E2 and E3 are each independent of each other because the audio feedback does not convey any information about the instruction. However, the instruction (E3) mode should take place just after E2. A potential weakness of the above approach is that the volunteer may have become more familiar with the device after each experiment which is conducted sequentially. Fortunately, after a post experimental analysis, we found that this is not a particular concern. Finally, E4 should be conducted last because the user has to have the knowledge of interpreting the audio feedback and should have been given the instruction of taking a high quality selfie image.

Table 22. Some statistics of BSFDB divided in sessions and experiments. All the snapshots include bounding boxes and FDC. Valid images are those which contain a face according to the face detector

Session	*Experiment*	*Snapshots*	*#Valid images*
1	1	17200	4303
	2	17600	7386
	3	17600	9069
	4	17600	10816
2	1	16400	8478
	2	16400	9150
	3	16400	9911
	4	16400	11065
Total		135600	70178

8.0.1.2 Database contents

Each user underwent two sessions of experiment, each of which is separated by about two weeks apart. In each session, the user was then subject to the four experimental settings as described above. As a result, for each user, we have 8 unique experiment-session combinations; and hence 8 videos. Each video contains between 0 and 400 images. . The number of images varies from one video to another due to the time variability of the face detection. When an image is taken, it is saved with the corresponding bounding boxes around a detected face, a time-stamp of the image, and the FDC associated with each bounding box. More bounding boxes are detected this way since we have set the face detection threshold to a lower-than-usual value so that both frontal and slightly off-frontal face images can be detected. In addition to the raw images, the BSF database also contains pre-processed and extracted images. Table 22 summarises the content of BSF.

8.0.1.3 Challenges in the acquisition and processing

There are two challenges that we have to deal with during the acquisition stage as well as during the images processing stage. These challenges are mostly due to the difficulty of the blind or partially blind subjects in taking selfies. Several participants found it difficult to focus their face especially in the E1 setting where they had not audio feedback, for instance. The main problems are further elaborated below.

<u>During image acquisition</u>

The user moves his/her head whilst adjusting the mobile device constantly in order to take a selfie. There are effectively six degrees of freedom – three due to the head movement and another three due to the positioning of the mobile device. As a result, many images contain either a partially observable face or else no face at all. If a user fails to acquire an image at the beginning of the sequence, he is likely to spend more time in taking his/her selfie. When a user exceeds the fixed timeout of 45s, the entire sequence can contain no image at all. Figure 63 illustrates six examples of error during acquisition:

Figure 64. Examples of different problems found when acquiring face images

a) The camera is placed too closed to the camera and is off cantered. As a result, only part of the face can be observed; and even so, the image appears to be too dark.

b) The camera is placed too far from the face. Furthermore, it is placed slightly below the required height as it is looking from above.

c) The user places her hand between the camera; thus occluding her face. Furthermore, there is too much light and the image is blurred due to the rapid movement.

d) The face image is not frontal.

e) The user is too close to the camera and there is too much light.

f) The user wears glasses and the image is blurred.

<u>During image processing</u>

Due to errors occurred earlier on during the images acquisition phase, these problems need to be addressed in the subsequent processing phase. As a result, the system has to handle images without a face and images with partial faces, rotated faces, or those that are blurred. In addition, because of the use of low threshold on the Face Detector Confidence, which is necessary in order to improve the chance of acquiring images with faces, false positives are more likely to occur. For example, as shown in Figure 65, the face detector confusingly recognizes the nostrils as the two eyes.

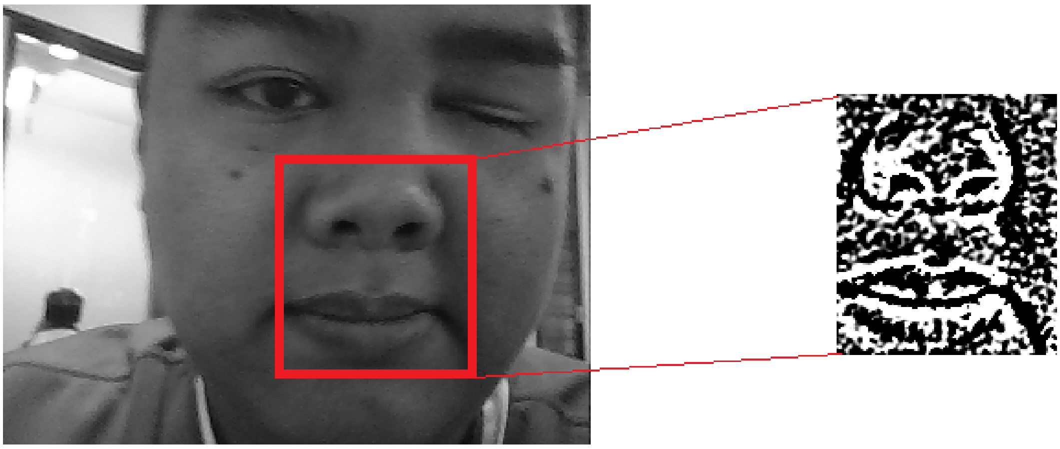

Figure 65. Example of false alarm due to the face detector. This is classified as an error in image processing. In this example, since the user is holding the camera too close to his face, the face detector has wrongly classified the nose and mouth as a face. The face detector is also operating at lower-than-usual face detection threshold in order to be able to tolerate for somewhat less frontal face images. As a result, false alarm is inevitable.

8.0.1.4 Analysis of the database

In this section we present a preliminary analysis of the BSF database. This consists of two metrics, namely, face recognition performance and FDC. We have also analysed the time influence in performance.

For the first metric, we measure the system performance obtained when applying two different face recognition algorithms, namely alignment and alignment-free algorithms. The advantage of using an alignment-free algorithm for face matching is that it does not require precise localisation of the face. Unlike alignment-based approach, an alignment-free algorithm can match two images containing only a partially observable face in each image.

In the second analysis, we analyse the evolution of FDC over time. This is important because we would like to measure and compare the efficiency of the four different modes of feedback (i.e., E1—E4 experiments). Even if two modes can acquire image of sufficiently good quality, the mode that can acquire a face image faster is considered a better one. For both analyses, a common image pre-processing procedure is used; and it is described next. The time (efficiency) influence in performance was measured trough the valid images and performance variations.

8.0.1.1 Pre-processing

The objective of the pre-processing module is to reduce the adverse impact of noise, e.g., to standardise varying lighting conditions; and to ensure that each detected face image has the same size for subsequent processing. During the acquisition process, each image is subject to face detection and a bounding box is associated with each detected face. Each image is cropped using the detected bounding boxes. Afterwards, photometric normalization is applied to the cropped face image [102] in order to correct for the lighting variation. Finally all the images are resized automatically to a common size of 120-by-142 pixels. Figure 66 illustrates an example of this pre-processing.

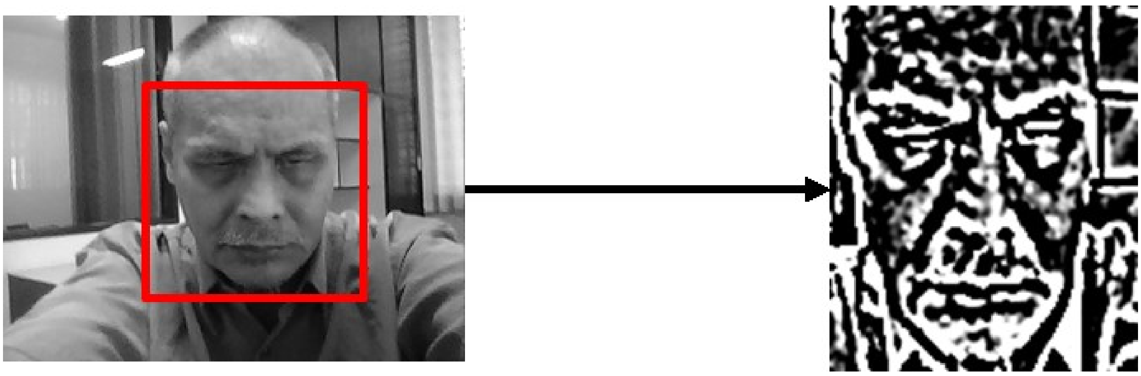

Figure 66. Image before and after the Pre-processing. The red square is the bounding box

8.0.2 EXPERIMENT 8: Alignment versus alignment-free

This experiment compares two approaches to face recognition, namely alignment and alignment-free face recognition. Since many of the BSF images are very noisy – due to, for instance, occlusion, blurriness, partial faces, etc. – they often do not contain key facial landmarks that are commonly required in traditional "alignment-based" algorithms. For this reason, we expect an alignment free algorithm to produce superior performance. We have selected 300 best images for each user from all the experiments in session 1. By best images, we understand that the top 300 images that has the highest FDC value in each session are retained for further processing. The same approach is applied to select another top 300 images for each subject in session 2. Effectively, we have designated session 1 as the enrolment data whereas session 2 as the probe data. The output of this process is a set of genuine scores from which FRR can be estimated.

In order to generate an impostor score set for a given target user, the images of the remaining users are used. Therefore, given 40 subjects, we have 40 sets of genuine scores, and 40x39 sets of impostor scores. The union of the 40 sets of genuine scores are used to calculate FRR whereas the union of the 40x39 sets of impostor scores are used to calculate FAR. In addition, we also summarise the performance in terms of EER. EER is a very useful performance metric to summarise a ROC curve because it enforces the prior class probabilities of false acceptance and false rejection to be equal. The algorithms applied are the following:

- Principal Component Analysis (PCA): The PCA is a method commonly used to reduce the dimensionality of an input vector which consists of a large number of interrelated variables while retaining as much as possible of the variation present over the entire data set [103]. For this experiment, we used the Pretty helpful Development (PhD) face recognition tool [104] [105] which implements the Eigenfaces approach to obtain the PCA vectors. In order to derive the PCA matrix, we use the ORL (AT&T) [106] as a separate training set. Then, we project all the detected face images from the BSF database into their corresponding PCA subspace. Finally, the matching is performed by the dot product of two image samples represented in their corresponding PCA subspaces.
- Linear Discriminant Analysis (LDA): The LDA is used to find a linear combination of features in such a way that two or more classes of objects are maximally separated [107], i.e., with large inter-class distances whilst having minimal intra-class variability. Again, the PhD tool which implements the Fisherface approach is used. Prior to projecting the data into the LDA space, a PCA step is performed in order to avoid the singularity issue due to a small number of training samples. Again the ORL (AT&T) database is used to derive the PCA and LDA project matrices. Finally, the matching is also performed by computing a dot product between two samples in their corresponding LDA subspaces.
- Scale-Invariant Feature Transform (SIFT) [108]: Unlike the PCA and LDA which are considered holistic face recognition that requires a proper alignment of face images,

SIFT is considered an alignment free algorithm. SIFT is a representation that describes local features in images and is commonly used for object recognition. The alignment-free property in SIFT is due to its use of salient features which are both scale and rotation invariant. When applied to a face image, these salient feature scan then represent the face of a person without requiring any landmark provided by a localisation algorithm. In this work, we have used SIFT because it is resistant to occlusion, scale and orientation changes. A SIFT descriptor for a salient point is effectively a histogram of local gradients that have been trained from a large corpus of facial parts. A face image is then represented by many SIFT descriptors. To compare two images represented in their respective sets of SIFT descriptors, a Euclidean distance between these descriptors is calculated. If their distance is below a prefixed threshold, the local patches of both images are considered a match. The final matching score is computed as the number of paired descriptors divided by the number of available descriptors.

8.0.2.1 Results

The performance results showed in Figure 67 in terms of ROC curves reveal that the BSF database is challenging for face recognition. This is not surprising due to the variation in the images in terms of lightning, blurriness, and low percentage of detected faces. In terms of face recognition performance, the result due to SIFT is significantly better, with an EER: 12.49% than in the traditional approaches based on PCA or LD, which has an EER of 38.04% and 35.10%, respectively. This shows that an alignment-free face recognition algorithm is likely to be better than any alignment-dependent face recognition for this database due to the difficulty of finding reliable and complete landmarks of a face.

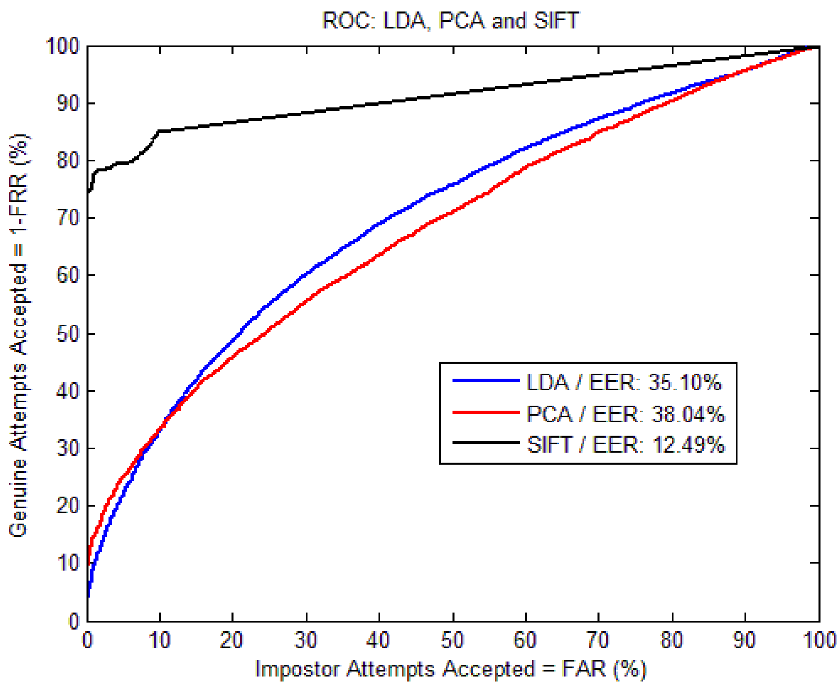

Figure 67. LDA, PCA and SIFT ROC curves including the EER

8.0.3 EXPERIMENT 9: The evolution in time of the FDC

In the second experiment we measured the evolution of the FDC in time through the 4 experiments. It is expected that this value increases along with the time. As more time has passed, the subject is also more likely to improve his or her FDC score. In addition, we also expect better FDC scores from condition E1 through to E4 in a progressive manner since the subjects have access to audio feedback (as in E2), or would have given instruction (as in E3), and have both feedback and received instruction in taking selfie (as in E4). For each experimental setting, the entire duration of image acquisition session is divided into three identical time slots, which corresponds to initial, medium, and last phases in order to better analyse the evolution of interaction over time. Consequently, the FDC is calculated in each time periods as described above.

8.0.3.1 Results

The results of this experiment is summarised in the form of a box plot of the FDC score across the four experimental conditions as well as between the initial and last time periods (labelled as time 1 and time 3, respectively) in Figure 68. The differences in FDC are not highly significant between Time 1 and Time 3 in the E1 to E3 but an improvement is somewhat observed in E4. This means that the quality of the images improves when users receive instructions and audio feedback. It is also worthy to note that the median FDC value is similar in E2 and E3 (close to 11) for both Time 1 (FDC_E2 = 10,5 / FDC_E3 = 12) and Time 3 (FDC_E2 = 10,5 / FDC_E3 = 11). This experiment reveals the differences in quality among images when users receive different feedback modes. These differences have also repercussion in the final performance and mean that when users receive both audio and visual feedback the results are better.

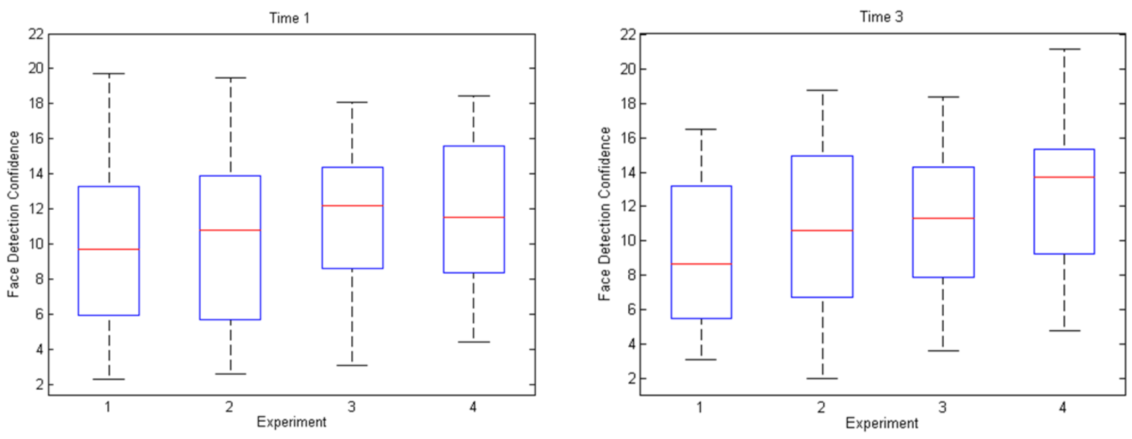

Figure 68. FDC boxplots by Experiments in Time 1 and Time 3. In the boxplots: the central box represents the central 50% of the data. Its lower and upper boundary lines are at the 25%/75% quantile of the data. A central line indicates the median of the data. Two vertical lines extending from the central box indicate the remaining data outside the central box that are not regarded as outliers. These lines extend maximally to times the height of the central box but not past the range of the data. Outliers (+): these are points indicating the remaining data.

8.0.4 EXPERIMENT 10: Performance evolution in time

This experiment studies how is the performance evolution in contrast with the time spent in the recognition. The main target is to contrast several hypotheses related to the efficiency in this evaluation. Next, the suggested hypotheses and the methodology followed to process the data are described. The algorithm applied is the SIFT, as long as it is the most reliable of the 3 used in the previous experiments.

8.0.4.1 Hypotheses

On the basis of the literature about efficiency [53] and face recognition [17] we had suggested several hypotheses that are all derived in the section 4. Those are the following:

i) Performance results are better in the last experiments (i.e. E4>E3>E2>E1) because a) the user has already acquired more habituation and b) the user has received more information about the process. This should involve better images' quality and therefore, better results.

ii) The audio feedback is more useful than the information because it is provided in real time and the users can correct bad postures on the spot.

iii) There are less performance variations in the last experiments (i.e. E4<E3<E2<E1) because the user is more habituated to the system and makes less mistakes.

iv) The number of valid images is higher in the last experiments because of the habituation because a) the user has already acquired more habituation and b) the user has received more information about the process.

8.0.4.2 Methodology

The SIFT algorithm returns the number of matches between two images. A match is obtained when the distance between the first and the second nearest neighbour between 2 descriptors is under a given threshold. Notice that comparing image A with image B can return a different number of matches than comparing image B with image A. Then, once the SIFT had returned the number of matches between two images, they were normalized to the number of descriptors of both to obtain a score:

$$Score = mean\left(\frac{matchesAB}{descriptorsA} + \frac{matchesBA}{descriptorsB}\right)$$

Applying this kind of normalization the score will be low as the number of matches between two images is always too much lower than the number of descriptors. In order to calculate genuine and impostor rates we have obtained the templates from the session 1, being the image with the highest confidence the template for each combination of user-experiment. The genuine rates were computed matching each user-experiment template with all the images from the same user-experiment of the session 2.

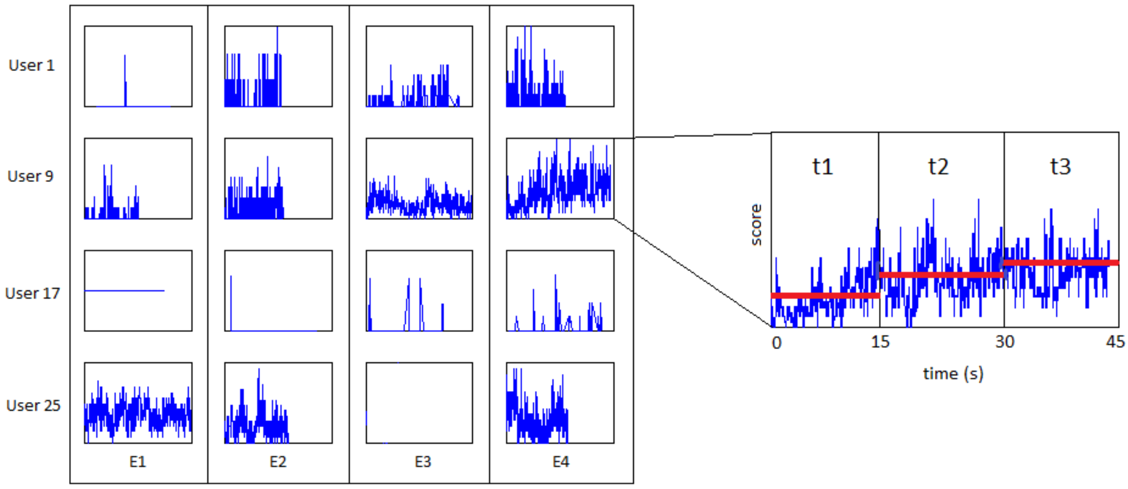

Figure 69. Four random users' performance results in the four experiments. The x axis is the time and the y axis is the genuine score. The red lines are the scores mean for each time slot (μ' ∈ R).

8.0.4.3 Scores in time

A first approach to the genuine scores has shown meaningful differences from some users to others representing no consistence as to obtain broad conclusions (Figure 69 left). Thus it is necessary to normalize the results on a user by user basis with the baseline in order to obtain reliable results. To obtain the scores evolution in time, we have divided each user interaction (each user-experiment) into three parts of the same length, namely t1 (first part), t2 (second part) and t3 (third part). Then, we have calculated the mean score per each user-part and normalized it (μ' ∈ R) to the same part in the baseline (E1) (Figure 69 right).

$$\mu_k^{t\,\prime} = \frac{\mu_k^t - \mu_1^t}{\mu_1^t} \qquad \textit{for } k \in \{2, 3, 4\}, \textit{ for } t \in \{1, 2, 3\}\,.$$

Then, we have measured the differences in the genuine scores between t1 and t3 (the beginning and the end of the interactions).

8.0.4.4 Results

In this section we contrast and discuss all the suggested hypotheses with the results obtained. First, we have used the t−test between the pairs t1−t3 for each experiment (except for Experiment 1 which is not normalized) to validate the data. The t−test showed these p−values: pExp2 = 0.12, pExp3 = 0.09, pExp4 = 0.25. Authors conclude that the database size should be increased in order to obtain more reliable and statistically significant results. Using the current database and following the initial order of hypotheses:

i) According to the results obtained, the experiments order (according to the performance evolution in time) is not as expected and the correct order is E4>E2>E3>E1 as shown in Figure 70b. The fact that the results in E2 outperform those of E3 points out that the audio feedback could be more useful than the previous instructions. As E2 has obtained better scores, we also suggest that the audio feedback is even more effective than the habituation.

ii) The audio feedback was more successful than the instructions given: as shown in Figure 70, E2 and E4 (those experiments where the audio feedback was provided) are better in performance.

iii) The Figure 70, where the variance is represented by the boxplots, shows that there is not an experiment dependency regarding the performance variance as it does not tend to change in any experiment. This fact shows that users did not get habituated to the system as much as to gain consistency in the results.

iv) The experiments order with regards to the amount of valid images in this case is as expected (according to the number of valid images): E4 – 6598 > E3 – 6216 > E2 – 5538 > E1 – 5238.

Figure 70 shows also that performance increases in time in all cases. Therefore, though the performance evolution is not consistent in all users, its tendency is to increase in time.

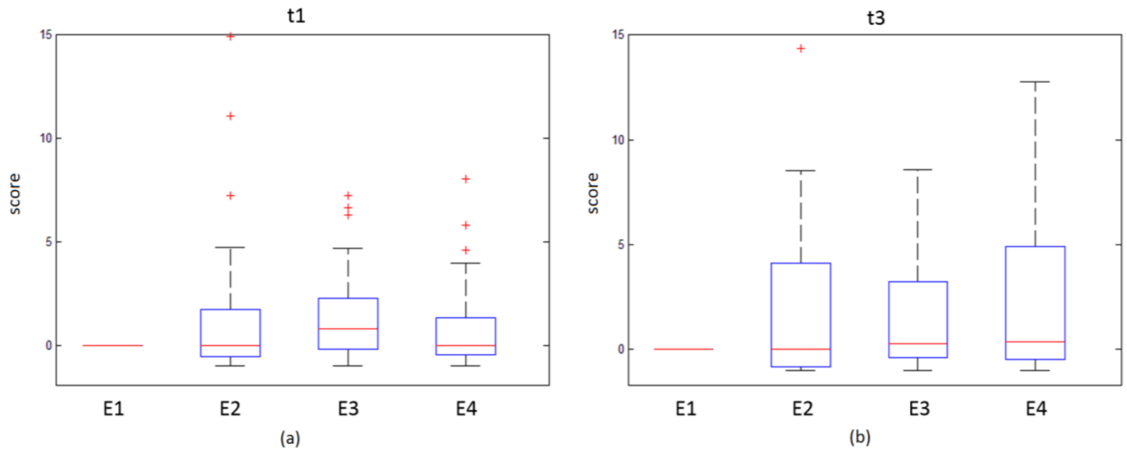

Figure 70. A boxplot of normalized genuine scores values for each of the four experiments for t1 (a) and t3 (b). The y axis is the genuine score and the x axis is the experiment. In the boxplots: the central box represents the central 50% of the data. Its lower and upper boundary lines are at the 25%/75% quantile of the data. A central line indicates the median of the data. Two vertical lines extending from the central box indicate the remaining data outside the central box that are not regarded as outliers. These lines extend maximally to times the height of the central box but not past the range of the data. Outliers (+): these are points indicating the remaining data.

8.0.5 Conclusions obtained for future improvements

The BSF database is the first biometric face database that addresses the accessibility of the face recognition technology for the visually impaired, in the context of taking selfie images using a mobile device. The database contains four experimental conditions following two dichotomies of factors: with and without audio feedback; and with and without instructions. Moreover the database contains additional data sets such as the timestamps and face detection information, including the bounding boxes and face detection confidence values that are associated with each detected face image. The additional meta-data is useful to understand issues such as usability, accessibility in face recognition and/or face detection. This approach effectively records the interaction between the user and the device.

Our experiments on the BSF database can be summarised as follows. First, visually impaired users find that it is difficult to take selfies. The images so-obtained are often blurred, off centre, sometimes only partially visible, and occluded in some cases. Because of the ill positioning of the face images, the alignment-free approach based on the SIFT features gives significantly better face recognition performance than the classical alignment-based approach. For instance, the SIFT approach attains an EER of 12.49% of EER compared to PCA and LDA, each attains 38.04% and 35.10%, respectively. In any case, these results are far from the state of the art results, where EERs applying LDA or PCA are close to 15-20% [107] and ~95% of recognition accuracy applying SIFT [108], both using the ORL Database. This highlights the challenging scenarios inherent in the BSF database that we have collected. The superiority of the alignment-free face recognition algorithm over its alignment-based counterpart is, of course dependent on the kind of scenarios. Further experiments should be made using other databases in order to compare both methods.

The analysis based on the FDC over time shows an increase in FDC value when the user is more habituated to the system and receives feedback. Particularly, the detected face images are likely to be better when the user receives both instructions and audio feedback, i.e., the E4 setting. One the other hand, the differences between E2 (audio-only feedback) and E3 (instruction-only feedback) are not noticeable in the later period of an acquisition session (Time 3). Once the user has received audio feedback and understood how to properly place the device, instructions do not appear to further improve the efficiency of capturing a good-quality face image as gauged by the FDC value. In either E2 or E3 setting, a marginal improvement in FDC is observed over the baseline default E1 setting, i.e., without any feedback or instruction. Therefore, face recognition can be made more accessible for the visually impaired with a careful and considerate design.

Regarding the influence of the time in performance, this work shows a high influence on usability concerns in biometrics, specifically for visually impaired people: the longer the interaction the better the performance. It also covers a gap as we did not find in the literature any other study of the time spent in the face recognition for disabled people (the accessibility studies are scarce in biometrics). Regarding the feedback, the previous works in biometrics (mainly carried out by NIST) were based on provide images to users. In this experiment, we have successfully applied audio feedback in real time and we suggest that it is more effective than previous instructions, but even more effective is to use both modes of feedback jointly.

Regarding accessibility, we have found big variations from one user to another when processing the received feedback. Then, for some of the users the audio is more helpful and other users process the previous instructions better. This fact strengthen our suggestion of provide both feedback modes at the same time (E4). We have also found that users acquired skills on taking self-photos during the process because the number of valid images increases from one experiment to another. Nevertheless, it is not consistent with the performance as we have obtained more valid images in E3 than in E2 but the performance in E2 is better. This fact reinforces our suggestion that audio feedback is more effective than instructions and even that habituation. The performance results obtained are under the state of the art as expected due to evident reasons.

Although our findings suggest that a continuous use of this technology by visually impaired people would lead to improve the final results. It is necessary to extend this work including other different kinds of feedback (e.g. mobile vibration, different sounds, etc.). Another future project could be to implement this work in a real application for common mobile devices including both feedback modes.

Part V: Conclusions and Future work

Chapter 9

Conclusions and Future work

Several conclusions have been extracted during this Thesis and have already been written down in previous sections. Those conclusions were attached to specific experiments. In this section are general conclusions gathered from the continuous work during this Thesis on usability and accessibility in biometric recognition systems. Future works including both topics are explained afterwards.

The contributions of this Thesis include:

- Improvements to the H-B interaction methodology, including several usability evaluations, where further contributions were made in various usability aspects: ergonomics, UX, user acceptance, HCI, etc. This is shown in Chapter 5 (usability experiments) and in Chapter 6 (modifications to the H-B interaction).
- Improvements on the accessibility of the ICT products by means of the integration of biometric recognition systems. This include the development of products starting from simple concepts. This is shown in Chapter 7 (improvements in accessibility).
- Adaptation and application of the EN 301 549 to biometric recognition systems. Including the proceedings to conduct a proper accessibility evaluation (following the standard) and indications for overcoming the incompliances. This is shown in Chapter 7 (accessibility experiments).

9.1 Usability conclusions

The work developed on usability during Thesis is oriented to 2 directions. On one hand, our intention is to improve and make easier the use of biometric recognition systems. On the other hand, validate and complete the H-B interaction methodology was scheduled in our roadmap. Then, first the lessons learned are summarized in a “best practices” section and finally we describe the conclusions regarding the H-B interaction.

9.1.1 Best practices on usability in biometrics

Through all the experiments done, we have acquired skills in the line to design usable biometric systems and evaluate them properly. Next are the main lessons learned during this Thesis:

Usability is not always the performance’s best friend. Even when performance results are excellent, maybe users are not that happy. It is clear that a biometric system which do not recognize people correctly is useless, but a poor usability system too. Nevertheless, when the performance is really good, the probability of having satisfied users is high. The influential factors on usability are too many as shown in Part III.

The time as a key factor. Along with the mobile revolution arrival, biometrics are now embedded in several mobile devices. Then, the development of reliable and fast apps is necessary. When interacting with mobile devices, time (efficiency) may influence other factors (effectiveness, satisfaction, user’s mood, etc.) and involves rejects and disuses of the technology if interactions are too long. According to user’s opinions gathered during this work, time could be as important as performance.

Training to win. Poor usability results come several times due to misuses or lack of practice. Training increase user’s skills and therefore decrease errors and time employed. Users feel more confident and satisfied when proceed smoothly and mistake less. It is highly important to carry out a training session before enrolment, but also before using biometrics after long periods of time.

Subjective over objective? Knowing user’s feelings during the experiments are highly valuable. In the end, when testing usability, the user’s opinions are maybe the most important factor. It is also true that sometimes these opinions may be biased by other factors such as environment, personal situations, etc. A combination of both subjective and objective measurements is convenient.

Leave them alone. Operators and external people may influence users during biometric recognition. Above all else, during evaluations, when users may repeat tasks several times and get distracted easily. It is important that users stay as alone as possible when testing usability. Operators may interact with users only to instruct them at the beginning and in case of systematic errors.

Smartphones boost usability in biometrics. The massive penetration of smartphones in the society involves not only the increase of biometrics usage, but a better usability too. As long as people got skills using smartphones (including camera, microphone, stylus and so on), they already got skills using biometrics (camera, microphone, stylus and so on). The inclusion of biometrics in smartphones made the biometrics penetration smooth.

I trust biometrics…but not always! As we learned during the user acceptance experiment, users feel confidence using biometrics in low-risk tasks (e.g. building entrances) but not in high-risk tasks (e.g. ATM). Biometrics jointly with other security solutions may overcome distrusts (e.g. biometrics and smart cards).

Moods and biometrics. It is clear that moods influence biometric recognition. Overall stress and nervousness have negative effects on algorithms (especially on behavioural modalities!). Incorporate different moods into evaluations may improve databases, making them more complete and realistic.

Testing usability may be a nightmare. As behaviour and moods are relevant factors in HCI, they are highly influential in biometrics also. Moreover, some moods influence other moods and involve changes in people behaviour. Testing these subjective factor is close to impossible.

9.1.2 H-B interaction conclusions

The H-B interaction methodology is a proper way to test usability in biometrics: it is based on several well-known and validated standards and methodologies: the ISO/IEC 19795, the ISO 9241, the HBSI and Common Criteria. Furthermore, the methodology has been previously validated also.

During this work, we have analysed the whole methodology and insisted on those parts considered as weakest. After several experiments made for testing and validating the technology, we have incorporated new updates, which in our point of view, enrich the methodology.

Unfortunately, we have found difficulties. But those difficulties are not a novelty and occur always when measuring usability (in all fields). Those difficulties come from the subjective factors influence, which are indeed hard or impossible to measure.

As long as new biometric modalities arise and the paradigm is constantly changing, the H-B interaction methodology may be regularly updated and keep to date.

9.2 Accessibility conclusions

We started working on accessibility in biometrics encouraged by the idea of increasing the use of the technology. Finally, we are currently designing biometric systems for easing people's life and making some cumbersome tasks less complicated.

Our milestones included the application of the new standard EN 301 549 (which incorporates WCAG 2.0), the updating of the H-B interaction, the testing of biometrics on real users (having different real concerns) and the acquisition of skills on accessibility testing and designing. After all these lessons learned, we realize there is still a long way to cover in order to reach accessible biometrics.

9.2.1 Best practices on accessibility in biometrics

In the same line as the usability best practices, this section shows some of the main outcomes we have collected from the works on accessibility:

Universal accessibility…is that possible? Developing universal accessible apps is nearly impossible due to the wide range of different accessibility issues that exist. Our findings suggest that a convenient accessible design must rely on the subject characteristics. Even when many users succeed in a task, many other could fail. Systems which adapt to users characteristics overcome this problem.

Follow the standards to succeed. Our first design of the ATM app (Experiment 6) was apparently accessible, as it followed the *Libro verde de la accessibilidad* and allowed users to modify the interface. Nevertheless, as it did not follow the standards, did not allow users to properly use the voice over. Using the EN 301 549, the design is much simpler and the coverage of user's needs was made automatically.

Know your audience. As sometimes designing systems for everyone is nearly impossible, the best choice is to get focus on the specific collective who is using the design. Having feedback since early stages (as we saw during the usability experiments) may benefit final designs.

Usable is not accessible and vice versa. Reaching good marks in usability does not involve that a system is accessible. As we saw in Chapter 7, the EN 301 549 has several requirements that usable systems do not need to accomplish. In the other way around, accessible systems may be cumbersome or "too much easy" for some other users.

One little step for the man... Even a little improvement in accessibility may ease people's life and it is extremely rewarding. There are not meaningless steps.

9.3 Future works

Future works start following the biometric recognition paradigm and acting in consequence: analysing usability and accessibility of new designs and products. In accordance, the H-B interaction methodology must be updated frequently. We have recently suggest the H-B interaction methodology as a new ISO/IEC/SC 37 standard, which could keep the methodology to date and add further metrics, factors and so on and so forth.

New modalities and biometric usages bring new usability and accessibility challenges that may be overcome. In the same line, the inclusion of biometrics in mobile devices brings several challenges in usability which are not covered yet. Further usability and accessibility evaluations including new modalities and scenarios should be done.

During this Thesis, we had users with different accessibility concerns: elderly people, blind people and people with physical and cognitive problems. There are other many groups of people having other accessibility concerns who must be taken into account. Future works may include them, making future databases more complete and representative of the population.

The inclusion of biometric recognition in daily tasks may involve the use of biometrics by young people or children. Usability evaluations (specially focusing on user acceptance), are required to properly design friendly systems for these new users groups.

Other new scenarios for biometric recognition may be smart environments. These scenarios require transparency and low intrusiveness. Designs of biometrics for smart environments require several usability analysis and evaluations as well.

Most of the experiments carried out during this Thesis were made using concept products. Following steps must pursuit the commercialization of products designed through a proper usability evaluation (e.g. the H-B interaction after this Thesis).

Other relevant future work is to make people aware of the biometrics benefits and avoid the distrust. The lack of confidence caused many people to reject participating in biometric evaluations, and therefore, rejecting the use of biometric recognition. This may be solved by previous meetings, explanations and longer training sessions.

References

[1] M. F. Theofanos, R. J. Micheals, and B. C. Stanton, 'Biometrics Systems Include Users', *Syst. J. IEEE*, vol. 3, no. 4, pp. 461–468, Dec. 2009.

[2] B. Fernandez-Saavedra, 'Evaluation Methodologies for Security Testing of Biometric Systems beyond Technological Evaluation', Carlos III University of Madrid, 2013.

[3] International Organization for Standardization. ISO/IEC 19795:2006 Biometric performance testing and reporting.

[4] E. Kukula, 'Design and evaluation of the Human-Biometric Sensor Interaction Method', *ETD Collect. Purdue Univ.*, pp. 1–512, Jan. 2008.

[5] International Organization for Standardization. ISO 9241:2010 Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.

[6] The Common Criteria. https://www.commoncriteriaportal.org [Accessed: 27-Nov-2015]

[7] ETSI - New European Standard on accessibility requirements for public procurement of ICT products and services. http://www.etsi.org/news-events/news/754-new-european-standard-on-accessibility-requirements-for-public-procurement-of-ict-products-and-services. [Accessed: 27-Nov-2015].

[8] R. Blanco-Gonzalo and R. Sanchez-Reillo, 'Biometrics on Mobile Devices', in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Springer US, 2014, pp. 1–8.

[9] *Younger users would rather have biometrics than passwords*. http://www.computerweekly .com/news/2240238497/Younger-users-would-rather-have-biometrics-than-passwords [Accessed: 27-Nov-2015]

[10] A. Adams and M. A. Sasse, 'Users Are Not the Enemy', *Commun ACM*, vol. 42, no. 12, pp. 40–46, Dec. 1999.

[11] WHO. *World report on disability*. http://www.who.int/disabilities/world_report/2011/ report.pdf [Accessed: 27-Nov-2015]

[12] V. Ravi and K. Nishanth, 'A Computational Intelligence Based Online Data Imputation Method: An Application For Banking', *J. Converg.*, vol. 9, no. 4, pp. 633–650, Sep. 2013.

[13] P. R., Singh Singh and M. Duhan, 'An effective implementation of security based algorithmic approach in mobile adhoc networks', *Hum.-Centric Comput. Inf. Sci.*, vol. 4, no. 1, p. 7, 2014.

[14] I. Y. Lee and S. W. Park, 'Anonymous Authentication Scheme based on NTRU for the Protection of Payment Information in NFC Mobile Environment', *J. Inf. Process. Syst.*, vol. 9, no. 3, p. 461,476, Sep. 2013.

[15] S. Marcel et al., 'On the Results of the First Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation', in *Recognizing Patterns in Signals, Speech, Images and Videos*, vol. 6388, D. Ünay, Z. Çataltepe, and S. Aksoy, Eds. Springer Berlin Heidelberg, 2010, pp. 210–225.

[16] A.K. Jain, P. Flynn, and A.A. Ross, *Handbook of Biometrics*. NJ, USA: Springer-Verlag, 2007.

[17] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[18] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, 'The FERET database and evaluation procedure for face-recognition algorithms', *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[19] ICAO TAG/MRTD, 'Guidelines on electronic - Machine Readable Travel Documents & Passenger Facilitation'. 17-Apr-2008.

[20] P. Campisi, E. Maiorana, M. Lo Bosco, and A. Neri, 'User authentication using keystroke dynamics for cellular phones', *IET Signal Process.*, vol. 3, no. 4, pp. 333–341, Jul. 2009.

[21] E. Sesa-Nogueras. M. Faundez-Zanuy, *Biometric recognition using online uppercase handwritten text*, *Pattern Recognit.*, vol. 45, no. 1, pp. 128 – 144, 2012.

[22] Z. Jian and S. Wan-juan, ‘Handwritten Numerical String Recognition Based on SVM Verifier’, in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2011 International Conference on*, 2011, vol. 2, pp. 186–188.
[23] R. Guest and O. Miguel-Hurtado, ‘Enhancing static biometric signature verification using Speeded-Up Robust Features’, in *Security Technology (ICCST), 2012 IEEE International Carnahan Conference on*, 2012, pp. 213–217.
[24] M. Faundez-Zanuy, ‘On-line signature recognition based on VQ-DTW’, *Pattern Recognit.*, vol. 40, no. 3, pp. 981–992, Mar. 2007.
[25] ‘Off-line signature verification and forgery detection using fuzzy modeling’, *Pattern Recognit.*, vol. 38, no. 3, pp. 341 – 356, 2005.
[26] L. Hu and Y.-D. Wang, ‘On-line signature verification based on fusion of global and local information’, in *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR ’07. International Conference on*, 2007, vol. 3, pp. 1192–1196.
[27] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, ‘MCYT baseline corpus: a bimodal biometric database’, *Vis. Image Signal Process. IEE Proc.* -, vol. 150, no. 6, pp. 395–401, Dec. 2003.
[28] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, ‘SVC2004: First International Signature Verification Competition’, in *Biometric Authentication*, vol. 3072, D. Zhang and A. Jain, Eds. Springer Berlin Heidelberg, 2004, pp. 16–22.
[29] R. Cappelli, D. Maio, D. Maltoni, J. L. Wayman, and A. K. Jain, ‘Performance evaluation of fingerprint verification systems’, *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 28, no. 1, pp. 3–18, Jan. 2006.
[30] S. Eliabeth S, B. Thomas, and J. J. Kizhakkethottam, ‘Analysis of effective biometric identification on monozygotic twins’, in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, 2015, pp. 1–6.
[31] R. Blanco-Gonzalo, R. Sanchez-Reillo, R. Ros-Gomez, B. Fernandez-Saavedra, ‘User Acceptance of Planar Semiconductor Fingerprint Sensors’, in *2015 IEEE International Carnahan Conference on Security Technology (ICCST).* To be published.
[32] V. Sasikala and V. LakshmiPrabha, ‘A comparative study on the swarm intelligence based feature selection approaches for fake and real fingerprint classification’, in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, 2015, pp. 1–8.
[33] A. Rattani and A. Ross, ‘Minimizing the impact of spoof fabrication material on fingerprint liveness detector’, in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014, pp. 4992–4996.
[34] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, ‘FVC2002: Second Fingerprint Verification Competition’, in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, vol. 3, pp. 811–814 vol.3.
[35] J. Sanchez del Rio, C. Conde, A. Tsitiridis, J. R. Gomez, I. Martin de Diego, and E. Cabello, ‘Face-based recognition systems in the ABC e-gates’, in *Systems Conference (SysCon), 2015 9th Annual IEEE International*, 2015, pp. 340–346.
[36] M. D. Marsico, C. Galdi, M. Nappi, and D. Riccio, ‘FIRME: Face and Iris Recognition for Mobile Engagement’, *Image Vis. Comput.*, vol. 32, no. 12, pp. 1161 – 1172, 2014.
[37] Y. Kim, J.-H. Yoo, and K. Choi, ‘A motion and similarity-based fake detection method for biometric face recognition systems’, *Consum. Electron. IEEE Trans. On*, vol. 57, no. 2, pp. 756–762, May 2011.
[38] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, ‘Face recognition vendor test 2002’, in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003*, 2003.
[39] P. J. Phillips, P. J. Flynn, T. Scruggs, and B. et al, ‘Overview of the face recognition grand challenge’, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, 2005, vol. 1.

[40] N. Houmani, A. Mayoue, S. Garcia-Salicetti, B. Dorizzi, M. I. Khalil, M. N. Moustafa, H. Abbas, D. Muramatsu, B. Yanikoglu, A. Kholmatov, M. Martinez-Diaz, J. Fierrez, J. Ortega-Garcia, J. Roure Alcobé, J. Fabregas, M. Faundez-Zanuy, J. M. Pascual-Gaspar, V. Cardeñoso-Payo, and C. Vivaracho-Pascual, 'BioSecure signature evaluation campaign (BSEC'2009): Evaluating online signature algorithms depending on the quality of signatures', *Pattern Recognit.*, vol. 45, no. 3, pp. 993–1003, Mar. 2012.

[41] S. Mansfield-Devine, 'Biometrics for mobile devices struggle to go mainstream', *Biom. Technol. Today*, vol. 2011, no. 9, pp. 10 – 11, 2011.

[42] D. H. Cho, K. R. Park, and D. W. Rhee, 'Real-time iris localization for iris recognition in cellular phone', in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on*, 2005, pp. 254–259.

[43] K. Cheng and A. Kumar, 'Contactless finger knuckle identification using smartphones.', in *BIOSIG*, 2012, pp. 1–6.

[44] D. Gentles and S. Sankaranarayanan, 'Biometric secured mobile voting', in *Internet (AH-ICI), 2011 Second Asian Himalayas International Conference on*, 2011, pp. 1–6.

[45] SecurePhone Project. http://www.secure-phone.info [Accessed: 27-Nov-2015]

[46] A. Goode, 'Mobile Phone Biometric Security; Analysis and Forecasts 2011 - 2015'. Jun-2011.

[47] P. N. A. Fahmi, E. Kodirov, D.-J. Choi, G.-S. Lee, A. Mohd Fikri Azli, and S. Sayeed, 'Implicit authentication based on ear shape biometrics using smartphone camera during a call', in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2012, pp. 2272–2276.

[48] R. Blanco-Gonzalo, R. Sanchez-Reillo, O. Miguel-Hurtado, and J. Liu-Jimenez, 'Performance evaluation of handwritten signature recognition in mobile environments', *Biom. IET*, vol. 3, no. 3, pp. 139–146, Sep. 2014.

[49] E. Vazquez-Fernandez, H. Garcia-Pardo, D. Gonzalez-Jimenez, and L. Perez-Freire, 'Built-in face recognition for smart photo sharing in mobile devices', in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 2011, pp. 1–4.

[50] J. G. Casanova, C. S. Ávila, A. de Santos Sierra, G. B. del Pozo, and V. J. Vera, 'A Real-Time In-Air Signature Biometric Technique Using a Mobile Device Embedding an Accelerometer', in *Networked Digital Technologies*, vol. 87, F. Zavoral, J. Yaghob, P. Pichappan, and E. El-Qawasmeh, Eds. Springer Berlin Heidelberg, 2010, pp. 497–503.

[51] R. Sanchez-Reillo, R. Blanco-Gonzalo, J. Liu-Jimenez, M. Lopez, and E. Canto, 'Universal access through biometrics in mobile scenarios', in *Security Technology (ICCST), 2013 47th International Carnahan Conference on*, 2013, pp. 1–6.

[52] A. de-Santos-Sierra, C. Sánchez-Ávila, G. B. del Pozo, and J. Guerra-Casanova, 'Unconstrained and Contactless Hand Geometry Biometrics', *Sensors*, vol. 11, no. 11, pp. 10143–10164, Oct. 2011.

[53] International Organization for Standardization. ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. 1st ed, 1998, pp. 28.

[54] M. El-Abed, R. Giot, B. Hemery, and C. Rosenberger, 'A study of users' acceptance and satisfaction of biometric systems', in *2010 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2010, pp. 170–178.

[55] ACM Curriculum Committee on Computer Science. (1968). Curriculum 68: Recommendations for academic programs in computer science. Communications of the ACM, 11, 3 (March), 151-197.

[56] J. Nielsen, 'Introduction to Usability'. Apr-2012. http://www.nngroup.com/articles/usability-101-introduction-to-usability [Accessed: 27-Nov-2015]

[57] J. Nielsen, '10 Usability Heuristics for User Interface Design'. Jan-1995. http://www.nngroup.com/articles/ten-usability-heuristics [Accessed: 27-Nov-2015]

[58] N. US Department of Commerce, 'NIST Manuscript Publication Search'. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=50911&division=894. [Accessed: 27-Nov-2015].
[59] M. F. Theofanos, B. Stanton, S. Orandi, and R. Micheals, 'Effects of scanner height on fingerprint capture', presented at the Proceedings of human factors and ergonomics society annual conference, 2007.
[60] M. F. Theofanos, B. Stanton, C. Sheppard, and R. Micheals, 'Usability Testing of Face Image Capture for US Ports of Entry', in *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008. BTAS 2008*, 2008, pp. 1–6.
[61] M. F. Theofanos, Brian Stanton, and Cari A. Wolfson, *Usability and Biometrics - Ensuring Successful Biometric Systems*. .
[62] 'UKPS biometric enrolment trial', *Biom. Technol. Today*, vol. 13, no. 7, pp. 6–7, Jul. 2005.
[63] E. Kukula and S. Elliott, 'Implementing Ergonomic Principles in a Biometric System: A Look at the Human Biometric Sensor Interaction (HBSI)', in *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International*, 2006, pp. 86–91.
[64] Adam Wamsley, Stephen Elliott, Carl Dunkelberger, and Mitch Mershon, 'Analysis of Slap Segmentation and HBSI errors across different force levels'.
[65] M. Brockly, R. Guest, S. Elliott, and J. Scott, 'Dynamic signature verification and the human biometric sensor interaction model', in *2011 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2011, pp. 1 –6.
[66] 'European Concept for Accessibility' http://www.ub.edu/escult/Water/N05/eca_full.pdf [Accessed: 27-Nov-2015]
[67] International Organization for Standardization. ISO 26800:2011. Ergonomics -- General approach, principles and concepts.
[68] W3C Consortium. http://www.w3.org [Accessed: 27-Nov-2015]..
[69] *EN 301 549 Accessibility requirements suitable for public procurement of ICT products and services in Europe*. http://www.etsi.org/deliver/etsi_en/301500_301599/301549 /01.00.02 _30/en_301549v010002v.pdf [Accessed: 27-Nov-2015].
[70] P. Korn, L. Martinez-Normand, M. Pluke, A. Snow-Weaver, and G. Vanderheiden, 'Guidance on Applying WCAG 2.0 to Non-Web Information and Communications Technologies (WCAG2ICT). W3C Working Group Note 5 September 2013. http://www.w3.org /TR/wcag2ict [Accessed: 27-Nov-2015].
[71] R. Wong, N. Poh, J. Kittler, and D. Frohlich, 'Interactive quality-driven feedback for biometric systems', in *2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010, pp. 1–7.
[72] R. Wong, N. Poh, J. Kittler, and D. Frohlich, 'Towards inclusive design in mobile biometry', in *2010 3rd Conference on Human System Interactions (HSI)*, 2010, pp. 267–274.
[73] International Organization for Standardization. ISO/IEC TR 29194:2014 Biometrics - Guide on the inclusive design and operation of biometric systems.
[74] C. M. Barnum, 'Conducting a usability test', in *Usability Testing Essentials*, Boston: Morgan Kaufmann, 2011, pp. 199–237.
[75] N. Nielsen, How Many Test Users in a Usability Study? 'http://www.nngroup.com/articles/how-many-test-users [Accessed: 27-Nov-2015].
[76] International Organization for Standardization. ISO 19794-7 Signature/sign time series data.
[77] B. L. Van, S. Garcia-Salicetti, and B. Dorizzi, 'On Using the Viterbi Path Along With HMM Likelihood Information for Online Signature Verification', *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 37, no. 5, pp. 1237–1247, 2007.
[78] O. Miguel-Hurtado, 'Online Signature Verification Algorithms and Development of Signature International Standards', University Carlos III of Madrid, 2011.
[79] R. Blanco-Gonzalo, O. Miguel-Hurtado, R. Sanchez-Reillo, and A. Gonzalez-Ramirez, 'Usability analysis of a handwritten signature recognition system applied to mobile scenarios', in *2013 47th International Carnahan Conference on Security Technology (ICCST)*, 2013, pp. 1–6.

[80] E. Kukula and R. Proctor, ‘Human-Biometric Sensor Interaction: Impact of Training on Biometric System and User Performance’, in *Human Interface and the Management of Information. Information and Interaction*, vol. 5618, G. Salvendy and M. J. Smith, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 168–177.

[81] J. Fierrez-Aguilar, S. Krawczyk, J. Ortega-Garcia, and A. K. Jain, ‘Fusion of local and regional approaches for on-line signature verification’, in *Proceedings of the 2005 international conference on Advances in Biometric Person Authentication*, Berlin, Heidelberg, 2005, pp. 188–196.

[82] R. Blanco-Gonzalo, L. Diaz-Fernandez, O. Miguel-Hurtado, and R. Sanchez-Reillo, ‘Usability Evaluation of Biometrics in Mobile Environments’, in *Human-Computer Systems Interaction: Backgrounds and Applications 3*, Z. S. Hippe, J. L. Kulikowski, T. Mroczek, and J. Wtorek, Eds. Springer International Publishing, 2014, pp. 289–300.

[83] T. H. Falk, C. Tam, H. Schwellnus, and T. Chau, ‘Grip force variability and its effects on children’s handwriting legibility, form, and strokes’, *J. Biomech. Eng.*, vol. 132, no. 11, p. 114504, Nov. 2010.

[84] W. E. Morrison, E. C. Haas, D. H. Shaffner, E. S. Garrett, and J. C. Fackler, ‘Noise, stress, and annoyance in a pediatric intensive care unit’, *Crit. Care Med.*, vol. 31, no. 1, 2003.

[85] C. N. Karson, ‘Spontaneous eye-blink rates and dopaminergic systems’, *Brain J. Neurol.*, vol. 106 (Pt 3), pp. 643–653, Sep. 1983.

[86] R. Blanco-Gonzalo, O. Miguel-Hurtado, A. Mendaza-Ormaza, and R. Sanchez-Reillo, ‘Handwritten signature recognition in mobile scenarios: Performance evaluation’, in *2012 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2012, pp. 174 –179.

[87] R. Blanco-Gonzalo, R. Sanchez-Reillo, and N. Poh, ‘On the effect of time efficiency in DSV under stress’, in *2014 International Carnahan Conference on Security Technology (ICCST)*, 2014, pp. 1–5.

[88] Public Report on an Evaluation of 3 fingerprint sensors and 2 algorithms. http://idtestinglab.uc3m.es/Publications [Accessed: 27-Nov-2015].

[89] E. Tabassi, C. L. Wilson, and C. I. Watson, *Fingerprint image quality: Nistir 7151,” tech. rep*. 2004.

[90] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. Mccabe, S. Janet, and K. Ko, *User’s Guide to NIST Biometric Image Software (NBIS)*. .

[91] W. Rankl and W. Effing, *Smart Card Handbook*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[92] R. Sanchez-Reillo, R. Alonso-Moreno, and J. Liu-Jimenez, ‘Smart Cards to Enhance Security and Privacy in Biometrics’, in *Security and Privacy in Biometrics*, P. Campisi, Ed. Springer London, 2013, pp. 239–274.

[93] What Is NFC?, *NFC Forum*. http://nfc-forum.org/what-is-nfc [Accessed: 27-Nov-2015].

[94] R. Blanco-Gonzalo, R. Sanchez-Reillo, L. Martinez-Normand, B. Fernandez-Saavedra, and J. Liu-Jimenez, ‘Accessible Mobile Biometrics for Elderly’, in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, New York, NY, USA, 2015, pp. 419–420.

[95] Libro verde de la accesibilidad en España: diagnóstico... (SID). http://sid.usal.es/libros/discapacidad/5999/8-1/libro-verde-de-la-accesibilidad-en-espana-diagnostico-de-situacion-y-bases-para-elaborar-un-plan-integral-de-supresion-de-barreras.aspx [Accessed: 27-Nov-2015].

[96] Are You Giving Your Users Positive Feedback? http://www.smashingmagazine.com/2012/07/17/are-giving-users-positive-feedback [Accessed: 27-Nov-2015].

[97] L. Martinez-Normand, ‘Prototype of EN 301 549 Decision tree (Demonstration). ETSI workshop on the Human side of technology: Human factors in ICT. June, 2014.

[98] M. Fairhurst. 'Age Factors in Biometric Processing' (Professional Applications of Computing, 2013). IET Digital Library

[99] R. Blanco-Gonzalo, R. Sanchez-Reillo, O. Miguel-Hurtado, and E. Bella-Pulgarin, 'Automatic usability and stress analysis in mobile biometrics', *Image Vis. Comput.*, vol. 32, no. 12, pp. 1173–1180, Dec. 2014.

[100] R. Blanco-Gonzalo, R. Sanchez-Reillo, O. Miguel-Hurtado, and J. Liu-Jimenez, 'Performance evaluation of handwritten signature recognition in mobile environments', *IET Biom.*, Sep. 2013.

[101] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, 'Sheep, Goats, Lambs and Wolves A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation', in *INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING*, 1998.

[102] X. Tan and B. Triggs, 'Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions', *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[103] 'Principal Component Analysis and Factor Analysis', in *Principal Component Analysis*, Springer New York, 2002, pp. 150–166.

[104] V. Štruc and N. Pavešić, *Gabor-Based Kernel Partial-Least-Squares Discrimination Features for Face Recognition* ∗. 2007.

[105] V. Štruc and N. Pavešić, 'The Complete Gabor-fisher Classifier for Robust Face Recognition', *EURASIP J Adv Signal Process*, vol. 2010, pp. 31:1–31:13, Feb. 2010.

[106] The orl database of faces'. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html [Accessed: 27-Nov-2015].

[107] G. Dai and Y. Qian, 'A Gabor direct fractional-step LDA algorithm for face recognition', in *2004 IEEE International Conference on Multimedia and Expo, 2004. ICME '04*, 2004, vol. 1, pp. 61–64 Vol.1.

[108] P. Kamencay, M. Breznan, D. Jelsovka, and M. Zachariasova, 'Improved face recognition method based on segmentation algorithm using SIFT-PCA', in *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, 2012, pp. 758–762.