Investigation into the Reliability of Contactless Biometric Systems

By

Mohammadreza Azimi

Thesis submitted to the

Institute of Control and Computation Engineering

in partial fulfillment of the requirements

for a Ph.D degree in

Electrical and Computer Engineering

Institute of Control and Computation Engineering

Faculty of Electronics and Information Technology

Warsaw University of Technology

Abstract

Biometrics are popular nowadays, as we cannot lose them and they are secure. A very significant problem with biometric solutions is their lack of performance, as the matching accuracy of biometric recognition systems can be affected by various social factors.

This thesis reports on our new findings regarding the influences of certain social factors on biometric recognition. Three methods were chosen and related questions were answered:

a- Iris: this part looks at the reliability test of the iris recognition system under the influence of diabetes. A new database has been collected. We have used various matchers in order to obtain similarity scores between the captured samples. We found that, while there is no obvious impairment on non-healthy irides, the accuracy of the recognition system is higher when dealing with healthy people. In other words, it is harder to recognize people who suffer from diabetes due to certain non-obvious disorders in their iris textures. Gender and age dependency studies are also available.

b- Voice: this part looks at the effect of "Morning Voice" on text-independent speaker recognition. It presents an investigation on the effect of the time of day on the matching accuracy of the voice recognition system. A new database has been collected and offered. The database contains a dataset of thirty people. We have collected 1780 voice samples. There were two different data collection sessions: a. participants were asked to record their voice after getting up, using their own smartphone devices (nine hundred and sixteen morning voice samples were recorded), and b. participants were asked to record their voice samples during the day (more than eight hundred and eighty-four samples were collected from the same users). Each sample lasts for six seconds, at a bit rate of 705 kbps. All the participants are native Persian speakers. In order to conduct numerical experiments, a pre-trained VGG-Speaker is used. An All-versus-All comparison scenario is carried out. The intrasession comparison scores are better than the intersession comparisons. For the evening versus evening comparison scenario, an equal error rate (EER) of 1.46% was achieved. For the morning versus evening comparison scenario, the EER is increased to 10.2%.

c- Face: this part explores the joint influence of makeup and facial expressions on the matching accuracy of the facial recognition system. We consider the question of whether or not the effect of makeup and facial expression are correlated. In fact, while a single effect of each is not significant, the joint effect is. We implemented three state of the art approaches, namely python face recognition dlib, Verilook and VGGFace. While the application of "light makeup" on angry faces showed no statistically significant differences, for fearful or happy faces the comparison score differs significantly.

For each of the mentioned cases, we have built a classifier to make the system more reliable. For instance, building a makeup detection algorithm can improve facial recognition.

Therefore, the central aim of this thesis is testing the performance of a biometric recognition system under the influence of social problems that have the potential to degrade a system's matching accuracy (the effects of diabetes on the iris, the morning effect on the voice and a combination of mood variation and makeup influences on facial appearance were considered). The next goal was to make mobile contactless biometric systems robust against the impacts of the predetermined influential parameters. The final goal of this work was to achieve an EER in all the mentioned cases.

In this thesis:

a- From a database of more than 1900 iris images from 509 eyes (723 diabetic iris images from 161 eyes and 1183 healthy iris images from 348 ones), we used three different matchers (open source) and found that accuracy was consistently higher in those images of eyes from people who do not have diabetes.

b- We present an investigation on the effect of time of day on the matching accuracy of a speaker recognition system. We have collected 1780 voice samples donated by 30 people. The intrasession comparison scores are better in than the intersession comparisons. For the evening versus evening comparison scenario an EER of 1.46% was achieved. For the morning versus evening comparison scenario, the EER increases to 10.2%.

c- An EER of 4.68% was achieved when identifying faces under the joint influences of full makeup and mood variation, while the EER under the effect of each of these factors separately is less than 1%.

In a nutshell, the study highlights the limits of unimodal biometrics and cautions against the widespread use of methods that only perform well under optimal circumstances, without taking into account certain relatively common conditions.

Improving the robustness of biometric systems can enhance the popularity of contactless mobile biometric systems. Robustness, which is defined as survivability under failure or attack, is one of the most important properties of a system. A biometric system should perform well under any circumstances. In this thesis, we have detected several issues and fixed them by using the same methodology.

To tackle the detected problems in all three cases, we have tried to build sophisticated diabetes, makeup and morning voice detection algorithms that can improve iris, face and speaker recognition.

While the accuracy of the makeup detector and morning voice detector was up to 95%, it was not possible for us to diagnose diabetes using the iris texture.

Acknowledgements

First and foremost, I am deeply indebted and especially grateful to Professor Andrzej Pacut, for his kind, gentle, and supporting supervision. It has been an exceptional honor for me to work under his supervision. I would also like to thank Professor Seyed Ahmad Rasoulinejad who helped me in connection with the iris database collection. Lastly, but by no means least, many thanks are owed to my family who supported me throughout.

I should mention that, I was sponsored by the funding source from AMBER with sponsorship from the Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020, under Grant Agreement No. 675087.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other university. I hereby confirm that this thesis is completely my own work.

I agree to the university library copying or lending this thesis on request.

Signed,

Mohammadreza Azimi

October, 2020.

Table of contents

List of Figures

List of Tables

1. Introduction

Biometric recognition systems are increasingly popular and prevalent in our everyday lives, especially on mobile cell phones. Biometrics technology is used to authenticate and verify users by measuring and analyzing their personal and behavioral traits. Owing to its high security and unique characteristics, it is used in a variety of applications in a wide range of industries, from banking and commercial security to healthcare. Today, there are many biometrics recognition systems that have been introduced by researchers and scientists. Each of these biometrics systems applies a special authentication/identification methodology and recognition approach that determines whether they can be used in practice for a specific purpose or not. The integration of biometric systems such as the face, voice and, iris in handheld mobile devices has been made easier due to the widespread availability of the necessary hardware, such as a camera, in modern phones and tablets. This chapter sets out an overview then looks at the biometrics characteristics, and finally discusses the current research in the application of mobile biometrics.

1.1. Overview

Using a biometrics identifier (which can be categorized as a physiological characteristic, including the iris [1], the ear [2], fingerprints [3], the retina [4], hand geometry [5], the face [6] and the voice [7], as opposed to behavioral characteristics, which include signature [8], gait [9] and keystroke [10]) is more reliable for verification and identification tasks than traditional methods such as passwords and identity cards [11]. Due to the strengths of mobile biometric authentication and identification in comparison with traditional methods such, as patterns, PINs and password, and due to the fact that the number of emerging options for this purpose on advanced smartphones is rapidly growing as they become smarter, the use of such technology is more popular than ever before. However, one very important issue concerning biometric solutions is their lack of performance and problems with reliability.

"System reliability testing" relates to testing the ability of a system to perform well over a particular period of time and under given environmental conditions. The same definition can be adopted for testing the reliability of biometric systems. The term "biometric system reliability" refers to the ability of the biometric system to perform verification/identification tasks regardless of the influences of various biological/behavioral/environmental parameters. In general, there are

three types of factors that can change the biometric data. The degradation of biometric performance occurs under the influences of three types of factors:

- Biological factors (such as aging): the biological factors relate to how the person looks physically. These factors usually have permanent effects on the biometric trait as they may change the physiological characters of biometrics (for instance, a retinal detachment can permanently change the trait of the iris, and therefore render biometric data from before the detachment unusable).
- Behavioral factors (such as facial expression and makeup): the behavioral factors relate to how the user behaves. These factors usually have temporary effects on the biometric traits, (for instance, facial expressions are visually detectable changes in facial appearance). Cultural restrictions are some well-known social issues that can be categorized in the same class, as they can change the way the user behaves.
- Environmental factors (such as noise or lighting conditions): the environmental factors relate mainly to the shortcomings of the sensors used and the obstacles imposed by unconstrained environments. We can control the environmental factors by controlling the conditions under which samples are acquired.

With respect to the type(s) of factor(s) involved, the differentiation between behavioral and biological biometrics can be of great importance for many reasons.

1.2. Mobile Biometrics

Biometric systems can be used in two different and well known ways: Firstly, as a verifier to be able to authenticate the same user, and secondly, as an identifier to prevent impostors from using the smartphone for unauthorized purposes [12]. The new generation of smartphones, offering so many new features and novel functionalities, give us this ability to use biometrics system for verification tasks. Smartphones can be used both as a mobile telephone and as a handheld computer, meaning that these devices have replaced the need to carry around other devices to perform tasks. However, much of the data are highly important, private and extremely valuable. Some of the most comprehensive surveys regarding mobile biometric systems can be found in references [13-21]. The implementation of biometric recognition on a mobile device is different from the environment on a dedicated device. For instance, to achieve better recognition

performance, the iris images captured using smartphone cameras should first be preprocessed [22, 23].

The urgent need for more secure mobile biometric systems is increasing as the popularity of such systems has grown, especially among the youth. It is completely rational to expect that the reliability of the system can be challenged under the impact of several influential parameters. Especially in real life scenarios, the reliability of biometrics recognition systems can be significantly affected by various social factors. These social factors include physiological/biological (such as age, gender, etc.) and behavioral (such as smoking, stress, and, alcohol consumption), as well as environmental factors (such as noise, lighting, etc.)

1.3. Objectives of the Thesis

In this chapter, a general description of mobile biometrics and examples of the successful application of these systems are presented. The main objectives of this thesis is to explore the limits of various kinds of biometric techs, and to explore cautions against the widespread use of the methods that only perform well under optimal circumstances, without accounting for relatively common conditions.

The main statement is that,

"The reliability of biometrics can be significantly affected by social factors, but that this can be mitigated by proper data analysis techniques"

This involves investigating and examining how the influential parameters can affect the usage of biometric recognition systems. In fact, analyzing the performance variation of biometric systems under different conditions is necessary in order to have a comprehensive examination of the reliability of the biometric recognition systems. Enhancing the matching accuracy of the system is the end goal.

In Chapter 2, the main factors that can challenge the reliability of biometric recognition systems will firstly be introduced and categorized. Then there is a brief review of the state of the art on the main challenges for the reliability of biometric systems under the influences of social factors (both physiological factors such disease, age, gender and aging, etc. as well as behavioral factors such as emotional state, smoking habit, alcohol consumption, makeup, etc.) As a result, a bibliographical review is presented in Chapter 2 using the following goals:

1- Reviewing the state of the art on biometrics recognition under the influence of physiological factors:
    - Reviewing the state of the art on the effects of time (both template aging and physiological aging).
    - Reviewing the state of the art on the effects of disease on the performance of biometric recognition systems.
2- Reviewing the state of the art on biometrics recognition under the influence of behavioral factors:
    - Reviewing the state of the art on the effects of variations in the emotional state on the performance of biometric recognition systems.
    - Briefly reviewing the state of the art on the effects of cultural differences on the acceptance of technology and other related issues (hesitation is one of the issues relating to biometrics. Users sometimes hesitate to use new technology: for instance, in Arab communities, it is very hard to ask female users to provide facial images, because they usually use a borga and cover their faces).

As explained above, we have categorized the social factors into different groups: including biological and behavioral factors. To give an example of the meaningful effect of the biological factor on the reliability of system, the following sub-statement will be proved:

S1. "Iris recognition is less effective for people with diabetes type II, regardless of their gender and age. By classifying the illness-affected samples, the performance of the system will improve."

To prove the first sub-statement (S1), Chapters 3-5 investigate iris recognition under the influence of diabetes. The age and gender dependency of the results is studied. In order to tackle the detected diabetes-related problem, we may build strong diabetes detection algorithms that can improve biometric recognition. Our classifier, with an accuracy of near 70%, is still insufficient to discriminate between healthy and illness-affected iris samples.

The next sub-statement is related to the combined effects of behavioral and biological factors on the performance evaluation results:

S2. "Morning voice can challenge the reliability of text independent speaker recognition systems. We can reduce this effect by detecting morning voice."

Even though the physiological features of speech may remain the same, there is little doubt that the behavioral part of the voice changes during the course of a day. To prove the second sub-statement (S2), Chapter 6 studies the effect of the time of day on the reliability of a text-independent speaker recognition system. A computer vision tool is then proposed to detect morning voice with a high degree of accuracy (more than 97%), which can then help to diminish the effects of morning voice. Interclass similarity scores were much lower than intraclass similarity scores. Hence, we enhanced the reliability of speaker recognition systems through the use of a support vector machine-based classifier.

The last sub-statement to prove is related to the effect of behavioral factors:

S3. "The simultaneous effects of makeup and facial expressions can affect the matching accuracy of a facial recognition system. Reliability can be enhanced using makeup detectors"

To prove the last sub-statement (S3), Chapter 7 presents an investigation into the reliability of facial recognition systems under the simultaneous influences of mood variation and makeup. We have proposed a SVM-based classifier to detect whether the user wears makeup or not, with an accuracy of 96%.

Chapter 8 offers a brief review of the main findings and reports on the conclusions. In Chapter 2, we discuss other challenging contemporary issues in mobile biometric recognition, and present the contributions of this dissertation that address these issues.

It should be noted that some of the chapters (Chapters 3, 5 and 7) have previously been published as journal articles. Chapter 4 is an extended version of a conference article that was also previously published. A short statement has been made beneath each chapter title in order to indicate that the chapter was previously published.

1.4. Publications

1.4.1. Peer reviewed journals

The following ISI papers were written under the affiliation of the Warsaw University of Technology (AMBER, EU Horizon 2020, under Grant Agreement No. 675087):

- M Azimi, A Pacut – An Investigation into the Reliability of Facial Recognition Systems under the Simultaneous Influences of Mood Variation and Makeup, Computers and Electrical Engineering, 85, 2020, https://doi.org/10.1016/j.compeleceng.2020.106662

- M Azimi, SA Rasoulinejad, A Pacut – Age dependency of the diabetes effects on the iris recognition systems performance evaluation results, Biomedical Engineering/Biomedizinische Technik, 2020, https://doi.org/10.1515/bmt-2019-0246
- M Azimi, SA Rasoulinejad, A Pacut – Iris recognition under the influence of diabetes, Biomedical Engineering/Biomedizinische Technik 2019, https://doi.org/10.1515/bmt-2018-0190

1.4.2. Peer-reviewed international conference papers

- M Azimi, A Pacut – The Effects of Social Problems and Human Factors on Biometric System Reliability: A Review, in: Intelligent Systems Conference (IntelliSys) 2020, September 2020 in Amsterdam, The Netherlands, https://link.springer.com/chapter/10.1007/978-3-030-55187-2_10
- M Azimi, SA Rasoulinejad, A Pacut – The Effects of Gender Factor and Diabetes Mellitus on the Iris Recognition System's Accuracy and Reliability, in: IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) 2019 https://ieeexplore.ieee.org/document/8936757
- M Azimi, A Pacut – The effect of gender-specific facial expressions on face recognition system's reliability, in: IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), 2018 https://ieeexplore.ieee.org/document/8402705
- M. Azimi, Effects of Facial Mood Expressions on Face Biometric Recognition System's Reliability, in: IEEE 1st International Conference on Advanced Research in Engineering Sciences (ARES), 1-5. https://ieeexplore.ieee.org/document/8723292

2. The Effects of Social Issues on the Reliability of Biometric Systems: A Review

Biometrics is defined as what we are, as opposed to what we have (e.g. smart cards), or what we know (passwords). Today, smartphones are equipped with biometric tech such as such as voice recognition as well as facial and fingerprint scanners. Within a person there is some variability in biometrics recognition, which is related to the impact of well-known influential parameters. These include: biological factors (such as aging), behavioral factors (such as facial expression and makeup), as well as environmental factors (such as noise or lighting conditions).

The literature review is a part of this thesis, with the main goals being to synthesize the authors' understanding of the research subject and to justify future research on the topic [24]. This chapter contains a report on the state of the art of the research being conducted on the influences of the most common social problems that can affect the matching accuracy of biometric systems.

In order to review the state of the art on the main challenges for biometric systems reliability under influences of the social problems (such as aging, mood variation and cultural differences) this chapter is written with the following goals in mind:

- Reviewing the state of the art on the effects of time passing on the body (physiological aging).
- Reviewing the state of the art on the effects of time passing on biometric samples (template aging).
- Reviewing the state of the art on biometric recognition under the influence of diseases.
- Reviewing the state of the art on the effects of habituation on the performance of the biometric recognition systems.
- Reviewing the state of the art on the effects of cultural differences on technology acceptance and other related issues. Hesitation is also one of the social issues relating to the variability of biometric recognition.

To cover these goals and to prove the sub-statements (the effects of biological and behavioral factors on the reliability of a system), this chapter consists of the following sections:

1. State of the art review on the effects of **biological factors** (related to S1 and S2)

1.1. Physiological aging: this section includes the state of the art on the impact of physiological aging on biometric systems.

1.1.1. Age-related biological changes in biometric characteristics

1.1.2. Age-related behavioral changes in biometric characteristics

1.2. Template aging: this term refers to the diminishing matching accuracy as a consequence of the time interval between the recognition events. This section includes the state of the art on the template aging effects.

1.2.1. Short-term template aging

1.2.2. Long-term template aging

1.3. Gender and Race: this section presents an investigation into the effects of factors such as ethnicity, gender and even eye color.

1.4. Disease: this section includes a brief overview of biometric recognition under the influence of disease.

2. State of the art review on the effects of **behavioral Factors** (S2 and S3)

2.1. Usability: this section reviews the usability aspects of biometric implementation.

2.2. Habituation: this section includes a brief overview of the effects of behavioral factors on a biometric system's usability and the state of the art on the effects of habituation on a biometric system's performance.

2.2.1. Familiarity

2.2.2. Makeup

2.2.3. Lifestyle

2.3. Emotional state: this section includes the state of the art on the effects of mood variation on the performance evaluation results of the biometric system.

2.4. Cultural Differences: this section discusses the effect of cultural differences on the popularity of a biometric system. The necessary methodology for investigating most of the methods is a questionnaire-based study approach.

3. Conclusions

The environmental factors can affect the quality of a biometric image. The behavioral factors can change the biometric features temporarily. The biological factors may change the biometric features permanently. For instance, the condition of a person's health has lasting effects on the

texture of the iris, while some temporary changes in biometric features can be observed under the influence of alcohol.

2.1. Biological factors

This section includes a state of the art report on the impact of physiological factors such as aging on the performance of biometric systems.
This part highlights the obstacles that factors of aging can impose on the way in which biometric systems are used. The systems should cope with changes in the biometric features of a user, such as vocal changes due to old age.

2.1.1. Physiological aging

Age has a high influence on how the user of a biometrics recognition system can authenticate him/herself appropriately. It has been reported in previous research that facial aging is mainly attributed to bone movement. On the other hand, due to the reduction in the skin's elasticity, the introduction of wrinkles is the main reason for skin-related deformations [25].

2.1.1.1. Age-related biological changes in biometric characteristics

The biological aging of individuals can obviously lead to changes in the biometric features. For instance, at different incidental lighting levels, the human iris's ability to adapt and accommodate will sharply decrease with age, which can potentially lead to issues about the availability of those features in an iris pattern that can be used to recognize a particular individual.

Pupil dilation is the most important factor for variation in an iris. This factor may be classified as an 'aging factor' (as the resulting biometric reference 'ages', independently of the 'aging' source), but there remains the urgent need to answer the question of whether aging also affects the iris texture, as this is the factor directly used by biometric systems. Pupil dilation cannot directly influence the unique features of the iris themselves. There is unconfirmed evidence that the feature of iris biometric exhibits significant differences as a function of the

biological age of individuals. However, it can confidently be claimed that age progression can decrease the capability of the eye for accommodating pupil [26]. The performance of online-signature and offline handwriting recognition systems are highly dependent on the age of the user. This is because the physiological age can affect the face. According to a previously published paper [27], older users displayed a diminished magnitude of pen dynamics (such as velocity). In [28], the relationship between the equal error rate of fingerprint recognition systems and the physical aging of individuals was investigated. Research showed that it was much less likely that image samples of fingerprints donated by users from the older age group would be of high quality, while users from younger age groups were much better able to provide qualified samples according to the previous research done by Merkel et al. [29]. Younger users can provide fingerprint biometric samples of higher quality, meaning that age progression can be considered as a source of error in evaluating the performance of biometric recognition systems, and that increasing in age may have an effect on the quality of the acquired biometric data. Sicker et al. [30] evaluated fingerprint quality across two populations, the elderly and the young, in order to assess age as a potential factor affecting the utility of image quality. The examination was conducted on a population over the age of 62, and a population between the ages of 18 and 25, using two fingerprint recognition devices (capacitance and optical). The results demonstrated statistically significant evidence that age affected the effective image quality of each index finger. In the work by Jain et al. [31], a database of the fingerprints of 309 subjects (maximum 5 years old children) was collected. This database was built up during four sessions over the course of a year. Fingerprints acquired from a child as young as six months old displayed the distinguishing features necessary for recognition. According to the results they achieved, it can be concluded that fingerprint recognition of young children (six months and older) is a viable solution based on the available capture and recognition technology. Liu [32] studied the recognition of infants by collecting footprint samples using a ridge sensor. Liu collected a database of 60 footprints of infants between one and nine months old over three sessions. The author demonstrated the effects of age and time gap on matching performance. According to Madry et al [33], reliable traces can be provided by using speech signals of younger participants, while the same features cannot provide reliable traces in the case of older adults. According to the results of numerical experiments using a database comprising the facial images of several users from different age groups [34], it was reported that the identification of a younger

population can be more difficult than of an older population. However, in another study [35], the authors claimed that young users are recognized as easily as older individuals. Fairhurst and Erbilek [36] split a database of iris images into three different age groups: a – less than twenty-five, b – between twenty-five and sixty, and c – more than sixty years old. Their work showed that age progression can affect the reliability of an iris-recognition system. Pupil dilation responsiveness decreases with age, due to the fact that the physiology of pupil dilation mechanisms is different for users from different age groups. Guest [37] presented an assessment of the age dependency of biometric dynamic signature verification systems. Some of the reported results indicate that there are no significant differences between the age groups with regard to their ability to satisfactorily enroll or be verified [36]. In [31], the authors noted that "it is more challenging to classify individuals who are in the younger or the older age groups than those in-between." In a paper by Faundez-Zanuy et al. [38], it was found that the false acceptance rate increases with the user's age. Older users were found to be more likely to be mistakenly verified as the genuine user. They reported that the performance of handwriting-based biometric systems degrades as age increases. One of the factors that can change the biometric data and challenge the reliability of a biometric system is the passing of time.

Speaker recognition is an interesting area in the field of biometrics and can be defined as verifying human subjects in various scenes from a voice sample or video source. Human beings can recognize and identify voices learned during their lifetime, even after a break of months. Thus, a major approach to speech processing is to recognize voices at a level approaching the capacity of an average human. Voice is one of the most distinguishable biometric cues that can be used for human identification purposes. Voice samples can be easily recorded and stored using the smartphone's microphones, meaning that voice recognition systems have become one of the most popular mobile biometric systems for cell phones. However, the reliability of speaker recognition systems can be affected by various social factors. Today, smartphones are equipped with biometric tech such as voice. The only problem with biometric solutions is their lack of performance. Voice biometrics has become popular, especially in smartphone or telephony applications where voice services are provided. Voice is one of the only methods to use a combination of physiological and behavioral biometrics. The physiological features of human speech stay the same, but the behavioral part of the speech changes over time and as a person ages. The gradual changes that occur in the human voice due to aging create challenges for

speaker verification. The main effects of aging on the voice are as follows: a less efficient respiratory system (speech slows down) [39], less flexible larynx cartilages [40], changes in pitch [41, 42], and shakiness in the voice [43].

2.1.1.2. Age-related behavioral changes in biometric characteristics

Age has a strong psychological basis and is also theorized to play an important role on behavioral changes.

In paper [44], the age dependency of electroencephalographic (EEG) brain signals was investigated. For this purpose, the EEG resting state activity was collected in 40 users. The participants were healthy adults aged between 16 and 85. They reported that the results of a complexity analysis showed that neuronal electrical activity is significantly different for the users from different age groups.

Our reaction times become slower as we age, with this factor being a well-researched phenomenon, the effects of which have been investigated in a number of different conditions.

System timeout is a common problem that older users face when trying to authenticate themselves. Akatsu and Miki [45] used an experimental approach to answer the question of how much more time is needed for older adults to authenticate themselves in comparison with individuals from younger age groups, in order to carry out certain tasks involving an ATM. According to their results, university students were able to transfer money and perform other transactions in half the time needed by the elderly. This means that older users will face the forced withdrawal of a payment more often than younger people due to the timeout imposed by a system. In that study by Akatsu and Miki [45], the authors suggested that, by making the buttons larger and easier to press and by increasing font size, the usability of a system for adults would be enhanced and more convenient. With age, the capacity to understand and remember the spatial relations among objects becomes slower and less precise.

To assess spatial abilities, Ziefle and Bay [46] used an online paper-folding test using mobile phones. In the study, 16 younger (23–28 years) and 16 older adults (46–60 years) had to solve nine common phone tasks twice consecutively in order to measure learnability.

According to their results, the younger group answered between 8 and 19 of the answers correctly (out of 20), reaching a mean performance of 13.4 (SD = 3.5). On the other hand, the

older adults performed worse that of the younger adults, with 5 to 18 correct answers and a mean of 9.6, (SD = 3.6). Ziefle and Bay [46] showed that if we ask older individuals to have some training before performing a task, this can reduce the failure rate significantly. They reported an average of six per cent performance enhancement when taking a test for a second time, in comparison with the first run without any pre-training.

A paper written by Obrist et al. [47] presented the findings of a usability evaluation study in combination with eye-tracking conducted for an information-oriented interactive television application. The investigation looked at the usability of an interactive television menu and navigation through it for two age groups (elderly users and users aged between 20 and 30). They reported that elderly participants took longer to solve the tasks than younger participants, indicating that higher contrast and font sizes could support elderly people in such situations.

In a study using a small-screen diabetes assistant, Calero-Valdez et al. [48] found that having a simplified spatial representation of the device's menu improved user performance.

As the use of computers and mobile smartphone devices is seeing a constant increase, there is an urgent need to make mobile biometrics more usable for elderly people. By the next decade, the ownership of tablets, personal computers and mobile phones will be more ubiquitous than today, and advances in technology may have left even more people struggling to keep up. Technology is complicated and elderly users are likely to face several problems while working with devices without referring to instructions and user manuals. Unfortunately, such manuals tend to contain multiple pages of sophisticated text in a small font. Additionally, written instructions are useless for the elderly who cannot read.

In a study by Theofanos et al. [49], the interaction of user and operator is investigated. The paper reports that, with the help of operator assistance, only two per cent of older participants were unable to successfully provide fingerprint samples. Interaction with an operator influences the ease of collecting ten fingerprints from users. However, only slightly more than half of the users (56 per cent) could follow various forms of instruction from posters, with 44 per cent of users requiring help from operators.

Within the realm of the cultural and the social, Sasse and Krol [50] identified three major factors that influence elderly people's interaction with biometric systems. These are: (i) tension between generations, (ii) habits and experience and (iii) beliefs and convictions.

The elderly, who are accustomed to more traditional means of communication, for example using pencil and paper, have difficulty keeping up with the rapid pace of developing technology in the modern era [51]. Although usability standards are continually being developed and adapted to the changes in both software and computing hardware, the increasing diversity of elderly populations, both culturally and educationally, requires the development of a specific set of criteria [51].

Zouaoui et al. [52] investigated the contribution of the co-training approach as an enhancement procedure for age range prediction from handwriting analysis, using samples extracted from the IAM dataset.

In thirty years' time, around one-third of the European population will be aged 65 or over. In the United States, the population of the elderly is increasing significantly all the time, with nearly 10,000 adults reaching the age of 65 every day. Aging-related research looks not only at the biological source of this phenomenon, but also the subjects' capabilities, habituation and trust in the technology gained while using it. There are a large number of articles concerned with the effects of physiological aging on the performance of biometric recognition systems. However, in all cases where users are from different age groups, this factor can play a role. The effect of physiological aging can sometimes be neglected in experiments, but we have taken the age factor into consideration.

2.1.2. Template aging

Biometric traits such as voice, signature, and keystroke (i.e. behavioral biometric traits) are liable to change and vary over time, whereas the potential for change in a person's face, iris and fingerprint takes place much more slowly and less perceptibly. One of the factors that has an impact on the performance of a system is human aging. The biometric data changes with the passage of time and this variation leads to a reduction in the performance of the biometric recognition system.

Differences between the biometric sample used as reference in a system and the person's biometric trait when they next use that system to verify their identity are a well-known phenomenon called "template aging". It has been demonstrated that this is likely to affect the verification process by changing the matching score between new data and the biometric

reference template taken from the user some weeks, months or in some cases years ago [53]. In other words, the reduced similarity between different samples from the same user is a direct function of time. Genuine match scores tend to significantly decrease when the time interval between two samples in comparison increases. However, in most cases, despite decreasing genuine scores over time, the average subject can still be correctly verified. Template aging is defined as an increase in error rates caused by time-related changes in the biometric data. This is an important research problem that requires longitudinal study (a research design that involves repeated observations of the same people over short or long periods of time. This type of study can take place over a period of weeks, months or years) [54].

2.1.2.1. Short term template aging

Komogortsev et al. [55] presented a template-aging study of eye movement biometrics, considering three distinct biometric techniques on multiple stimuli and eye-tracking systems. Short to midterm aging effects were examined over two weeks and seven months.

They reported that aging effects are evident as early as two weeks after the initial template collection, with an average 28% (±19%) increase in equal error rates and a 34% (±12%) reduction in rank-1 identification rates. An average 18% increase in EER and 44% reduction in rank-1 identification rates were observed after seven months.

Czajka et al. [56] investigated the diurnal change of the iris and concluded that daily fluctuations have an impact on the density distribution of genuine and impostor similarity scores between iris images. For the purposes of their study, a subset of 18 people was selected, from Canadian border control, where the individuals had been encountered at least four or more times and at time spans of at least two hours during a day. They concluded that the changes during the day, in both pupil dilation and eyelid opening, are statistically significant. Hence the diurnal change of the iris can affect the reliability of the system.

To investigate the potential impact of aging processes on various samples of written content within a biometric handwriting system in terms of authentication performance, scheidat et al. [57] presented the results of an experimental evaluation on the effect of changes in handwriting biometrics by acquiring data from writers during three sessions one month apart. The equal error rate obtained two months after the reference date (EER = 0.162) was four times higher than that

calculated based on the benchmark and the verification data from the first session (EER = 0.041) [57].

In [58], an analysis of template-aging effects on the performance of a speaker recognition system was presented. The collected database contains voice samples from 22 fluent English speakers. The data were captured from the participants during three different sessions spanning 45 days. The results indicate that short-term vocal template aging exists. The observed differences in the distribution of similarity scores between the voice samples were statistically significant.

The purpose of a study by Sanoma et al. [59] was to verify the conservative ECG of humans in their activities, to determine whether it is suitable for use in biometric devices. The experiment involved six participants aged between 21 and 23 years old. The robustness of the ECG was tested under various situations, including variations in health condition, emotional state and heart rate. The results indicated that the ECG is insufficiently stable and seems to vary with daily activity and emotional state.

2.1.2.2. Long-term template aging

A paper written by Manjani et al. [60] studied template aging in three-dimensional facial biometrics. The database used for the study contained two-and three-dimensional facial images taken from 16 participants. The authors investigated the effects of short-term and long-term aging on the performance of facial recognition systems. They concluded that the passage of time between enrollment and use would lead to genuine scores being lower. According to the reported results, the differences between the distributions of intraclass and interclass comparison scores are statistically significant.

Maiorana and Campisi [61] reported and discussed the results achieved from several experimental examinations conducted using a database of EEG signals provided by 45 individuals. The users' EEG signals were collected during at least five different sessions spanning a minimum period of three years, using various protocols of elicitation.

Kelly and Hansen [62] investigated the effect of short-term (between two months and three years) and long-term (up to 30 years) aging on the reliability of speaker recognition systems. They reported that relative reductions in the cost of the log-likelihood ratio of 1-4% and 10-43%

are obtained at short- and long-term intervals respectively. According to the results presented in a paper written by Galbally et al. [63], the effects of age and aging have an influence on the matching accuracy of a fingerprint recognition system. Limited research has been done in the area of aging biometrics in children. It is generally agreed that voice F0 in normal children decreases with age [64]. However, rigorous studies analyzing the variability of speaker recognition in children over time are lacking. Johnson et al. [65] presented an iris aging analysis based on comparison results obtained for a commercial iris matcher – "Verieye". They indicated that the iris biometric characteristic is stable over time, at least as early as the age of four.

Biometric-template-aging studies have been carried out with respect to methods involving the iris [66], the face [67], the voice [68], signatures [69], and fingerprints [70].

In a paper written by Erbilek and Fairhurst [71], a template-aging study with respect to the iris method was presented. The study analyzes the effects of time between enrollment (when the biometric template is provided and stored) and use (when the user wants to verify their identity) on the performance evaluation results. The results showed a significant difference in performance over time.

It was traditionally accepted that biometric template aging does not occur for fingerprint biometrics. It was claimed that fingerprints have enough long-term stability for reliable automatic person recognition.

The results of a statistical analysis [72] of AFISs' similarity matching scores of fingerprints recorded over a span of 21 years showed that an average nine per cent of the variability of the AFISs matching scores can be explained by template aging, and it was found that fingerprint template aging has a statistically significant negative impact on the performance of the AFISs, even on a relative young white-male population aged from 14 to 53 years.

In a study by Yoon and Jain in 2015 [70], fingerprint-matching scores were analyzed. Longitudinal fingerprint records were sampled from a preexisted database. The database contained five 10-print records for about 15,597 individuals over a time span of at least five years. The study claimed that, while the matching accuracy of the fingerprint recognition system tends to remain stable, genuine comparison scores will decrease in a statistically significant way as the time between enrolling and use increases up to twelve years, which was the longest time lapse between sessions in the published dataset.

Kirchgasser and Uhl [73] performed experiments on datasets including a time span of four years. Non-minutiae fingerprint recognition methods were utilized to verify earlier findings related to fingerprint template aging. The analysis found that there are very similar effects in terms of fingerprint template aging detectable for all the recognition methods considered. The same authors [74] investigated the impact of "ghost" fingerprint and minutiae information fingerprint datasets separated by a four-year time span. The presumption was that a high amount of ghost fingerprints within the data might be responsible for recently reported template-aging effects. However, with respect to detected increased error rates, the analysis exhibited very similar effects for all the considered methods, regardless of whether ghost fingerprint information was removed or not. Thus, they concluded that ghost fingerprints were not responsible for the observed effects.

Best-Rowden and Jain [75] conducted a numerical experiment using a database of up to 150 thousand mug shots of more than 18 thousand criminal offenders. The longitudinal database contains at least five facial images taken from each individual over a minimum time span of five years. A longitudinal analysis of the scores showed that, despite the genuine similarity scores decreasing over time, facial recognition systems can still verify true users at a false acceptance rate of 0.01 per cent across up to 16 years of elapsed time, which was the longest span in the mentioned database.

Since biometric features alter over time, within-person variation and template aging lead to significant system performance degradation [76]. Baker et al. [77] used an iris database with a four-year time lapse and a two-year time lapse to investigate the effect of template aging on the iris recognition system. Both sets of experimental results indicated that there is a significant increase in the average Hamming distance between images after one year.

Czajka [78] studied the iris aging effect on interclass matching scores using four different matchers. The database contained samples collected during three sessions: the first session in 2003, the second session in 2007 and the last session in 2013. Using this database, it was possible for the author to investigate aging in iris biometrics for both in the mid-term (less than two years) and long-term time span (from five to nine years). Czajka reported around a 14 % degradation in the average genuine similarity scores because of template aging. It was also demonstrated that there is a meaningful difference with respect to the time between the acquired iris samples.

Browning and Orleans [79] investigated the effects of template aging on the reliability of an iris recognition system. The numerical experiments were conducted using a subset of 14,227 individuals who had used the system three or more times over a time span of at least three years. The set included 892 individuals who had used the system three or more times over a time span of at least five years.

The paper reported that no evidence had been found to prove that there is a clear and continuous trend of degradation in similarity scores as the time intervals increased. An investigation into the relationship between age and the Failure to Enroll (FTE) Rate was also presented. The authors reported that the overall template generation failure rate was 1.5% (1,783 instances). The FTE rate was lower for older individuals than for users between twenty and thirty years old. With respect to facial biometrics and template aging, the Face Recognition Vendor Test 2002 [80] (which is an independently administered technology evaluation of mature face recognition systems that characterizes performance as a function of elapsed time between enrolled and new images of a person) performed experiments on a dataset of 121,589 images from 37,437 individuals, with at least three images obtained from each person. The results show that performance decreases linearly with the increasing lapse of time.

Dynamic Signature Verification (DSV) is unique among other biometric recognition technologies as there is no clearly defined method of creating a forgery. Galbally et al. [81] investigated the effect of aging on the dynamic signature system's performance using a consistent and reproducible database over time.

For conducting a long-term and/or short-term longitudinal study (with appropriate data collection) related to biometric aging, using interfaces for various methods operating on mobile platforms, it is necessary to provide an experimental evaluation of the influence of changes in recognition ability by acquiring data from the same speaker. It is very important that the biometric samples of the users are captured by using different devices, which means that the effects of "sensor aging" should be considered as negligible.

Sensor aging is an influential issue that is very important to address, as it can challenge the reliability of a system in a meaningful way. Due to these circumstances, there are only a few suitable data sets available when presenting aging analysis [82]. In [83], the evidence for the template-aging effect on a hand biometric recognition system is presented.

In research carried out by Kirchgasser et al. [84], the achieved results do not suggest the existence of a template ageing effect for fingerprints.

The fact that the human voice changes due to the phenomenon of aging is a well-researched problem and it has been proven that vocal aging may be represented through acoustic changes in the voice. In study [85], the authors used an i-vector speaker verification framework to report on the impact of long-term aging on state-of-the-art speaker verification. In the work, the Trinity College Dublin Speaker Ageing (TCDSA) database was used, which contains the longitudinal speech recordings of 34 adults across a time span of 25-58 years per speaker (Release I contains samples from 26 users). The authors reported that the performance of the i-vector system, in terms of both discrimination and calibration, degrades progressively as the absolute age difference between training and testing samples increases.

For male adult users, the EER changed from 4.61 % for time lapses of up to one year, to 32.74% for samples from the long-term aging sub-database with a time span between 51 and 60 years. According to their results, the aging effect varies for different age groups and is also gender dependent.

In work done by Fernandez et al. [86], handwriting samples from 51 subjects were collected. After five years, the experiment was repeated for 25 of those subjects. They reported that the handwriting of users from the 39–65 age group remained constant, while for users from the 65+ age group, the effects of template aging on the performance of the system was significant.

Kelly and Hansen [87] stated that, as the time interval between train and test sessions increases, so too does the EER. Over a three-year interval, there is an approximate relative increase in EER of 60% for females and 80% for males. Given that the conditions in the experiment were controlled, this outcome suggests that vocal aging over three years does degrade speaker recognition performance.

In [75], Best-Rowden reported that ethnicity and gender can significantly change the performance evaluation results, from a statistical point of view.

Template aging may occur when the biometric characteristic undergoes significant changes over time. In line with the previously mentioned studies, it can be concluded that for each face, fingerprint, voice, signature and iris biometric characteristics, aging-based research has been performed. In recent years, speaker recognition across aging has become a very popular and challenging task. Many researchers have contributed to this area, but there is a significant gap to

fill in. It is obvious that, before talking about the effects of aging, the most important precondition is the need for a data set including a time span. This requirement can be met, but there is an additional problem. The amount of non-aging-based variability within the data set should be as low as possible. The voice we go to sleep with is significantly different from the one we wake up with. This study investigates the effects of very short-term template aging on the performance of speaker recognition systems.

2.1.3. Gender and Ethnicity

Classifying individuals based on ethnicity using their facial images can be very useful and can provide an ability to make a pre-estimation about the criminal background of the user, and potentially also lead towards facial identification [88].

Amayesh et al. [88] used a classification methodology in their work, in order to study the relationship between the gender of participants and the specific traits in the shape of their hands. For this computer vision approach, they calculated the function represented by the distance from a chosen part to two different Eigenspaces, for both male and female users in order to classify the users to two different genders. They reported 98% accuracy for their methodology of classification, though they used a very small data set containing 20 male and 20 female participants.

In [89], a model for the classification of users based on ethnicity from frontal facial pictures of individuals was proposed. The authors of that paper reported 95.1% accuracy for their proposed model of classifying race from facial images. The study also used the same model to classify users based on their ethnicity, using facial images in an unconstrained environment. The reported results showed a 6.2% drop in the accuracy of the model.

We can use iris features for user recognition, but very little work has been focused on using iris texture patterns to determine ethnicity or gender. The prediction and classification of ethnicity based on iris texture patterns, using image processing, artificial intelligence and computer vision techniques, is possible. Most of the work carried out by researchers, through delving and comparing the various techniques, algorithms and results achieved over the last decade, was consolidated in a review article [90].

Howard and Etter [91] carried out research to investigate whether factors such as gender, race and even eye color can play a role in the performance of an iris recognition system for different participants across the population. According to their results, African American participants with dark eyes (brown or black eyes) are the most likely groups to be mistakenly rejected as false users by iris verification systems. According to their results, the observed degradation in the performance of an iris recognition system may be also due to the eye color of the users.

Daughman et al. [92] used IrisCode in order to achieve more than 100 billion impostor scores between the iris samples of hundreds of thousands of participants from different nationalities. According to their results, the iris recognition system performs well. Biometric recognition systems are very popular nowadays. However, the limits of various kinds of biometric techs must be highlighted. Based on previous research works [93], there is a good deal of biological evidence to prove that the general effects of physiological and behavioral factors on the performance of biometric systems can also be considered gender dependent. So for the investigation of our chosen special cases, we will check whether the reliability of a biometric recognition system tends to be different for female and male users under the influence of biological/behavioral factors.

2.1.4. Disease

Disease is another social problem that can affect the reliability of a biometric system. Trokielewicz et al. [94] studied the effect of eye diseases on the reliability of iris recognition system using images taken from illness-affected eyes of 92 participants (184 eyes). The illness-affected irides were then categorized into four different categories based on the shape of the disorders. The collected database contains up to three thousand near infra-red images and visible wavelength pictures that were captured during routine ophthalmological practices. According to the results, eyes from the obstruction group are the hardest to recognize.

Parkinson's disease can affect the loudness of the voice and the clarity and speed of speech. For this reason, we can consider it one of the social issues that can cause a reliability challenge for a speaker recognition system. Parkinson's disease is also a social issue that can make signature verification tasks harder, posing a challenge to the reliability of online signature recognition systems due to the tremors associated with the disease [96-98].

Seyyedin et al [99] conducted several numerical experiments to investigate the reliability of an iris recognition system following iatrogenic pupil dilation. They collected a database during two sessions spanning 2-24 hours after phacoemulsification and intraocular lens implantation, as well as before and after iatrogenic pupil dilation. They concluded that standard cataract surgery does not seem to be a limiting factor for iris recognition in the large majority of cases, though it can reduce the mean value of genuine matching scores distribution, and as a result it can diminish the reliability of a system.

Pupil size can affect the accuracy of recognition system for various degrees of dilation. According to the results reported in [100-102], when the degrees of dilation of the pupil in the probe and gallery iris images have same values, better results can be achieved. The performance of a system can be defined as its matching accuracy. It can be concluded that, as pupil dilation can be induced by drugs, the matching accuracy of the iris recognition system may decrease as a result of drugs. [103].

The skin color or the papillary lines of fingertip can be changed tragically by skin diseases. As the main and most popular fingerprint scanners are not dependent on the color of the skin on the finger, the problem of changing color can be resolved easily during the data collection phase. However, Drahansky et al. [104], when conducting a numerical experiment, presented a paper entitled, "Influence of Skin Diseases on Fingerprint Recognition", for which they had collected a new database of fingerprints taken from users with skin disease using various fingerprint scanners. They concluded that using color-based scanners may lead to the system experiencing a greater error rate. The widespread and well-known disease diabetes quite frequently leads to eye conditions such as retinopathy (retinal damage), cataracts (clouding of the lens) and glaucoma (optic nerve damage), hence we can hypothesize that the impact of this medical condition on the accuracy of iris recognition is significant.

2.2. Behavioral Factors

This section includes a brief overview of the effects that behavioral factors can have on a biometric system's usability, and sets out the state of the art on the impact of behavioral factors on the performance of biometric systems.

2.2.1. Usability

Human-Biometric Sensor Interaction (HBSI) focuses on the interaction between the user who wants to authenticate herself/himself, the sensors that can be utilized to collect data and the biometric recognition system. A usability study of a biometric system aims to find the answers to three questions: a- Is the system fast to learn?; b- How can qualified samples be provided using the system?; and c- Can the system be considered user-friendly from the participants' perspective?

According to ISO 13407:1999, the definition of usability 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." [105].

According to ISO 13407:1999, a usability study of a system can be undertaken by investigating the effectiveness, efficiency and user acceptance of the system.

In the current project, the usability of the system is not of the main interest and, as human-computer interaction can influence the data collection procedure but cannot change the performance results directly, we will be more focused on investigating the effects of social problems. The occurrence of social issues could lead to nonlinear changes in the way a system works. There are several user-related parameters, and these factors have a direct effect on the evaluation of the performance results of the recognition system. Under the influence of these parameters, the matching accuracy of the system might be affected directly.

Back to usability, the International Organization for Standardization classifies usability into three components [105]: a- effectiveness (can users successfully provide a high-quality sample?): the Failure to Acquire rate represents this parameter; b- efficiency (are users able to quickly accomplish goals): by measuring the user action time, or the time the user spends on providing the sample, we can report how efficient the system is. The enrollment time or the time required for a system to train itself to recognize the users represents the efficiency of a system; c- satisfaction (are users intimidated by using the biometric system): this can be assessed using feedback questionnaires.

There are two more components that are also important: learnability (are users able to use the system to some defined level of competence after instruction or training?): the Failure to Enroll rate represents this parameter, and memorability (how do experienced users differ from infrequent/novice users?): which examines the repeatability of an experiment in different sessions.

In order to present a detailed analysis of the experiment, the framework of Human Biometric Sensor Interaction [106], as proposed by researchers at Purdue University, must be applied.

However, the performance evaluation of a biometric system concentrates on the system level. In other words, the measurements of factors that can represent the error rates of a biometric system are more popular among researchers. The raw basic metrics for error rate, such as the True Accept Rate (TAR), which refers to the percentage of times a system correctly recognized genuine as genuine; the False Accept Rate (FAR), which refers to the percentage of times a system incorrectly recognized impostors as genuine; the False Reject Rate (FRR), which refers to the percentage of times a system incorrectly recognized genuine as impostors; and the True Reject Rate (TRR), which refers to the percentage of times a system correctly recognized impostors as impostors. By contrast, for an investigation using Human Biometric System Interaction, the metrics of Failure to Enrol (FTE) and Failure to Acquire (FTA) will be used [107].

In a study by Hurtado et al. [108], the usability evaluation of PIDaaS for a speaker recognition system using the HBSI model was presented. In their study, the interaction of each user was captured on four cameras. Although the users were from different age groups and nationalities, there was no extended investigation regarding the effects of the factors mentioned here.

2.2.2. Habituation

Lifestyle habits, including addiction, smoking and the abuse of pharmaceutical drugs can cause biometrics to fail. In some cases, the differences are not significant enough to lead to failure, but the influential parameter can challenge the reliability of the biometric system in a meaningful way. The habits of a daily routine can also be considered as important factors that can play key roles.

2.2.2.1. Familiarity

Smejkal et al. [109] reported that using the first attempted signature as “practice”, not to be included in the database, will reduce the variability of a participant’s signature. They also demonstrated that shorter signatures (abbreviated signature), like Japanese, show very high variability of conformity and non-conformity between individual signatures. It was also

concluded that the quality of recognizing a signature rises with the length of the information written down.

In paper [110], Gunetti et al. tried to answer the question of whether the rhythm of keystrokes can be a reliable way of identifying a user while typing sentences and words in a different language from enrollment phase. To study this problem, they collected data from users from different nationalities, speaking different languages but also fluent in English. The users were asked to type free text for authentication purposes. The results proved that keystrokes can indeed be used successfully to authenticate an author, even if the user was enrolled using the text with different words and letters. However, due to the different photonic structure, pronunciation and accents, language can play a key role for text-dependent investigations of voice recognition. In the only study to date looking at the role of language in voice processing by children, young English-speaking children aged 7-9 are significantly more accurate at discriminating voices when listening to English speech than German [111].

The typing rhythm of a user would be totally different for familiar and new words that can be categorized as unfamiliar words. In an investigation conducted by Seyd et al. [112], the performance of such a biometric identification system was tested for passwords using familiar words from a dictionary or from the user's memory, and for passwords using nonsense, random combinations of letters. For trained words from the memory of the user vs. unfamiliar words used as a password, the results showed that as participants are step by step trained in using the practiced word, the stability of typing rhythm in sessions will definitely be increased, and by becoming habituated to the chosen word, users will repeat the same rhythm for typing the password after several times of practicing. In their conclusion, they also declared that, in comparison with dictionary-based words, complex and random passwords with no meaning will cause a change in the reliability of the biometric system, and can increase the equal error rate of the system as a result.

2.2.2.2. Make up

Using makeup for beautification has long been a way of life for many people and only grows in popularity. In order to answer the question of whether face makeup can affect the matching accuracy of facial recognition systems or not, Dancheva et al. [113], collected and gathered two

different databases from the face pictures of individuals, before applying makeup and while made up. The presented results from the experiment reveal that makeup can indeed influence the performance evaluation results of a facial recognition system significantly, but is unlikely to lead to a failure in the system.

Kohl et al [114] presented an in-depth analysis of the effect of wearing contact lenses on iris recognition performance. The IIIT-D Contact Lens Iris database was collected, with over 6,500 images from 101 subjects. For each subject, images were captured without wearing lenses, while wearing transparent lenses, and while wearing color cosmetic lenses, using two different iris sensors. The results computed using VeriEye suggest that a color cosmetic lens significantly increases the false rejection at a fixed false acceptance rate. The authors concluded that building a sophisticated lens detection algorithm may be a way of improving iris recognition.

Osman et al. [115] contributed a paper present an investigation into the effect of facial plastic surgery on the reliability of a facial recognition biometric system. The database used was collected with the help of plastic surgeons from all over the world. The authors reported a verification rate in excess of 91%.

2.2.2.3. Lifestyle

In a paper written by Bours and Evensen [116], the reliability of a biometric system looking at gait was investigated using a database containing data from 40 users. The users in the experiment were asked to wear two different shoes. The results showed that the accuracy of the system varies, both for the same shoe and different shoe setting.

Smoking is another social issue that can lead to deviations in a person's voice and vocal changes, as reported in [117], which used a subset of data from the NIST telephone recording database.

In a paper by Satori et al. [118], an HMM automatic speech recognition system was created to detect speaker who are smokers. The results obtained from the experiment imply that biometric systems can be adapted to identify smokers and confirm whether a speaker is a smoker, even when the observed recognition rate is below 50%.

The influence of alcohol on a person is very common and can change the evaluation of how a biometric system performs and can be considered as an important parameter, especially for the

identification of drunk drivers. Alcohol can cause the iris to dilate and constrict at a very much slower speed, and pupil dilation and constriction can change the iris texture pattern, meaning there is a link between pupil dilation and variation in the region of interest.

In a study conducted by Arora et al. [119], the "IIITD Iris Under Alcohol Influence" database, containing iris images from before and after the consumption of alcohol, was used to investigate the influence of alcohol on the reliability of an iris recognition system. The authors hypothesized that as pupil dilation can change the segmented and normalized iris pattern and can be the main reason for shape deformation in the iris texture, hence alcohol consumption can affect the accuracy of the recognition system significantly. The presented results from the performed experiment using the "IIITD Iris Under Alcohol Influence" database proved that, in comparison with the pre-alcohol consumption iris photos, the average value of similarity scores for the post-alcohol consumption were lower. The genuine matching score experienced a 20 % degradation for post-alcohol consumption sub data. This means that the eye under the influence of alcohol is significantly different than normal ones.

In [120], the authors provided a methodology to measure the difference in a handwritten signature before alcohol consumption and after, along with a change in the stability of samples using a database gathered from 30 users. They also used a detector to measure the alcohol rate in a participant's breath, to be sure of the level of intoxication.

Although there are already a large number of papers looking at the effects of makeup on facial recognition, and there are many papers considering variations in facial recognition accuracy due to expression, it seems that there has not been any previous investigation into the effects of a combination of both expression and makeup, and there is no published paper to answer the question whether a combination of full makeup and facial expression makes a statistically significant change in the biometric results.

2.2.3. Emotional State

While they are awake, a person's mood may vary considerably. The human face can be described as a window into the emotional experiences of a person and the expressions of facial emotion are reactions that include many interconnecting elements of movements by facial muscles. The expressions of the face can give us an ability to discern a person's emotional frame

of mind. Facial expressions are visible changes in response to a person's internal emotional states, intentions, or social communications. Expressions of the face are commonly categorized into seven classes: anger, disgust, fear, happiness, sadness and surprise, along with the neutral state. This visible change in a person's face means that the reliability of a biometric recognition system is influenced by the mood of the users.

Facial expressions have long been studied by clinical and social psychologists, medical practitioners, actors, and artists. However, in the last quarter of the 20th century, with advances in the fields of robotics, computer graphics and computer vision, also animators and computer scientists started showing an interest in the study of facial expressions. This has continued and is especially true for the last two decades, during which the study of facial expression recognition has been employed in various Human Computer Interface (HCI) applications.

Wan and Aggarwal [121] investigated the accuracy of facial expression recognition using the MFP database. Their work improved the overall accuracy by increasing the training features and the probability distance inside each class. The authors of that work concluded that geometric-based methods have high complexity, as they require accurate and reliable algorithms in order to detect the various subtle changes in facial expression.

Ivanovsky, et al. [122] used a convolutional neural network-based method in order to detect a smile and recognize facial expression. They reported 84.9% and 94.73% accuracy of their classifiers.

Li et al. [123] presented an image processing pipeline to recognize facial expression by first using a face template to identify a set of feature points on faces, and then applying a neural network to classify facial expression into one of six categories: happy, surprise, sad, distracted, focused, and plain. They also tested the pipeline on a standard database and found that it can achieve a satisfactory performance.

In [124], experiments were conducted on the Yale Database [125], the JAFFE (Japanese Female Face Expression) database [126] and on the CMU AMP Expression database [127]. The proposed method gave a recognition rate of 94% in the Yale database, 99.52% in the JAFFE database and 100% in the CMU AMP database.

The studying of the effect of facial mood on the reliability of facial recognition systems from the point of view of biometrics, science is very important when investigating the performance of a system.

In our daily lives, we may face a range of different emotional conditions, (when we are in courts, hospitals, in the case of accidents, etc.) and the biometric system must be able to recognize and verify a user perfectly, despite them being under the influence of a mood variation.

As facial recognition is influenced by mood, the matching accuracy of a system will be degraded and so a major problem is that the matching score between samples in a neutral condition and those facial images taken under the influence of emotions is lower than comparison scores achieved by comparing non-expressive face images with themselves.

In the context of using the voice as a biometric, it is important to assess whether these reported variations in a population affect the performance of a standard system. However, the key challenge when verifying a speaker is that the performance of speaker recognition systems will be significantly degraded when talking under the influence of emotional environments. In [128], an emotional speaker recognition system, based on different feature extraction methods, focusing on the diversities between simulated and natural emotional speech databases (BERLIN and IEMOCAP), was assessed. The reported results showed that the context of emotional speech impact significantly speaker recognition rates. Intra-speaker variations, generated by factors such as the emotional and physical state of the speaker, also affect the speaker recognition performance. Only a few studies have considered the effect of non-speech sounds, such as a whistle, scream etc., on speaker recognition [129]. Revathy et al. [130] discussed the effectiveness on the use of a Hidden Markov Model tool kit (HTK) for recognizing speech, speaker and emotion, based on emotional speeches using Mel frequency cepstral coefficients (MFCC) as a feature. They claimed that the accuracy of the system can be improved if an additional preprocessing technique for noise reduction is used prior to conventional preprocessing. Parthasarathy et al. [131] evaluated a speaker verification system trained with the i-vector framework and with a probabilistic linear discriminant analysis (PLDA) back-end. In a paper written by Krothapalli et al. [132], a neural network based feature transformation framework for developing an emotion independent speaker identification system was proposed. An investigation of the effect of emotional state upon text-independent speaker identification was presented in [133].

Chang et al., [134] reported that keystroke rhythm and keystroke duration will change under the influence of muscle fatigue and other fatigue-related factors by measuring the duration of muscle twitch in the finger flexor muscles.

Komandur et al., [135] established a study for the same purpose, but they used the mouse click rhythm trait for their experiment. They showed that the duration of the mouse click is a direct function of finger flexor muscle twitch durations.

To think clearly and have a high mental availability, it is necessary to be fresh. Tiredness may be considered as the main reason for a lack of concentration, and employees in the workplace can lose their focus when performing tasks because of fatigue. The loss of concentration, lack of accuracy and reduction in mental availability and awareness can cause accidents and put people in trouble. Hence, fatigue, just like stress and emotional moods, can affect the reliability of a biometric system by changing the biometric characteristics of individuals.

In a study by Al-libawy et al. [136] at the University of Liverpool, a methodology for fatigue detection on a mobile device was proposed, by using a typing task on a screen. The app on a cellphone stores the time features of each keystroke and can be used for non-intrusive human fatigue detection in order to classify users. Several parameters, such as the rhythm of text typing, as well as the pressure and duration of pushing a key, have been measured for alertness detection. A support vector machine classifier is also proposed to show fatigue status. The achieved reported results showed an 88.8% accuracy of the classification tool.

In paper [137], Haque et al. investigated the relevance of expressions of pain from facial video to be used as a biometric or soft-biometric trait. They employed a biometric person recognition scenario by using features obtained from the pain expression pattern found in the temporal axis of subjects' videos. They demonstrated that pain can reduce a system’s reliability.

Previous studies concerned with gender differences in psychology have reported that women are more emotional in comparison with men. However, in terms of the science of biometrics, this fact has not yet been investigated. However, in this study we hope to answer the following questions:

a- Are, as psychologists suggest, facial expressions more intense in females?

b- For which mood classes does the application of lipstick in combination with facial expressions change the results in a statistically significant way?

c- Which facial mood images/ expressive voice samples are most dissimilar to non-expressive pictures/ voice samples of the same person?

2.2.4. Cultural Differences

Acceptance of technology may vary between countries and cultures. There are several works on the influence that a person's cultural background might have on the acceptance of biometrics technology. This has been measured through the use of questionnaire surveys. Krup et al. [138] presented a comprehensive questionnaire regarding the social acceptance of biometric technologies in Germany. Furnell et al. [139] proposed a questionnaire on the acceptance of biometric technologies that was completed by 175 respondents living in the UK. In [140], a survey was proposed in order to measure perception based on various uses of biometric technology. It was found that the 141 respondents in the survey were most willing to employ biometric identifiers into the United States passport system (almost half of the respondents). Mok and Kumar [141] investigated privacy-related concerns in the deployment of biometrics and data protection technologies in China. Rashed et al. [142] presented user acceptance of biometrics in the Arab culture (more precisely, in Yemen). In [143, 144], the authors explored the acceptance of biometrics as an authentication tool in e-commerce in Saudi culture. An extended Technology Acceptance Model (TAM) and Hofstede's cultural dimension theory were combined in a quantitative survey to compare familiarity, knowledge and acceptance of biometrics between Brazil and Finland [145]. Brazilian respondents perceived biometrics as more useful when they had a higher need for security, and they found involving their body in the authentication process to be less invasive than their Finnish counterparts. Unlike Brazilians, the Finnish respondents did not favor biometrics as a way of increasing their personal security, and did not see ease of use as a strong enough factor to directly affect usage intentions. We believe that some of the recognition problems caused by influential factors can be solved by system owners. For instance, by asking the users of a biometrics system to re-enroll themselves from time by time, the problem of template aging can be easily resolved. However, users should be willing to re-enroll themselves, or to provide several templates. To judge whether the obstacles imposed by cultural differences can be overcome by the users themselves, it is necessary to have better understanding of biometric technology acceptance in various societies. This problem can be easily dealt with by

using a questionnaire-based study. A comprehensive questionnaire regarding technology and the acceptance of biometric technologies will be presented for future studies.

2.3. Conclusions

Biometric techniques, such as face, voice and eye recognition, have gained immense popularity (and attracted much controversy) in recent years. This thesis reports on new findings regarding the influences of certain social factors on biometric recognition. We investigate how influential parameters have an effect on the use of verification/identification systems, and we attempt to analyze the variations in the performance of biometric systems under various conditions (by reporting the results of several special cases). The field of "system reliability testing" relates to testing the ability of a system to perform well over a particular period of time under given environmental conditions. The same definition can be adapted for testing the reliability of biometric systems. The term "biometric system reliability" refers to the ability of a biometric system to perform its verification/identification tasks regardless of the influences of a number of biological/behavioral/environmental parameters.

In general, there are different types of factors that can influence the biometric data. The biological factors (related to S1 and S2) usually have permanent effects on the biometric trait, as they may change the physiological characters of biometrics (for instance, retinal detachment can permanently change the iris trait, and thereby render it unusable against a template collected previously). Behavioral factors usually have temporary effects on the biometric traits (for instance, facial expressions are visually detectable changes in facial appearance, but can change and change back rather quickly). Cultural restrictions cover certain well-known social issues that can be categorized in the same class, as they can change the way a user behaves. Finally, there are environmental factors, which can be mitigated by controlling the condition under which a sample is acquired.

With respect to the type(s) of factor(s) involved, the differentiation between behavioral and physiological biometrics can be of significant importance for many reasons.

3. Iris Recognition under the Influence of Diabetes

This Chapter was previously published in the Journal of Biomedical Engineering/Biomedizinische Technik in 2019, Authors: [M Azimi, SA Rasoulinejad, A Pacut], Title: "Iris recognition under the influence of diabetes".

3.1. Overview

Chapter 2 reviewed the effects of social issues and the human factor on the reliability of biometric recognition systems. In this chapter, iris recognition under the influence of diabetes will be investigated. As mentioned above, Iris recognition is one of the most reliable methods of identification [146]. The iris is a thin, circular structure in the eye that can control the diameter of the pupil, and therefore the amount of light reaching the retina. According to the results of research carried out during last two decades [147], it has been proven that the pattern of the iris is unique to every individual, and so the detailed structure of the front layer can be used to identify people. Even identical twins will have a completely different iris pattern. It is also worth mentioning that the structure of the iris in a person's left eye will be completely discriminable from the same person's right eye. Among the factors that can challenge the reliability of an iris recognition system and degrade its accuracy are abnormalities developing in the iris pattern due to ocular conditions. A new database containing 1,318 pictures from 343 irides – 546 iris images from 162 healthy eyes (62% female users, 38% male users, 21% < 20 years old, 61% 20–40, 12% 40–60, 6% > 60), and 772 iris images from 181 diabetic eyes with a clearly visible iris pattern (80% female users, 20% male users, 1% < 20 years old, 17.5% 20–40, 46.5 % 40–60, 35% > 60), was collected. All of the diabetes-affected eyes have clearly visible iris patterns without any visible impairments, and only patients with type II diabetes for at least two years were considered for the investigation. Three different open-source iris recognition codes will be used to achieve the performance evaluation results for the iris recognition system under the influence of diabetes.

3.2. Introduction

In this part, we will investigate the biometric system performance evaluation results degradation due to the non-obvious distortions in iris texture caused by diabetes type II.

Diabetes is a risk factor for many well-known diseases and it is a growing epidemic especially among elderly people [148]. This social issue can be the main reason for eye diseases includes diabetic retinopathy [149], diabetic macular edema (DME) [150], cataract, and glaucoma. According to previous researches, strong academic evidence suggests that diabetes can be diagnosed by examining the iris texture (by conducting classification based study). Hence, here, it has been hypothesized that the iris recognition system's accuracy for healthy and diabetic eyes must be different and the reliability of biometrics recognition system can be influenced by the health condition of participants [151-153].

Samant and Agarwal [151], provided an automated tool with machine learning techniques to access the correlation between distortion of iris tissues and diabetes mellitus. They have also proposed a diagnostic tool along with the mainstream diagnosis methods for discrimination of healthy patients and those who are suffered from diabetes. The results show best classification accuracy of 89.63% calculated from RF classifier.

Obviously, identification of irises suffering from ophthalmic diseases would be harder than healthy ones due to the problems which would be occurred for system during segmentation and encoding process. Segmentation and feature extraction of those iris patterns which are occluded due to surrounding tissue pushing the pupillary boundary is harder than segmentation of healthy iris images [154-156].

In order to answer this question: "how iris recognition methods perform in the presence of ophthalmic disorders?," Trokielewicz et al. reported [94] that because of several problems for diseased eyes, iris segmentation phase is the most sensitive part of recognition whole process. Nigam et al. showed that cataract surgery affects the discriminative nature of the iris texture pattern. This finding raises concerns about the reliability of iris-based biometric recognition systems in the context of subjects undergoing cataract surgery [157].

3.3. New Database

For this study, a new database was collected to investigate that if there is any relation between the iris recognition system's accuracy and the health condition of the users or not? Hence, we needed volunteers to acquire datasets for investigation of iris recognition system's accuracy under influence of diabetes. The experiment specifically designed to investigate that if there is

any relation between the iris recognition system's accuracy and the health condition of the users or not. Before the experiment, consent agreements were signed by the participants. The participants were also asked to provide non-biometric data, including their names , gender, age, and the duration (if applies) of their diabetes illness. The personal data are kept separately to guarantee additional security of the personal data. As a result, all of the participants were fully aware of the experiment as we provided full detailed information on the study and it is important to note that the signed consent forms are also obtained from all of the individuals. The experiment protocol has been approved by the Ethics Committee of Warsaw University of Technology. The collected database includes near infra-red (NIR) iris images taken from volunteers. All those participants were from the same ethnic group (Iranian) but different age groups. The data samples have been captured using a commercial iris capture device: "Iri Shield USB MK 2120U " [158], which was connected to a galaxy A5 smartphone for storage of iris samples. Iri Shield USB MK 2120U is a mono iris NIR capture device widely used for capturing iris texture pictures and the captured iris images are compliant with ISO/IEC 19794-6 standard. For each user, the whole acquisition time was less than one minute. All of the taken samples have been provided during one session. The data collection took approximately one month (25 Aug – 15 Sept). As it is shown in Table.1, the new offered database contains 546 pictures from healthy 162 irides (62% female users, 38% male users, 21% < 20 years old, 20 < 61% <40, 40 <12% < 60, 6% more than 60 years old) and 772 iris images from 181 diabetic eyes but with a clearly visible iris pattern (80% female users, 20% male users, 1% < 20 years old, 20 < 17.5% < 40, 40 < 46.5 % < 60, 35% more than 60 years old). For healthy users (who were not suffered from diabetes mellitus) no medical examination was done and for collection of the non-healthy iris samples we have requested the individuals with known diabetes to participate in our experiment.

The experiment specifically designed to investigate that if there is any relation between the iris recognition system's accuracy and the health condition of the users or not. Before the experiment, consent agreements were signed by the participants. The participants were also asked to provide non-biometric data, including their names, gender, age, and the duration (if applies) of their diabetes illness. The personal data are kept separately to guarantee additional security of the personal data. As a result, all of the participants were fully aware of the experiment as we provided full detailed information on the study and it is important to note that

the signed consent forms are also obtained from all of the individuals. The experiment protocol has been approved by the Ethics Committee of the Warsaw University of Technology. Regarding to the process of preparing the diabetic iris database, we hereby, confirm that:

a. The process of taking pictures with the NIR camera has no harm for the subjects and has not damaged the volunteers.

b. Volunteers have informed us for how long they are taking diabetes Pills/Insulin and no tests have been performed to confirm their diabetes states.

c. No drugs have been used for testing.

d. There is no need for medical examination to record and save photos.

In Fig.3-1, four samples from each group have been presented and as it can be illustrated the samples captured using near infra-red camera. The pictures are gray scale and the resolution of pictures is 640*480 pixels. It is worth mentioning that all of the volunteers are from same ethnic group (Iranian), and it is also necessary to note that volunteers are living in Iran and they are having the same dietary culture.

3.4. Methodology Description

Despite of diabetes diagnostic techniques which need just a small part of iris texture's region, for iris recognition purpose the overall structure of the iris must be segmented.

3.4.1. Iris Segmentation (Weighted Adaptive Hough and Ellipsopolar Transform) [159]:

In this manuscript, for iris segmentation Weighted Adaptive Hough and Ellipsopolar Transform (WAHET) methodology has been used. Weighted Adaptive Hough and Ellipsopolar Transform technique is a two-stage iris segmentation technique;

a. Finding center point: the center of multiple approximately concentric rings at iteratively refined resolution can be determined by removing the detected reflection mask, detecting the edge, and finally by applying the weighted adaptive Hough transform.

b. Extracting the region of interest: the center point must be used for this purpose. Firstly initial boundary must be detected and after first iteration, Ellipsolar transform will be applied for Inner and outer boundary detection. After this stage the extracted iris texture will be normalized using Daugman's rubber sheet model.

3.4.2. Feature Extraction:

For feature extraction, three descriptors have been selected and used: Iris coding method based on Differences of Discrete Cosine Transform (DCT) [160],1D-LogGabor Feature Extraction or Masek et al algorithm (LG) [161] and Algorithm of Rathgeb and Uhl (CR) [162].

a. Differences of Discrete Cosine Transform:

Due to Differences of Discrete Cosine Transform's much lower complexity, it can be considered as a computationally intensive replacement for the Karhunen Loeve Transform (KLT). DCT is a real valued transform and it calculates the truncated Chebyshev series processing minimax properties.

b. One dimensional log – Gabor feature extraction

The algorithm proposed by Masek et al. [161] examines 1D – intensity signals applying a dyadic wavelet transform and a Log-Gabor filter, respectively.

The frequency response of a Log-Gabor filter is given as (Eq. 3-1):

$$G(f) = exp\left(\frac{\left(-log\left(\frac{f}{f_0}\right)^2\right)}{2\left(log\left(\frac{s}{f_0}\right)^2\right)}\right) \quad (3\text{-}1)$$

Where, f_0 is the center frequency, and s is the bandwidth of the filter.

c. Algorithm of Ratgheb et al.

This feature extraction method is based on comparisons between gray scale values. The features can be extracted by examining the local intensity variations in the iris texture. This technique also includes a post iris texture image processing stage in order to eliminate the small peaks of pixel paths by determining threshold. The Ratgheb et al. 's descriptor needs no complex calculation. In order to segment the iris and calculate the comparison score between samples the University of Salzburg Iris Toolkit (USIT), available at http://www.wavelab.at/sources/ Rathgeb12e/ [163] has been successfully used.

3.4.3. Matching:

Using USIT, the comparison scores can be achieved based on calculation of Hamming distance. Hamming distance is the measure of similar bits between two bit patterns. The formula is given as (Eq.3-2):

$$\mathrm{HD} = \frac{1}{N}\sum_{j=1}^{N} X_j \; ! \; Y_j \tag{3-2}$$

The all vs. all comparison scenarios has been used for achieving the maximum possible number of matching scores.

3.5. Results and Discussions

In this section, the obtained comparison scores between samples will be presented and discussed. According to number of possible comparison scores, up to 900 K results have been achieved. As it can be illustrated in Figure.2, the empirical cumulative distribution functions of genuine scores obtained by a- DCT b- 1d-log gabor and c- CR codes show statistically significant differences between those comparison scores obtained by comparing diabetic iris samples and those matching results achieved by comparison of iris pattern images taken from healthy irises. Based on the obtained biometric results, user identification tends to be harder under influence of diabetes and the accuracy of iris recognition system for healthy irises is higher. The results show that if the gallery and probe images are taken from healthy eyes, every recognition system yields the best performance. But for identification of users under influence of diabetes the reduction in performance is observed.

This is due to the fact that, there are some non-obvious disorders in diabetic eyes which appears to identification task harder. In order to test if first null hypothesis which states that: "the mean values of distributions are same" and second null hypothesis which claims that: "the obtained samples are drawn from same distributions" can be proved or not?, we have used t-test and Kolmogorov Simonov tests respectively. According to chosen confidence threshold (0.01) and obtained p-values which are near zero the both null hypothesizes can be definitely rejected and we can concluded that the performance of iris recognition system for healthy eyes is better than diabetic ones (Table.3-3). In the other words, it is harder to recognize people who are suffered from diabetes according to some non-obvious disorders in their iris textures. The ROC curves for healthy and diabetic groups have been presented in Figure.3. According to Figure.3-3, the healthy eyes are easier to be recognized in comparison with non-healthy eyes. It is also worth mentioning that the DCT has best performance while the 1d log Gabor is worst descriptor for feature extraction based on the obtained results.

Donating image samples of iris with high quality from old population was less easy than young age group, however, younger users can provide iris biometric samples in higher quality so as a result, age progression can be considered as a source of error and the age increasing may have an effect on the quality of the acquired biometric data. Hence, maybe, part of the observed differences between the obtained comparison results for the mentioned two groups would be the consequence of the difference in mean age of healthy and diabetic groups. Area under curve (AUC) can be defined as:

$$\frac{\text{The Coverage under ROC curve-Emprical}}{\text{The Coverage under ROC curve-Ideal}} \times 100 \tag{3-3}$$

According to Eq.3-3, as the accuracy of recognition system increases the coverage will increase. In ideal case, when we can enhance the accuracy of system to one hundred percent, (when true acceptance rate equals with one at false rejection rate = 0), the AUC would be 100%. As a result higher AUC means better performance evaluation results. Table.3-2 presents the AUC for different ROC curves which have been presented in Fig.3-3. The biological age of individuals can obviously make changes in the biometric data. For instance, for different incident lighting level, the human iris ability for adopting and accommodating will be sharply decreases with age, and this in turn can lead to issues about the availability of the features used to recognize a particular individual's iris pattern. Hence, as the users are coming from different age groups, so maybe this factor can be considered as one of the influential parameters. Our results indicates that the DCT has best performance. However, the iris recognition is less effective for people with diabetes. To check whether the algorithm itself is sufficient enough or not, in Appendix.A, the same matcher is used for another special case. The EER of 1% is achieved.

3.6. Conclusions

This chapter was concerned about the reliability test of iris recognition system under influence of diabetes. A new database has been collected and offered in this manuscript. We have used four different matchers, in order to obtain the similarity scores between the captured samples. Although there is no obvious impairment on the non-healthy irides, but according to the results achieved by all four matchers (3 open source codes and one commercial closed one), the

accuracy of system is higher when we want to recognize healthy people using their iris texture images. The Best performance has been observed by using the methodology proposed by Monro et al. (USITv.2). We have just used the codes and we didn't modify/change any of them. It must be also noted that, due to the physiology of pupil dilation mechanisms differences, (pupil dilation responsiveness decreasing with age) for users from different age groups, the difference in mean age of the chosen groups must be considered as an additional error source.

Condition	Age Group (%)	Gender (%)	Total number of iridies	Total number of pictures
Healthy	Less than 20 years old: 21%	Female: 62%	162 irides	546
	Between 20 and 40 years old: 61%			
	Between 40 and 60 years old: 12%	Male: 38%		
	More than 60 years old: 6%			
Diabetic	Less than 20 years old: 1%	Female: 80 %	181 irides	772
	Between 20 and 40 years old: 17.5%			
	Between 40 and 60 years old: 46.5%	Male: 20 %		
	More than 60 years old: 35%			

Table.3-1. Demographics of collected database

Methodology	Healthy AUC	Diabetic AUC
DCTC	0.9658	0.9027
1D-LogGabor	0.8272	0.7721
CR	0.8927	0.8203

Table.3-2. AUC for different matchers. It shows that results are better for healthy eyes

Method	KS - test	T - test
DCTC	5.368 e -10	2.875 e-11
1d-LogGabor	2.921 e-7	9.531 e -17
CR	6.676 e-15	1.568 e-16

Table.3-3. Hypothesis test results. It shows that we can reject the null hypothesis.

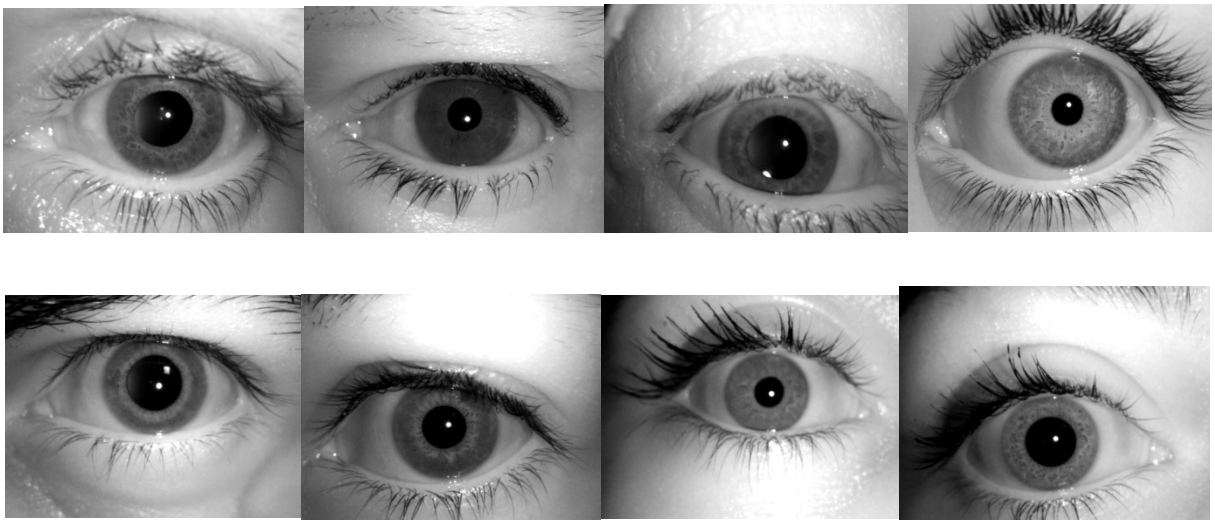


Figure.3-1. Samples from the database - First row: Diabetic eyes, Second row: Healthy eyes - Captured by mono ocular Iri Shield USB MK 2120U.

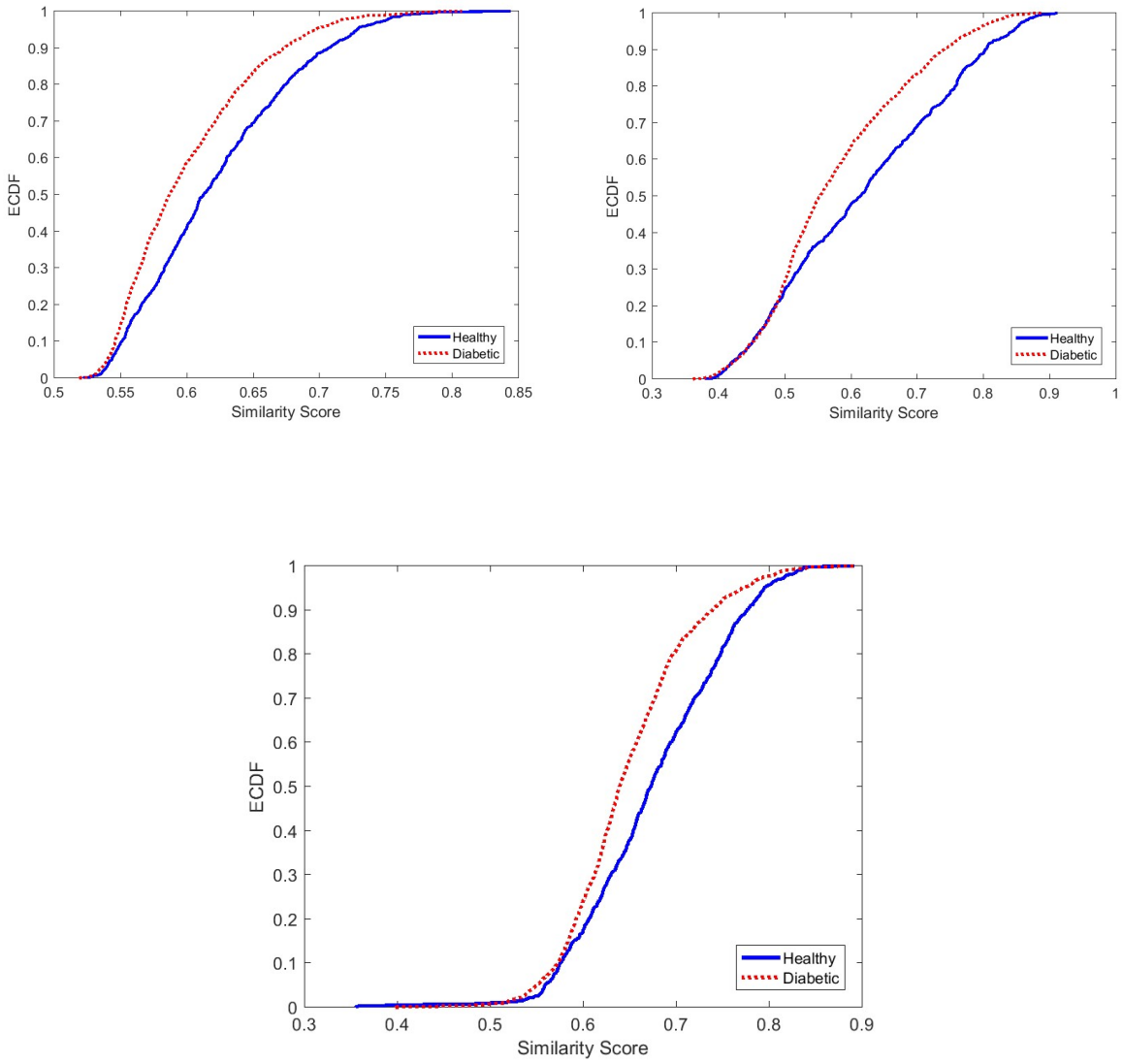
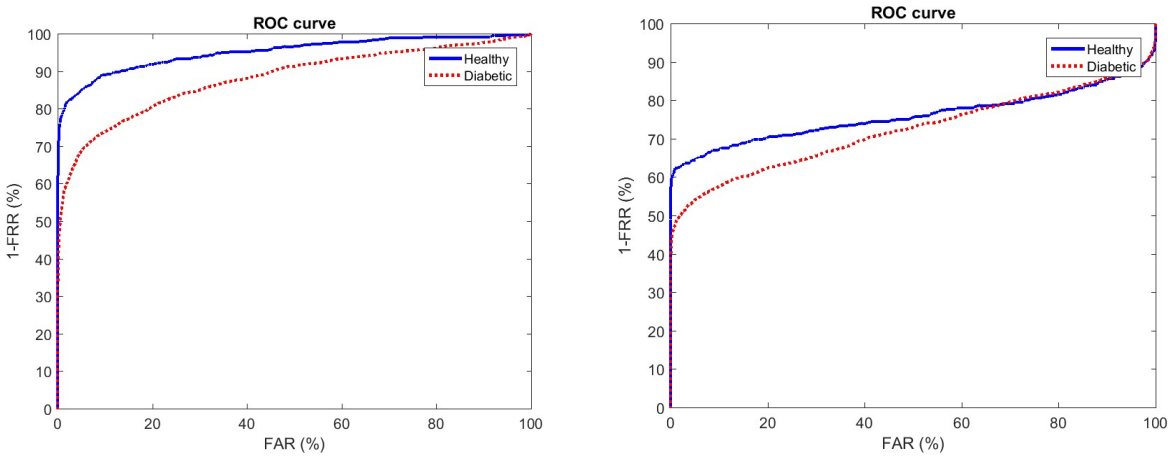
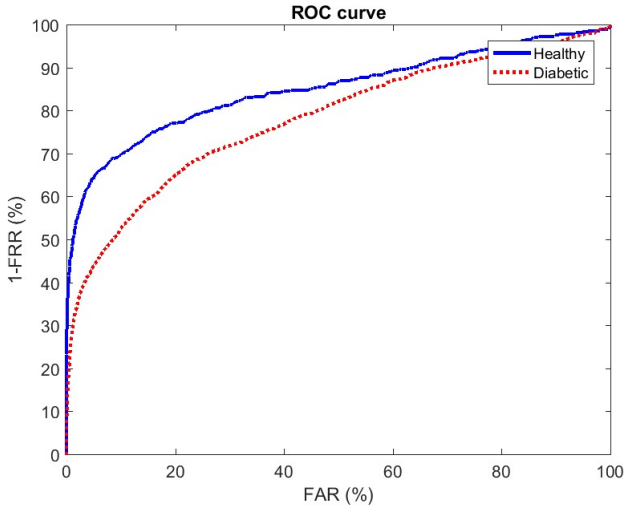


Figure.3-2. Empirical Cumulative Distribution Function - upper left: DCTC , upper right: 1D-LogGabor, down: CR, The figures show that similarity scores between healthy iris samples are higher.





c

Figure.3-3. ROC curves - upper left: DCTC , upper right: 1D-LogGabor, down: CR, The figures show that iris recognition is less effective for people with diabetes type II and DCTC has the best performance.

4. Age-dependency of the Diabetes Effects on the Iris Recognition Systems Performance Evaluation Results

This Chapter was previously published in Journal of Biomedical Engineering/Biomedizinische Technik in 2020, Authors: [M Azimi, SA Rasoulinejad, A Pacut], Title: "Age dependency of the diabetes effects on the iris recognition systems performance evaluation".

4.1. Overview

In the previous chapter we demonstrated that iris recognition is less effective for people with type II diabetes. In this chapter, we attempt to answer the questions whether iris recognition task under the influence of diabetes is more difficult, and whether the effects of diabetes and an individual's age are correlated. We hypothesized that the health condition of the participants plays an important role in the performance of the iris recognition system. To confirm the obtained results, we reported the distribution of the usable area in each subgroup in order to have a more comprehensive analysis of the effects of diabetes. There was previously no study that had investigated for which age group (young or old) the diabetes effect is more acute on the biometric results. For this purpose, we created a new database containing 1,906 samples from 509 eyes.

We applied the Weighted Adaptive Hough and Ellipsopolar Transform techniques and contrast-adjusted the Hough transform for the segmentation of the iris texture, along with three different encoding algorithms. This time, by a visual examination of the samples manually, we inspected the iris segmentation results, in order to be definitely sure about the high success of the segmentation phase.

To test the hypothesis related to the physiological aging effect, Welches's t-test and Kolmogorov-Smirnov test were used to study the age-dependency of the influence of diabetes mellitus on the reliability of our chosen iris recognition system. Our results give some general

hints related to the effect of age on the performance of biometric systems for people with diabetes.

Therefore, this chapter turns to look at the reliability test of the iris recognition system under the influence of diabetes, with special consideration of an additional factor, namely the physiological aging parameter.

4.2. Introduction

Aging can be considered as one of the most influential parameters which can degrade the matching accuracy of the biometrics recognition due to the variation of biometric traits through time. This effect accounts for the loss in the system's performance due to the increase of the time-lapse between the different acquisition sessions. But even assuming a short time differences between the gallery and probe acquisitions, the variations in accuracy between different user groups according to their age must be taken into account. The absolute age has a high influence on how a user can authenticate him/her using a biometrics recognition system in an appropriate way. This problem arises for almost all biometric modalities.

Another problem with age progression is that the acquisition of high-quality biometric samples, especially for fingerprint modality, is not easily possible for an older population due to the fact that, as age increases, loss of moisture content from the skin is more probable which can change the skin into drier one, and hence it can reduce the reference quality.

The pupil dilation is one of the most important variation factors in iris features. This factor may be classified as 'aging factor'. There is a question if the aging relates also to the iris texture, i.e. a direct donor of the biometric features. Pupil dilation cannot directly influence the unique iris features themselves, but it is worth considering as the eye's capability for pupil dilation decreases with age.

Erbilek and Fairhurst [36] presented the experimental study to analyze the effects of time the separation between the enrollment (when the reference data are first recorded and stored) and authentication (a specific identification/verification event) for iris modality. They show a significant difference in performance across time. Erbilek and Fairhurst [36] have investigated the stability of other biometric modalities as a function of age and have highlighted problems in

terms of feature presence within an elderly population that cause difficulties for conventional systems for devices such as fingerprint, hand geometry and face.

In this chapter, we present a performance analysis of iris recognition system for healthy and diabetes affected irises, separately for younger and older users.

4.3. The extended database

The database used in this study is an extended and modified version of our database presented in the previous chapter. Due to the very low quality of some samples, we have firstly, excluded almost 3% of the diabetic iris samples and 2% of images taken from healthy eyes. The healthy sub database were collected with the same experiment protocol as the diabetes one, in order to be able to make a more powerful statement about the effect of diabetes on the performance evaluation results of the chosen iris recognition system. It is also important to clarify that 8 samples (0.6% of total samples) were labeled incorrectly in our first database, and the problem is fixed now (there was a mistake between left and right eyes). The new modified database contains iris images collected from volunteers of different age and gender groups. All the individuals participating in the experiment have been presented with a detailed information on the research by providing a form that contained all the project details, and the consent agreements were filled up by each volunteer. All of the users were Iranian with black/brown eyes. The only limitation for the eye was to have a clearly visible iris pattern. Volunteers have informed us for how long they are taking diabetes pills or insulin and no medical tests have been performed to confirm their diabetes states; also, no medical tests were performed and no drugs were used in the collecting of the database.

Hence, the new and modified version of our database is offered here. This database contains 1906 samples from 509 eyes (723 diabetic iris images from 161 eyes and 1183 healthy iris images from 348 ones). The data samples have been captured using a commercial iris capture device: “Iri Shield USB MK 2120U ” , which was connected to a galaxy A5 smartphone for storage of iris samples. Iri Shield USB MK 2120U is a mono iris NIR capture device widely used for capturing iris texture pictures and the captured iris images are compliant with ISO/IEC 19794-6 standard. The device also provides information related to the iris usable area in the

taken iris image. According to the user's guide documents released by the same company, the usable area which is an index for measuring the quality of the donated sample should be greater than 70 to be ideal for enrollment. They also suggested that in the case of the usable area is less than or equal to 50, the image should be rejected.

Fig. 4-1. First row: Diabetic eyes- (first from the left: sample with high usable area – young user, second from the left: donated by an old user), second row: (first from the left: sample with high usable area – young user, second from the left: donated by an old user) - Healthy eye – Captured by mono ocular Iri Shield USB MK 2120U.

It is also worth mentioning that the new experiment was done under the guidance of Professor Rasoulinejad at his ophthalmological office and also, the data collection protocol was approved by the ethics committee of the Warsaw University of Technology.

In Fig.4-1, One samples from each group have been presented. The first row contains healthy samples while the diabetic iris images are presented in the second row. The first sample of each row is a sample with high quality provided by a young user while the second iris sample of each row is a sample taken from an old eye. All of the volunteers are Iranian and are living in same city with same dietary culture.

4.4. Methodology

The objective of this experiment is to figure out whether the age of the user can play a significant role in the performance of iris recognition systems under the influence of diabetes. For this purpose, we have firstly partitioned the database into healthy and diabetic sub-databases and then, the diabetic samples were also partitioned into two different age groups.

We conducted the same test with three different iris recognition systems to make sure they were testing the eyes, and not the quality of different algorithms. The first step for recognition is to detect the region of interest and segment the iris pattern from the whole ocular image.

Like the previous chapter, in order to segment the iris and calculate the comparison score between samples we successfully used the University of Salzburg Iris Toolkit (USIT), available at http://www.wavelab.at/sources/ Rathgeb12e/ [162].

4.4.1. Iris Segmentation (Weighted Adaptive Hough and Ellipsopolar Transform):

Weighted Adaptive Hough and Ellipsopolar Transform technique is a two-stage iris segmentation technique;

a. Finding the center point: the center of multiple approximately concentric rings at iteratively refined resolution can be determined by removing the detected reflection mask and detecting the edges. Finally, the weighted adaptive Hough transform is applied at multiple resolutions to estimate the approximate position of the iris center.
b. The center is used for the purpose of extracting the region of interest. Firstly the initial boundary is detected and after the first iteration, the ellipsolar transform is applied for the inner and outer boundaries detection. The ellipsopolar transform finds the second boundary based on the outcome of the first.

To exclude gross segmentation errors, we inspected the iris segmentation results manually, by visual examination of the samples one by one. Some of the irises happened to be segmented incorrectly or incompletely, hence, for the rest of samples, we have used a different method of iris image segmentation called Contrast-Adjusted Hough Transform Segmentation Algorithm.

4.4.2. Iris Segmentation (Contrast-Adjusted Hough Transform):

This iris segmentation algorithm is a modified implementation of a Hough Transform approach. It uses contrast adjustment to locate pupillary and limbic boundaries in iris images, the

Canny edge detection for detecting curves of boundary, and finally some enhancement techniques for removing unlikely edges. The samples of successful segmentation and unsuccessful segmentation are presented In Fig.4-2. In summary, we are confidents about the success of the segmentation phase.

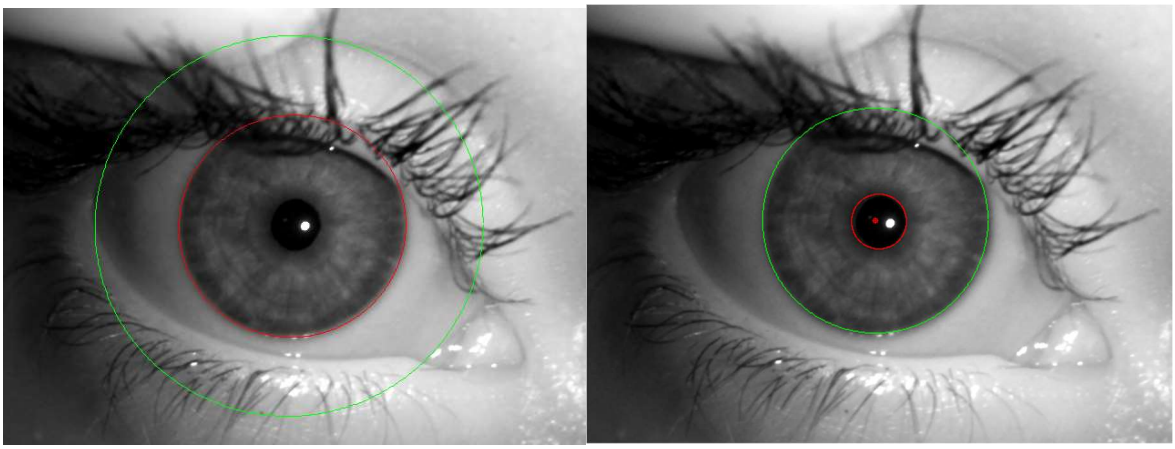


Fig. 4-2.a- Unsuccessful segmentation , b- Successful Segmentation

After this stage, the extracted iris texture will be normalized using Daugman's rubber sheet model. For extracting the features from the segmented iris patterns, we have used three different descriptors from the USIT package: a- the code using the Monro et al. algorithm for generating an iris code from iris texture - Differences of Discrete Cosine Transform (DCT), b- the code using the Rathgeb algorithm for generating an iris code from iris texture (CR) and 1D-LogGabor Feature Extraction or Masek et al algorithm (LG).

4.4.3. Differences of Discrete Cosine Transform (DCT):

As it has been mentioned in the chapter.3, Due to Differences of Discrete Cosine Transform's much lower complexity, it can be considered as a computationally intensive replacement for the Karhunen Loeve Transform (KLT). DCT is a real-valued transform and it calculates the truncated Chebyshev series processing minimax properties.

4.4.4. Algorithm of Ratgheb et al. (CR):

This feature extraction method is based on comparisons between grayscale values. The features can be extracted by examining the local intensity variations in the iris texture. This

technique also includes a post iris texture image processing stage in order to eliminate small peaks of pixel paths by thresholding the paths. Ratgheb et al.'s descriptor needs no complex calculation. Here, we are using a database in which the non-biometric data like age and gender are available for each of the samples.

4.4.5. 1D-LogGabor Feature Extraction (LG):

The algorithm proposed by Masek et al. examines 1D – intensity signals by application of the dyadic wavelet transform and a log-Gabor filter, respectively. The frequency response of a Log-Gabor filter is given as (Eq.3-1).

4.5. Results and Discussions

The structure of this section is as follows: firstly, we will present a figure to give some useful information related to the usable area of the donated samples for two groups: healthy and diabetes-affected. So the figure contains two plots: one for diabetic users and the other for the healthy ones. Then, ROC curves for healthy and diabetic groups will be presented. Finally, we use all vs. all comparison scenario by comparing all of the samples with each other, to investigate the simultaneous effects of diabetes and age factor. The information about the diabetic state was also available for the users but as we did not want to perform any medical testing (due to the restrictions of our approved experiment protocol), we did not know whether the duration of the diabetic state was the duration of controlled or uncontrolled diabetes. Consequently, no information related to the duration of the diabetic state is will be employed here. The numbers of genuine and impostor scores are tabulated in Table.4-1. In Fig.4-3 we present the distribution of the percentage of the usable iris area. The mean value of the same parameter for healthy eyes is visibly higher. The first reason is that the mean age of the users in the healthy group is lower than the mean age of those who are categorized as diabetic ones. For younger users, it's easier to keep their eyes open for a longer time and be less sensitive to the illumination and donate more qualified samples.

Type	Number of genuine scores	Number of impostor scores
Healthy	5 169	1 394 320

Diabetes	4 241	518 488

Table. 4-1. Number of comparison scores for two groups

As illustrated in Fig. 4-3, less than 10 percent of the captured images have a usable area less than or equal to 60%, which is acceptable. The effectiveness of our chosen iris recognition systems are represented graphically by the receiver operating characteristic (ROC) curves.

To calculate the comparison scores, each sample was compared to all of the samples in each subgroup and as the matcher returns symmetrical matching scores (the score between the images A and B is equal to the matching score between B and A).

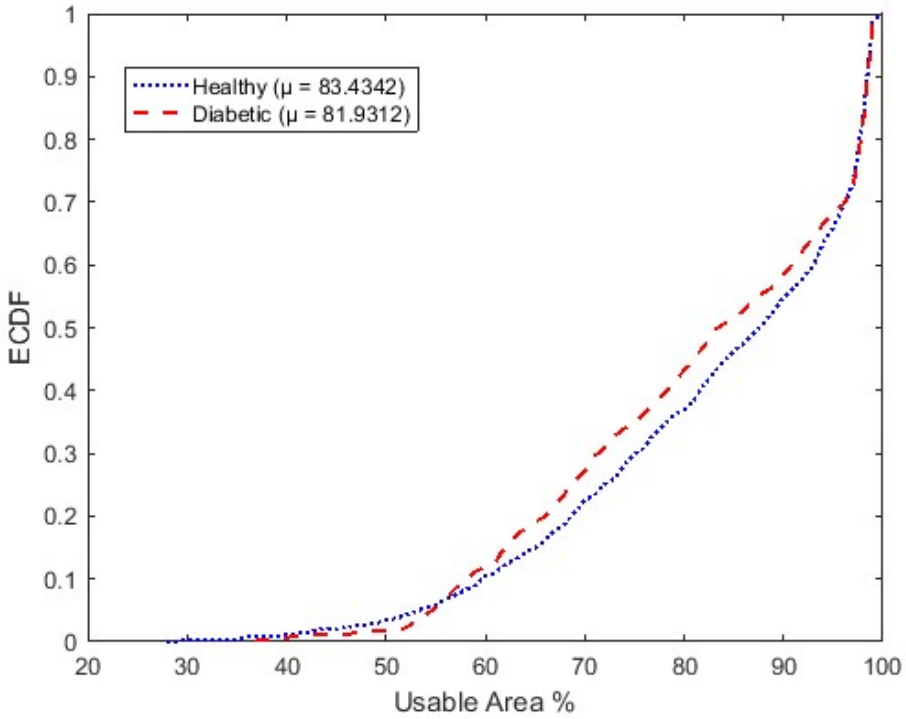


Fig. 4-3. Usable Area: An index for quality of the captured images. This figure shows the effect of environmental factor on the iris samples.

In Figs. 4-4, we presented the ROC curves for diabetic eyes and healthy ones to evaluate the performances of algorithms. The results presented in Figs. 4-4 are achieved by the use of DCT, CR , and LG codes, respectively. As it has been shown in Figs. 4-4, the equal error rate is higher for the diabetic group. The achieved equal error rate for the healthy group using the DCT method is around 5.6 % while the same value for the diabetic group is about 12.4%. According to the same figures, identification of healthy eyes is an easier task to do. We have conducted the same test with three different iris recognition systems to make sure we were testing the eyes, and not the quality of different algorithms. In each case, the results were the same. All three systems more easily identified the healthy irises and were less accurate when scanning diabetic eyes. To account for the results, we noted that diabetes can affect the eyes in a number of ways that may not be obvious, causing retinal damage, cataracts, and glaucoma.

In ideal case, when we can enhance the accuracy of system to one hundred percent, (when true acceptance rate equals with one at false rejection rate = 0), the AUC would be 100%. As a result higher AUC means better performance evaluation results. Table. 4-2 presents the AUC for different ROC curves.

In Figs 4-5 the empirical cumulative distribution functions of genuine scores for all versus all comparison scenarios have been presented. As are shown in Figs 4-5 (the empirical cumulative distribution function of genuine scores), for all of the three iris recognition algorithms (DCTC, 1D-LogGabor and CR), the biometric results for healthy eyes are better than the ones achieved for diabetes affected eyes. This indicates that there are some non-obvious disorders in diabetic eyes that makes the identification task more difficult. The observed differences can also be interpreted as the performance loss under the influence of the age factor.

It is important to clarify that for achieving the ECDF curves presented in Figs. 4-5, we have eliminated the zero values from the datasets, namely we did not compare identical samples with each other. The reason is that there are different numbers of samples in each group. Now we attempt to answer the main question: Is diabetes effect on the biometric results age dependent? We partitioned the diabetic database into two subgroups: (1) the samples donated by individuals who were less than 45 years old and (2) the iris images provided by those volunteers who were more than 45 years old. We removed from the sample the individuals age 45 or older since, after age 45, the prevalence of diabetic is higher than that of healthy (all the users are Iranian and they

are living in a same Babol region) due to the previously published paper by Rasoulinejad et al. [176].

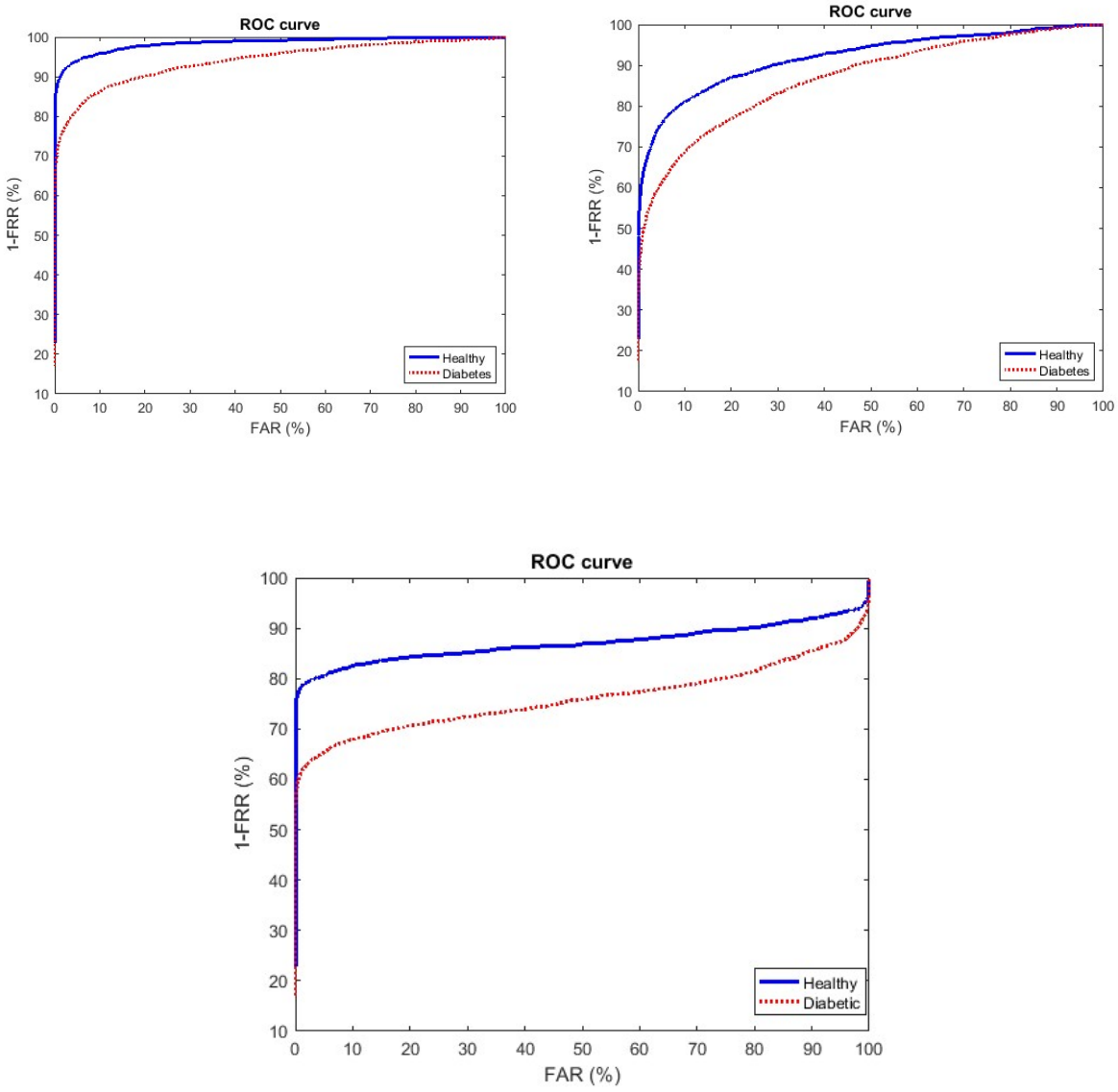


Figure.4-4. ROC curves - upper left: DCTC , upper right: 1D-LogGabor, down: CR, The figures show that iris recognition is less effective for people with diabetes type II and DCTC has the best performance.

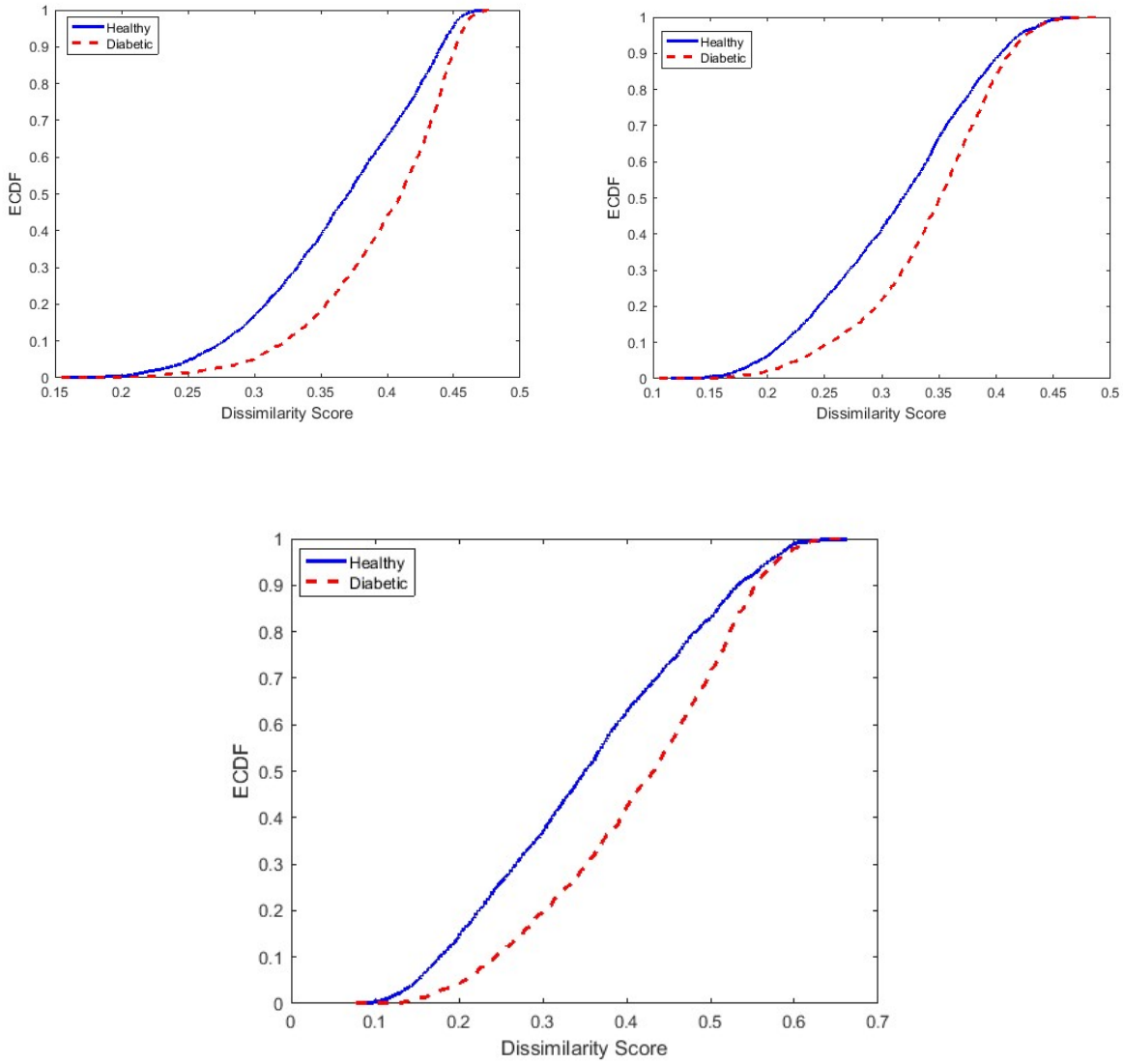


Figure.4-5. Empirical Cumulative Distribution Function - upper left: DCTC , upper right: 1D-LogGabor, down: CR, The figures show that similarity scores between healthy iris samples are higher.

Methodology	Healthy	Diabetic
DCTC	0.984711	0.940504
CR	0.919835	0.868877
LG	0.870452	0.760800

Table 4-2. AUC for different ROC curves – This table shows that the observed differences between the AUC for healthy and illness affected eyes were independent from the success of segmentation process.

In Fig. 4-6, we show the empirical cumulative distribution function of genuine scores for both young (blue line) and old volunteers (red line) who suffered from diabetes). In Fig 4-7, the ECDF of impostors' results is shown. As both figures show, the performance evaluation results of iris recognition systems tend to be better for the younger population.

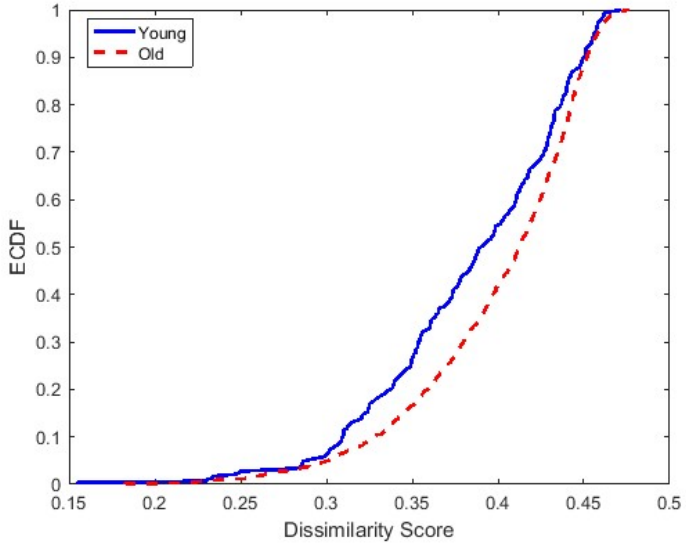


Fig. 4-6. The empirical cumulative distribution functions of genuine scores for younger and older volunteers with diabetes by DCTC. The figure shows that results for younger people are better.

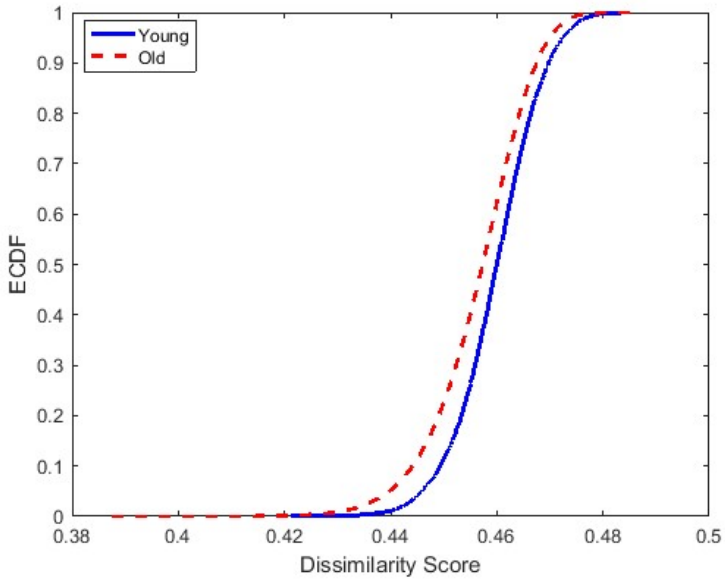


Fig. 4-7. The empirical cumulative distribution function of impostor scores for younger and older volunteers with diabetes by DCTC. The figure shows that results for younger people are better.

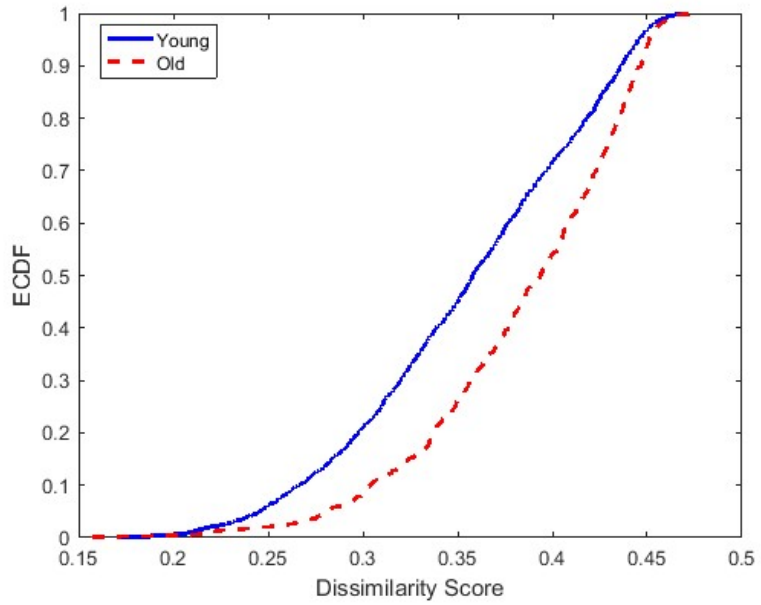


Fig. 4-8. The empirical cumulative distribution function of genuine scores for younger and older healthy volunteers by DCTC

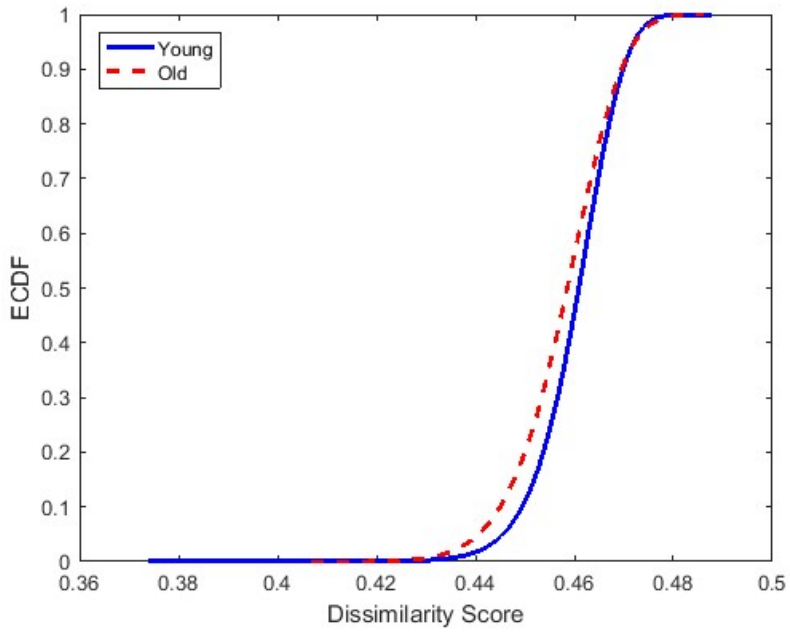


Fig. 4-9. The empirical cumulative distribution function of impostor scores for younger and older healthy volunteers by DCTC

However, to judge whether the observed differences in dissimilarity scores across partitions can be considered as samples drawn from the same distribution, a two-sample Kolmogorov-Smirnov test has been applied with the significance level $\alpha = 0.05$. The null hypothesis H0 in this test states that the samples originating from the two compared sub-groups are drawn from the same distribution. The alternative hypothesis claims that at least one value does not match the specified distribution.

We have also used t-test for judging if the observed difference between the mean values of the dissimilarity scores distributions for the mentioned subgroups are statistically significant different. For t-test, the null hypothesis is that the means are equal and the alternative hypothesis is: H1: F_diabetic(more than 45 years old) > F_diabetic(less than 45 years old).

We also analyzed the results in order to clarify whether the difference between biometric impostor scores for two different groups is statistically significant. In all cases (both for genuine scores and impostor scores distributions), two sample t-test and Kolmogorov Simonov test were applied (Table.4-3).

Test	P-value (genuine scores)	P-value (impostor scores)
t-test	1.9514e-10	~ Zero
Ks-test	3.4795e-10	~ Zero

Table 4-3. Results for t-test and K-S test (less and more than 45 years old). The results indicate that we can reject the null hypothesis.

According to the hypothesis tests results in Table. 4-3, the null hypotheses can be rejected. The empirical cumulative distribution function of impostor and genuine scores for healthy young (blue line) and healthy old participants (red line) are depicted in Figures 4-8 and 4-9 respectively. According to Figures 4-6 to 4-9, old individuals with diabetes type II are the most likely groups to be mistakenly not recognized as true users by biometric recognition systems.

4.6. Conclusions

This study highlights the limits of iris biometric techniques, and cautions against the widespread use of modalities that only perform well in optimal circumstances and do not account for relatively common conditions like diabetes.

In this paper, in order to investigate the age-dependency of diabetes effects on the iris biometric recognition system, the extended/modified version of our previously collected database and USIT popular iris recognition toolkit were used. Just by repartitioning of the new database and by the use of the same recognition system, we also made additional experiments to show whether the age of participants itself is an influential parameter for the purpose of verification under the influence of diabetes. According to our results, the biometric system's reliability will experience the lowest change for younger users under the influence of diabetes. In other words, the quality of the biometric system tends to be higher when we are trying to verify younger participants under the influence of diabetes. This paper also indicates that the diabetes effects on the reliability of the iris recognition system must be considered age-dependent. The final remarks are as follows:

- Iris recognition is less effective for people with diabetes regardless their age.
- Although diabetes cannot destruct the iris itself yet it can affect the eyes in a number of ways that may not be obvious, causing retinal damage, cataracts, and glaucoma.
- As the eyes capability for pupil dilation decreases with the age, and older users cannot donate high qualified samples in most of the cases, the problem with iris recognition might result from the difference in the mean age of groups.
- This chapter also gives a hint that the diabetes effect on the reliability of iris recognition can be considered age-dependent.
- The degradation in the accuracy can be also caused by of occlusion, illumination, etc. Hence, collection of a new database in an extremely restricted condition is encouraged.
- The observed degradation in the performance of the iris recognition system may be also due to the eye color of the users (in our research everybody has black eyes).
- The controlled group of participants was of the same race and living in the same area with the same dietary culture, so that is one of the advantages of using our newly offered database.
- The number of samples is not high enough to make a general statement but the number of classes is adequate enough to be confident about the achieved results.

- Iris recognition is a popular form of identification for government programs, however, according to our findings, we must be careful in the widespread use of this a specific modality that only perform well in certain optimal circumstances.

This research shows that the results of “iris scanning” can be less accurate when a person suffers from diabetes. So the statement would be: “From biometric science point of view, it is more probable for diabetic irises to be mistakenly rejected by the iris recognition system and the reliability of iris recognition system under the influence of diabetes tends to be higher for younger patients”.

5. Gender-dependency of the Diabetes Effects on the Iris Recognition Systems Performance Evaluation Results

This chapter is an extended version of a conference article presented in: IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Title: "The Effects of Gender Factor and Diabetes Mellitus on the Iris Recognition System's Accuracy and Reliability", 2019, Authors: [M Azimi, SA Rasoulinejad, A Pacut]

5.1. Overview

In the previous two chapters, we demonstrated that iris recognition is less effective for users with type II diabetes, regardless of their age. In this chapter, we want to discover whether iris recognition is less effective for people with diabetes regardless their gender too.

For this purpose, we analyzed the gender dependency of the effects of diabetes on the reliability of iris recognition systems, using a database containing a mixture of diabetic and healthy irides for both sexes. The matching scores (similarity scores between the samples) were obtained by implementing three image-processing algorithms. We used open-source codes applied in USIT. As mentioned above, the database contains 1906 samples from 509 eyes (723 iris images from 161 diabetic eyes and 1183 iris images from 348 healthy eyes).

However, for the purpose of conducting the numerical experiments in this chapter, we firstly partitioned the samples into four groups: healthy women (760 samples), Healthy men (423 samples), diabetic women (593 samples), and diabetic men (130 samples). For each of the mentioned groups, the empirical cumulative distribution functions of genuine and impostor results were presented. We have also given the area under the curve of a number of ROC curves. According to the results achieved by the methodology proposed by Zhang, Monro, and Rakshit, which is considered to give the best performance, iris recognition is less effective for people with type II diabetes, regardless of their gender. In summary, while the results are better for healthy women, iris recognition systems perform less successfully when we attempt to identify women with diabetes.

5.2. Introduction

Despite iris recognition being one of the most reliable methods of identifying individuals, there are many parameters that can attenuate the accuracy and reliability of iris recognition systems. There will always be a tradeoff between true rejection and false acceptance rates by the appropriate adjustment of thresholds. An appropriate threshold is closely related to the required security level. Additionally, there are a number of social issues that affect biometric technology. User-related parameters include physiological factors, such as age, gender, and behavioral factors including habituation and cultural restrictions, can influence the biometric sample characteristic.

Further to the previous chapters, diabetes can affect the matching accuracy of iris recognition systems, and healthy eyes are easier to recognize than diabetic ones.

On the other hand, based on the research works [110] concerned with predicting the gender of a person based on an analysis of the features of the iris texture, the iris can be used to indicate gender. Howard and Etter [91] carried out research to investigate whether factors such as gender, race and even eye color can play a key role in the performance assessment results of an iris recognition system for various participants across the population. According to the results they achieved, African American participants with dark eyes (brown or black eyes) are the most likely groups to be mistakenly rejected as false users by iris verification systems. They found that the observed degradation in the performance of the iris recognition system may also be due to the eye color of the users.

The construction of this chapter is as follows: First, Section 5.3 looks at the main challenges to the reliability of iris recognition systems and briefly reviewed the role of the gender factor. In Section 5.4, brief information about the pre-collected samples of the database will be presented. In Section 5.5, the methods will be explained and the results, discussions and conclusions will be presented in Sections 5.6 and 5.7 respectively.

5.3. Related works

There are several forms of physiological/biological evidence to prove that the general effects of diabetes on body organs are gender dependent.

The prevalence of impaired fasting glucose and impaired glucose tolerance, the prevalence and incidence of type 2 and type 1 diabetes, and the sex-specific effects of testosterone and estrogen deficiency and excess were discussed by Mauvais-Jarvis [164] in a recently published manuscript. According to a published review article written by Kautzy-Willer and Harreiter [165], diabetic women come under greater cardiovascular risk. In another review, Campesi et al. [166] discuss the sex-gender differences that are known to have an impact on diabetes, mainly by focusing on the cardiovascular complications. They also gathered scientific materials in order to investigate therapeutic approaches to managing diabetes-associated cardiovascular complications and how differences in sex-gender can influence the existing therapeutic approaches.

On the other hand, the non-obvious ocular abnormalities due to the diabetes mellitus in the iris tissue pattern can also affect the matching scores. Hence, in our study we hypothesize that the effect of diabetes on iris is gender dependent and we want to investigate for which gender the difference between the obtained scores is higher.

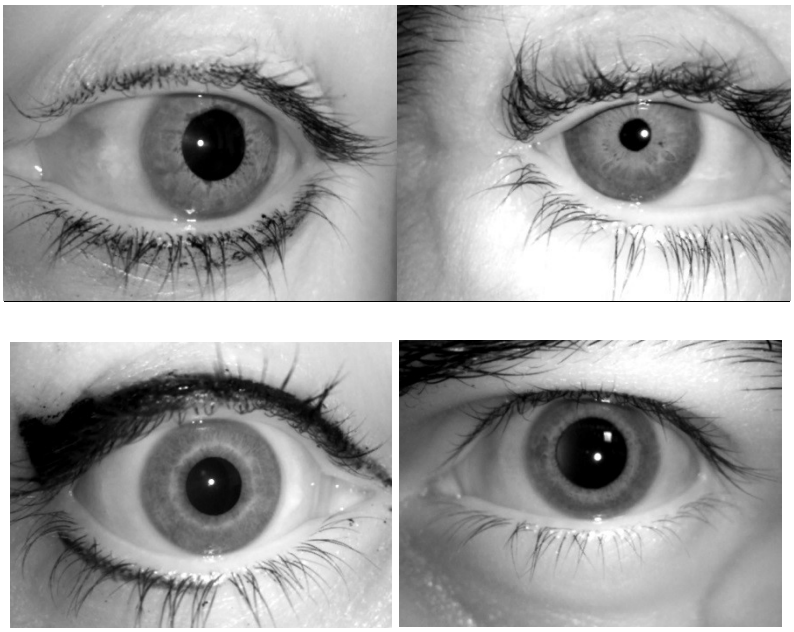


Fig. 5-1. Samples from the database - Top row: Diabetic eyes- (left: female user, right: male user), bottom row: Healthy eyes (left: female user, right: male user) – captured by a monocular IriShield USB MK 2120U.

5.4. Partitioning the pre-collected Iris Database

This database contains 1,906 samples from 509 eyes (723 iris images from 161 diabetic eyes and 1,183 iris images from 348 healthy eyes). The data samples were captured using a commercial iris capture device: "IriShield USB MK 2120U" connected to a Galaxy A5 smartphone for storing the iris samples. The database comprises iris images collected from volunteers before routine ophthalmology examinations and before applying a dilation drop. All the participants of the experiment were provided with detailed information on the research (on a form containing all the details of the project) and signed a consent form. The only limitation for participants was to have a clearly visible iris pattern without any ocular abnormalities. In Figure 5-1, four samples are shown to illustrate diabetic (top left: female user, top right: male user) and healthy (bottom left: female user, bottom right: male user) eyes.

We partitioned the samples into four groups: healthy women (760 samples), healthy men (423 samples), diabetic women (593 samples), and diabetic men (130 samples).

5.5. Methodology

In order to compare the sample scores (for all vs all comparison scenarios) we used the University of Salzburg Iris Toolkit (USIT). The iris segmentation method is the Weighted Adaptive Hough and Ellipsopolar Transform (WAHET). The WAHET technique is a two-stage iris segmentation technique. In the first step, the center point must be found. To exclude gross segmentation errors, we inspected the iris segmentation results manually, making a visual examination of the samples one by one. Some of the irides happened to be segmented incorrectly or incompletely, so for the rest of the samples we used a different method of segmenting the iris images called Contrast-Adjusted Hough Transform Segmentation Algorithm. The extracted iris textures were normalized using Daugman's rubber sheet approach. We selected three methods of extracting the features of the iris: the iris coding method based on 1D LogGabor (LG), the algorithm of Rathgeb and Uhl (CR) and the differences of discrete cosine transform method (DCTC). All the images were compared to the entire database. A lower distance means a higher similarity between the samples.

5.5.1. Statistical Distance

To calculate the discrepancy between iris images, we applied Bhattacharya distance D_B, namely for two distributions, p and q, over the same domain X.

$$D = -\ln \sum_{x=X} \sqrt{p(x)q(x)} \quad (5\text{-}1)$$

5.6. Results and Discussions

In this section we present the results derived using the technology mentioned above. The main question that this part of the study wants to answer is: "for which gender does the iris recognition system perform better under the influence of diabetes?" In order to answer this question, standard deviations from the genuine score distributions for four different groups: a- Healthy Female, b- Healthy Male, c- Diabetic Female, d- Diabetic Male have been tabulated in Table 5-1. We calculated genuine and impostor results presented for the three different matchers used, so twelve values of Bhatcharyya distances were obtained (six distances between healthy and diabetic genuine scores for female and male users, and six distances between healthy and diabetic impostor scores independent of gender).

Method	DCTC	CR	LG	**Method**	DCTC	CR	LG
Women - genuine	Healthy: μ=0.358 σ=0.061	Healthy: μ=0.311 σ=0.068	Healthy: μ=0.351 σ=0.133	**Men – genuine**	Healthy: μ=0.375 σ=0.060	Healthy: μ=0.316 σ=0.073	Healthy: μ=0.356 σ=0.125
	Diabetic: μ=0.397 σ=0.051	Diabetic: μ=0.3427 σ=0.059	Diabetic: μ=0.415 σ=0.115		Diabetic: μ=0.393 σ=0.056	Diabetic: μ=0.341 σ=0.061	Diabetic: μ=0.409 σ=0.121
	Distance: 0.0695	Distance: 0.0350	Distance: 0.0381		Distance: 0.0133	Distance: 0.0247	Distance: 0.0234
Women – impostors	Healthy: μ=0.459 σ=0.008	Healthy: μ=0.410 σ=0.031	Healthy: μ=0.499 σ=0.0243	**Men - impostors**	Healthy: μ=0.459 σ=0.008	Healthy: μ=0.410 σ=0.033	Healthy: μ=0.499 σ=0.0235
	Diabetic: μ=0.456 σ=0.009	Diabetic: μ=0.407 σ=0.031	Diabetic: μ=0.499 σ=0.0241		Diabetic: μ=0.460 σ=0.008	Diabetic: μ=0.413 σ=0.031	Diabetic: μ=0.4989 σ=0.0232
	Distance:	Distance:	Distance:		Distance:	Distance:	Distance:

	0.0175	9.23e-04	5.58e-05		0.0016	0.0019	1.386e-04

Table 5-1. Biometrics Statistics – the results are better for healthy people regardless of gender.

Further to what was mentioned above, a higher Area Under the Curve (AUC) means better performance evaluation results. Table 2 shows the AUC resulting from each group in the above four tests, using the original all vs all protocol, respectively.

According to the results presented in Table 5-2, it can be concluded that the results are better for healthy female users. We can also conclude that the biometric accuracy of Iris recognition for female users will be reduced more by the presence of type II diabetes.

Health Condition	Gender	Are Under Curve		
Diabetic	Female	0.7583 (LG)	0.8645 (CR)	0.9352 (DCTC)
	Male	0.7733 (LG)	0.8894 (CR)	0.9625 (DCTC)
Healthy	Female	0.8669 (LG)	0.9283 (CR)	0.9898 (DCTC)
	Male	0.8769 (LG)	0.9040 (CR)	0.9753 (DCTC)

Table 5-2. Area Under the Curve, the results are better for healthy people regardless of gender.

We conducted the same test with three different iris recognition systems to make sure they were testing the eyes, and not the quality of the algorithms used. In each case, the results were the same. All three systems had an easier time identifying healthy irides, but were less accurate when scanning the eyes of women with diabetes.

As shown in Figures 5-2 to 5-4 (for women and for men), the empirical cumulative distribution functions of genuine scores obtained by a- DCTC, b- CR and c- LG methods show statistically significant differences between the comparison scores obtained by comparing samples of diabetic irides and iris pattern images taken from healthy irides (for both gender groups). This claim that biometric user identification tends to be more difficult under the influence of diabetes and the accuracy of an iris recognition system for healthy irides is higher.

Moreover, our results show that the effect of diabetes on the accuracy of iris recognition system is higher for female users. If the gallery and probe images are taken from healthy eyes, the all three recognition systems yield the best performance. Yet, for identification of users under the influence of diabetes, the reduction in performance is observed. In Fig. 5-5 to Fig 5.7, (for women and for men), we present the empirical cumulative distribution functions of impostor scores for comparison scenarios.

As depicted in Figures 5-2 to 5-7, the biometric results for healthy eyes are better than the corresponding results for eyes affected by diabetes for all three iris recognition algorithms (DCTC, CR and LG: for women and for men). This may result from the fact that there are some non-obvious disorders in diabetic eyes. **In summary, iris recognition is less effective for people with type II diabetes, regardless of their gender.**

In this section, we have demonstrated that social factors such as diabetes can degrade the performance of mobile contactless biometric recognition systems. To tackle the detected problem, we need to build sophisticated diabetes detection algorithms that can improve biometric recognition.

In Support Vector machines, the basic idea behind dealing with a non-linear case is to map the input data into a higher dimensional feature space H via a kernel function.
The feature space is derived using the kernel function, instead of being strictly defined. In this way, the selection of the kernel is the key to determine the feature space. I have chosen the Gaussian radial basis function (RBF):

$$k(x, x^T) = e^{-\gamma \|x - x^T\|^2} \tag{5-1}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations (True Positives (TP); True Negatives (TN); False Positives (FP); and False Negatives (FN).

$$Precision = \frac{TP}{TP + FP} \tag{5-2}$$

Recall is the ratio of correctly predicted positive observations to all the observations in the actual class.

$$Recall = \frac{TP}{TP + FN} \tag{5-3}$$

The F1 Score is the weighted average of Precision and Recall. In order to train our model, we used 1,400 samples, with 506 iris images being used to test the model. Table 5-3 shows the statistics of the results achieved (in order to obtain the feature vectors, the DCT code was implemented).

	precision	Recall	F1-score	support
Diabetes	0.53	0.50	0.51	185
Healthy	0.72	0.75	0.73	321
Avg. / total	0.63	0.62	0.65	506

Table 5-3. Classification Results – Diabetes

We did not find diabetes diagnosis possible. Samant and Agarwal [151] have proposed a diagnostic tool for the discrimination of healthy and diabetic eyes. They claimed the best classification accuracy of 89.63%. They tried to validate the use of iridology. **However, the author of this thesis does not believe in the iridology method**.

5.7. Conclusions

We present a performance analysis of the iris recognition system for healthy irides and those affected by diabetes, separately for female and male users. To exclude gross segmentation errors, we inspected the iris segmentation results manually, through a visual examination of the samples one by one. Some of the irides happened to be segmented incorrectly or incompletely, so we used a different method of iris image segmentation, called Contrast-Adjusted Hough Transform Segmentation Algorithm, for the remainder of the samples. The Bhattacharyya distance was used to compare irides of diabetic and healthy men and women. We found out that the performance of

the system is higher for female users, but that for people with diabetes the system performs better for male users. After achieving the scores for both genders and for both diabetic and healthy groups, we used the Bhattacharyya distance formula in order to achieve the statistical distance between the cumulative graphs. According to the result shown in Table 5-2, we can conclude that it is harder to recognize people who are suffered from diabetes regardless to their gender, according to some non-obvious disorders in their iris textures. According to our results, for all three matchers, without considering the effect of diabetes, the accuracy of the iris recognition system is higher for female users. So the statement would be: "from a biometric science point of view, it is more probable for diabetic irides to be mistakenly rejected by the iris recognition system and the reliability of the iris recognition system under the influence of diabetes tends to be higher for male patients."

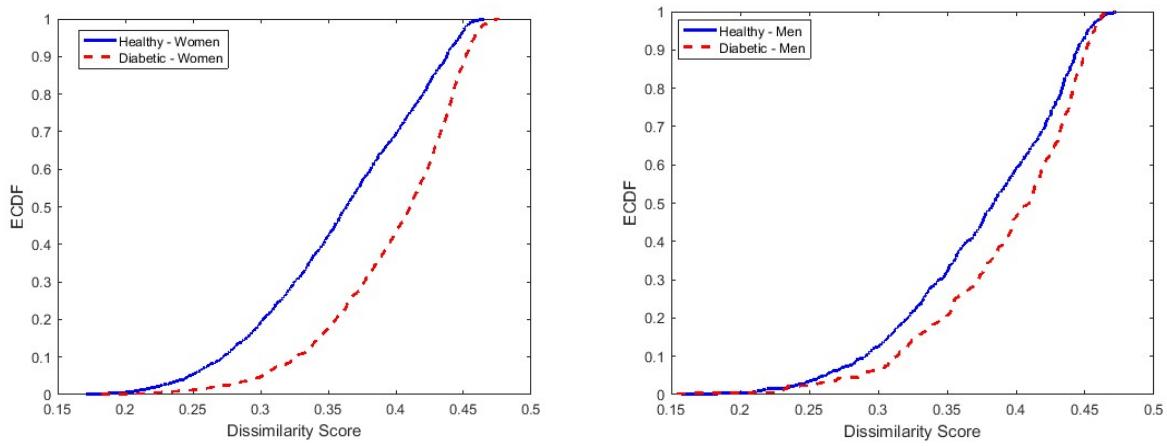


Fig. 5-2. ECDF of genuine scores obtained by the DCT algorithm – left: Women, right: Men. The figure shows that it is more probable for diabetic irides to be mistakenly rejected by the iris recognition system. The results are worse for people with diabetes type II regardless of their gender.

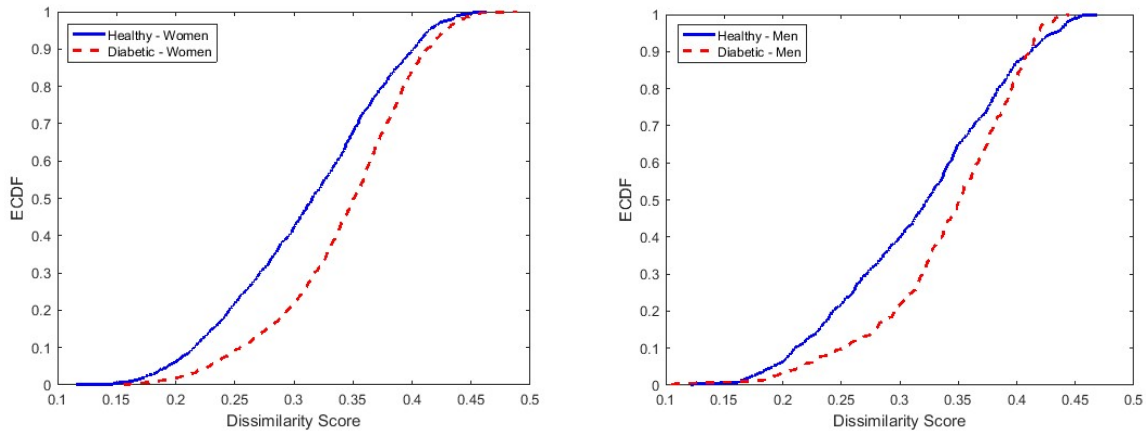


Fig. 5-3. ECDF of genuine scores obtained by the CR algorithm – left: Women, right: Men.

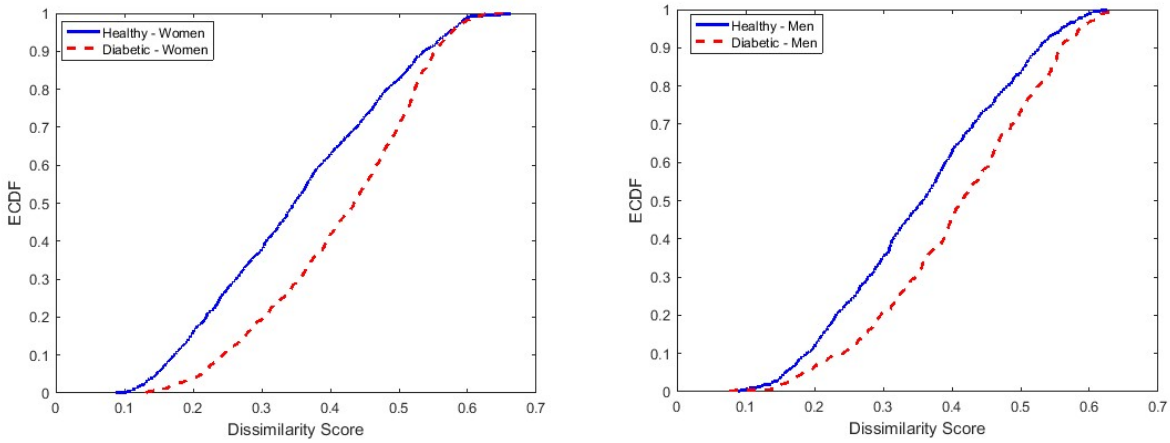


Fig. 5-4. ECDF of genuine scores obtained by the LG algorithm – left: Women, right: Men. The results indicate that iris recognition is less effective for people with diabetes type II regardless of their gender.

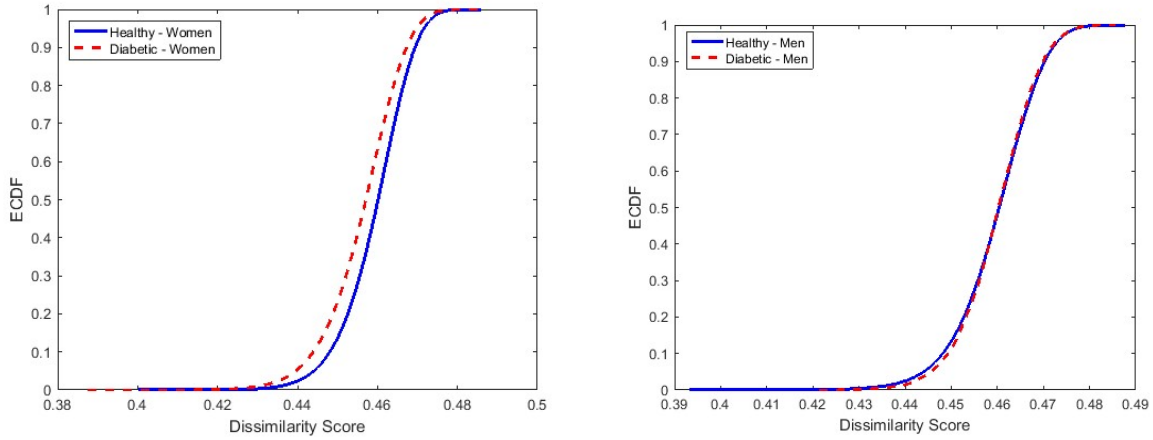


Fig. 5-5. ECDF of impostor scores obtained by the DCT algorithm – left: Women, right: Men.

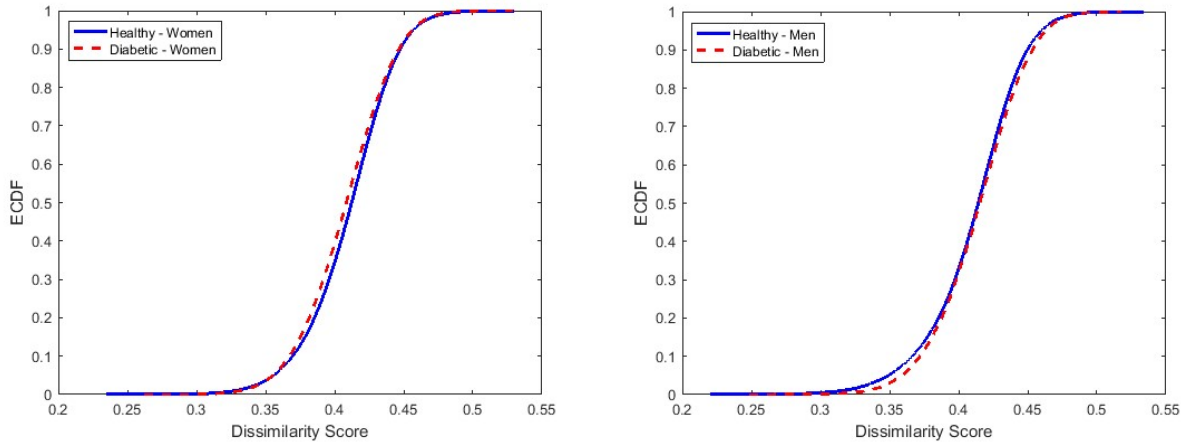


Fig. 5-6. ECDF of impostor scores obtained by the CR algorithm – left: Women, right: Men.

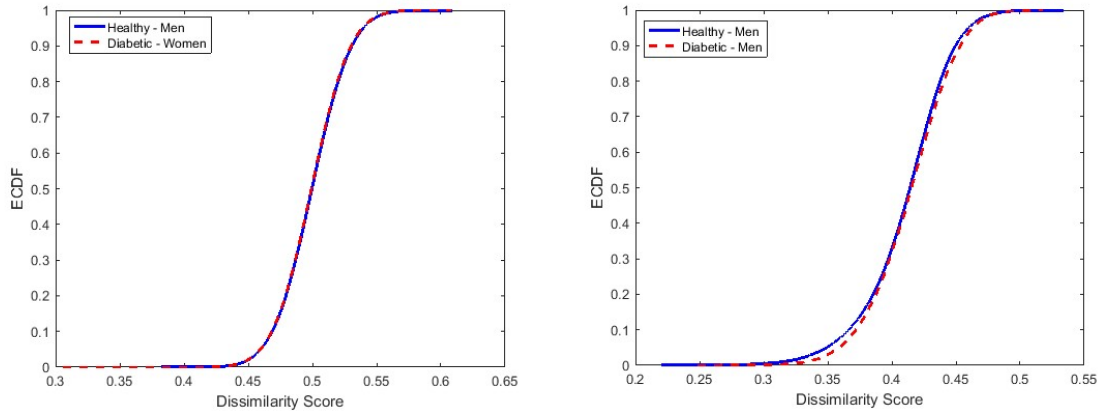


Fig. 5-7. ECDF of impostor scores obtained by the LG algorithm – left: Women, right: Men. The results indicate that, while impostor results are the same for healthy and illness-affected eyes, the genuine scores are better for healthy eyes.

6. Morning Voice

6.1. Overview

In the previous chapters, the effects of several biological factors (disease, physiological aging. and gender) on the performance of iris recognition systems were studied. We concluded by proving sub-statement No 1 (S1). This chapter presents the investigation on the effect of time of day on the matching accuracy of voice recognition systems (related to S2). In this part of the thesis, the effects of morning voice on voice biometrics as a contactless biometric modality, will be discussed. Morning voice can degrade the performance of the system due to biological and behavioral differences in the biometric data, from a morning acquisition to an evening one

In the early morning, the fundamental frequency of the voice is lower, as breathing through the mouth while asleep will dry out the vocal cords. On the other hand, in the early morning the person feels fresh, while during the day they will be more fatigued. This difference can mean that a person's voice in the morning sounds different to the voice they go to sleep with, both biologically and behaviorally. To the best of our knowledge, no analysis of this phenomenon has ever been presented or reported in published papers.

A new database was collected and offered. The database contains a dataset of thirty people. We collected 1,780 voice samples. There were two different data collection sessions: a- participants were asked to record their voice after getting up, using their own smartphone devices (916 morning voice samples were recorded), and b- participants were asked to record their voice samples during the day (884 samples were collected from the same users). Each sample is six seconds long at a bit rate of 705 kbps. All of the users are native speakers of Persian. In order to conduct the numerical experiments, a pre-trained VGG-Speaker was used. An all vs all comparison scenario was done. The intrasession comparison scores are better in comparison with intersession ones. For the evening versus evening comparison scenario, an EER of 1.46% was achieved. For the morning versus evening comparison scenario, the equal error rate increased to 10.2%.

6.2. Introduction

The voice we go to sleep with can be significantly different from the one we wake up with. Although a deep voice in the morning – a heavier voice after getting up in the early morning – is an abnormal change in voice, it should not be confused with hoarseness, which is generally caused by an inflamed larynx [168].

The throats tissues collect fluid during sleep. During the night, when people are asleep, the lack of use of the vocal cords causes mucus to build up. In addition, as most people breathe through their mouth while sleeping, this also has the effect of drying out the vocal cords. The vocal cords will be hindered from moving together without this lubrication during sleep. As a result, we can expect the lower pitch of voice in the early morning.

Voice is one the most distinguishable biometrics factors that can be used for human identification purpose. Voice samples can easily be obtained using smartphone microphones, meaning that voice recognition systems have become one of the most popular mobile biometric systems for cell phones. Today's smartphones are equipped with biometric tech such as voice. The only problem with this biometric solution is its lack of performance. However, the reliability of speaker recognition systems can be affected by various social factors.

Most people feel fresh after a night's sleep in the morning, while they would expect to be tired in the evening after spending a working day. Moreover, while a person is sleeping, the vocal cords will become dried out. This causes the characteristic lower-pitch sound of morning voice.

6.3. Related Works

One of the factors that has an impact on the system's performance is the time lapse between enrolling and later use. As the biometric data changes with the passage of time, this variation leads to a reduction in the performance of the biometric recognition system. This performance degradation is also termed "template aging".

Kelly and Hansen [62] investigated the effect of short- (between two months and three years) and long-term aging (up to 30 years) on the reliability of speaker recognition systems. They

reported that relative reductions in the log-likelihood ratio cost of 1-4% and 10-43% are obtained at short- and long-term intervals, respectively.

A very short time lapse between enrolling and use can also affect the performance of the speaker recognition system. **Even if we assume that the physiological features of speech would stay the same, the behavioral part of the voice definitely changes during the course of a day.**

In this chapter we will investigate the effect of morning voice on the matching accuracy of a text independent speaker recognition system.

6.4. New Database

The objective of this experiment is to analyze the variability of speaker recognition in adult users over the course of a day. For this purpose, we have collected a new database consisting of morning voice samples and evening voice samples all acquired with mobile devices. We also conducted several numerical experiments in order to have a better understanding of the system's performance under different conditions.

A new database was collected and offered containing 1,780 donated voice samples from 30 users in two scenarios: a. users were asked to donate voice samples after getting up in the early morning by using their own smartphone devices (916 morning voice samples) and b. the same users the collected voice samples from the evening (864 evening voice samples). Each sample is six seconds long at a bit rate of 705 kbps. All the users are native speakers of Persian. The other demographics were also asked at the beginning of the data collection, for future studies. Samples were recorded using the participants' smartphone devices and then the recorded samples were transferred to the author's telegram account by the participants.

The personal data are kept separately in order to guarantee additional security of the personal data. It is also important to note that all of the participants were fully aware of the experiment, full detailed information on the study was provided and all the participants signed consent forms. The experiment protocol was approved by the Ethics Committee of the Warsaw University of Technology.

6.5. Methodology

In order to obtain similarity scores between the voice samples, we used VGG speaker recognition methodology, which is a powerful speaker recognition deep network, using 'thinResNet' trunk architecture and a dictionary-based NetVLAD or GhostVLAD layer to aggregate features across time, which can be trained end-to-end.

The VGG speaker descriptor is generated by a Convolutional Neural Network, and has been proposed by researchers at Oxford University. The output of the penultimate layer (FC7) was used as the extracted feature vector of 512 elements. Then the Frobenius normalization was applied to vector spaces and the Euclidian distance between these normalized feature vectors was calculated. In this way the dissimilarity score would be the Euclidean distance between two extracted vectors, while a zero comparison score between identical pictures and a higher dissimilarity means a lower matching score between samples.

The model was trained end-to-end on the VoxCeleb2 [169] dataset (only on the 'dev' partition containing speech from 5,994 speakers) for identification and test on the VoxCeleb1 verification test sets [170, 171]. It is important to note that the development set of VoxCeleb2 is completely separate from the VoxCeleb1 dataset (i.e. no speakers in common).

The matcher may return symmetrical matching scores and the score between voice sample A and voice sample B are equal to a matching score between B and A).

6.6. Vocal Characteristics

The reported characteristics of voice are as follows:

6.6.1. Jitter:

The value of jitter can be measured in a number of parameters, such as absolute, relative, relative average perturbation and the period perturbation quotient.

Jitter absolute is the cycle-to-cycle variation of fundamental frequency, i.e. the average absolute difference between consecutive periods. In this work, however, first we detected the

speech activity, then we calculated the mean value of average jitter for each pulse of free speech. The average jitter value was calculated for each sustained vowel, then the average of all the gathered values was reported. The relative average or local jitter is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage (Eq. 6-1):

$$jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i-1}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \qquad (6\text{-}1)$$

Where Ti is the extracted glottal period lengths and N is the number of extracted glottal periods.

6.6.2. Shimmer:

The shimmer relative is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the amplitude, expressed as a percentage (Eq. 6-2):

$$shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i-1}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \qquad (6\text{-}2)$$

Where Ai is the extracted peak to peak amplitude data and N is the number of extracted glottal periods.

6.6.3. Harmonic to noise ratio:

The harmonic to noise ratio provides an indication of the overall periodicity of the voice signal, by quantifying the ratio between the periodic (harmonic part) and aperiodic (noise) components. This parameter is usually measured as overall characteristics of the signal and not as a function of frequency. Harmonic to noise can be calculated as a function of the autocorrelation coefficient:

$$HNR = 10\log_{10}\frac{AC_V(T)}{AC_V(0) - AC_V(T)} \qquad (6\text{-}3)$$

6.6.4. Fundamental Frequency:

In order to obtain the mean pitch (the fundamental frequency) of each sample, 10 seconds was split; the speech activity was detected first – by determining the pitch range between 75 to 500 hertz and using autocorrelation analysis method – then for each 10 ms of sample the F0 was obtained. At the next step, the average of the obtained F0 scores was calculated for each of the samples.

6.7. Results and Discussions

In this section, the obtained results will be presented through several figures. We will also discuss the achieved results in detail. In Figure 6-1, the plots depict the mean value of fundamental frequency of both evening and morning speech of every thirty users.

As can be seen in Figure 6-1, the fundamental frequency of speech is lower in the early morning. The user's voice pitch changes significantly during a day regardless of their gender. Harmonic to noise ratio, shimmer and jitter, are the cornerstones of acoustic voice measurement. According to Figure 6-1, the existence of meaningful differences in the value of the mean F0 parameter is clearly observable. The mean values of jitter, shimmer, and the harmonic to noise ratio of participants, are presented respectively in Figures 6-2 to 6-4.

According to Figures 6-2 to 6-4, we cannot make general statements regarding the difference between jitter, shimmer and the harmonic to noise ratio of the participants in the morning and in the evening.

It is worth mentioning that Figures 6-2 to 6-4 present the mean values for the same set of parameters for both morning (blue) and evening (red) voices.

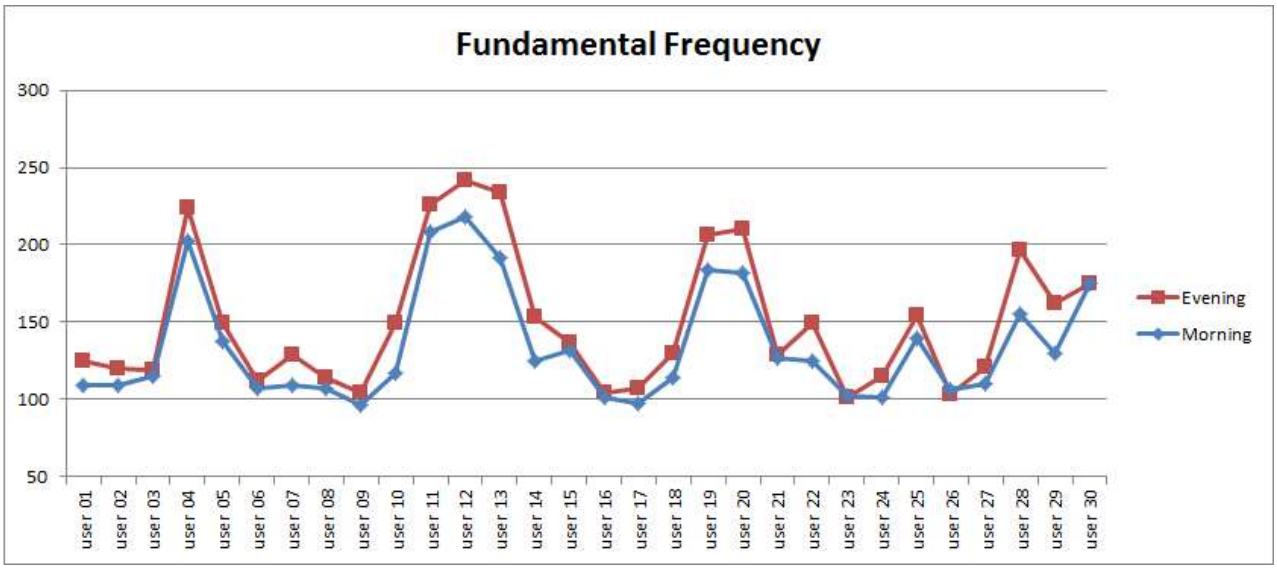


Fig. 6-1. Mean Values of Fundamental Frequency – it shows that F0 is lower in the morning.

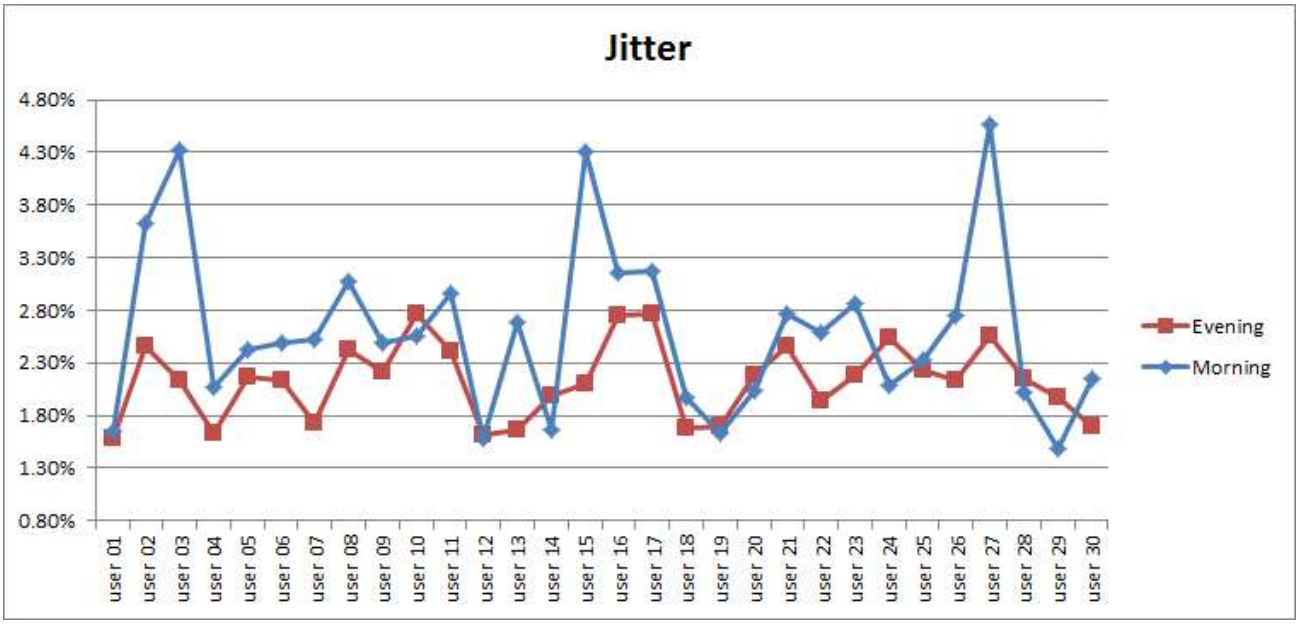


Fig. 6-2. Mean Values of Jitter – it shows that the jitter graph is nonlinear.

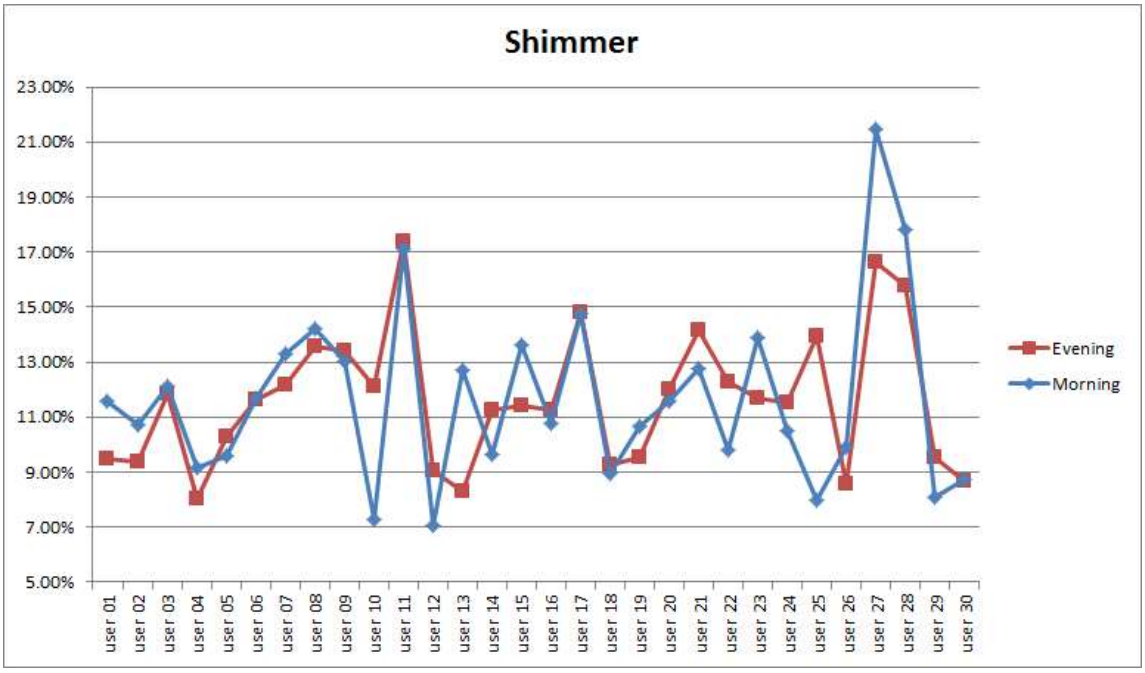


Fig. 6-3. Mean Values of Shimmer. It shows that shimmer in the morning is higher for some users and lower for the rest.

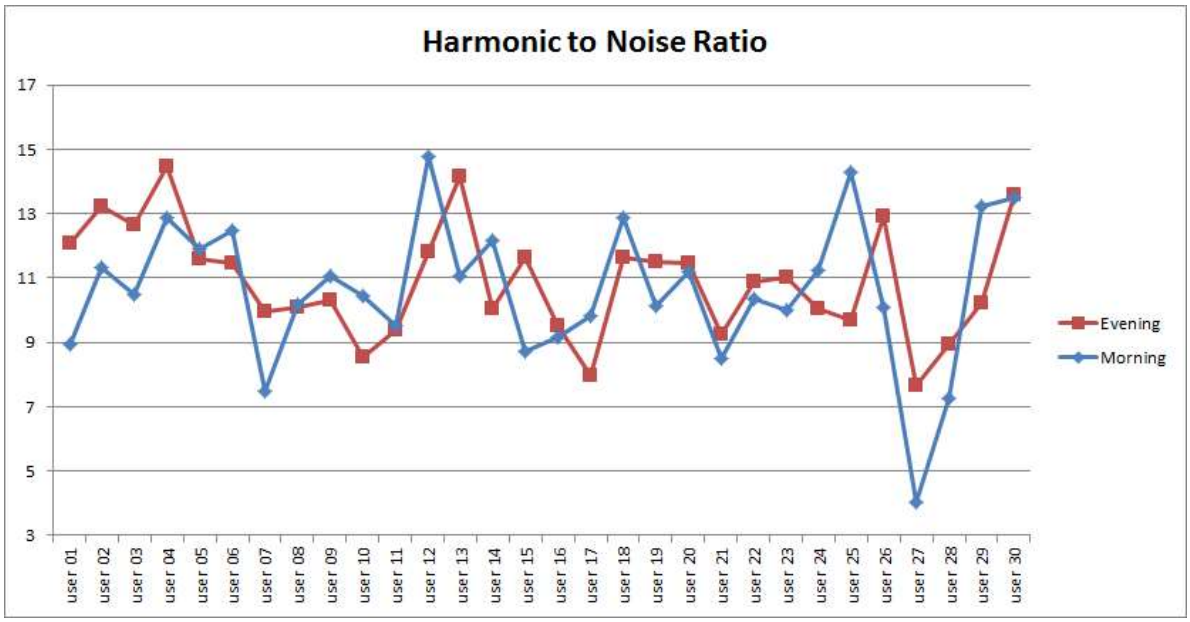


Fig. 6-4. Mean Values of the Harmonic to Noise Ratio. It shows that the HNR graph is nonlinear.

Figure 6-5 shows the empirical cumulative distribution function of similarity scores by comparing a- morning voice samples with morning voice samples (blue line), evening voice samples with evening voice samples (red line), and morning voice samples with evening voice samples (yellow line). According to Figure 6-5, the intraclass comparison results are better than the interclass comparison scores. This means that the "Morning voice" effects on the performance evaluation results of the speaker recognition system are significant.

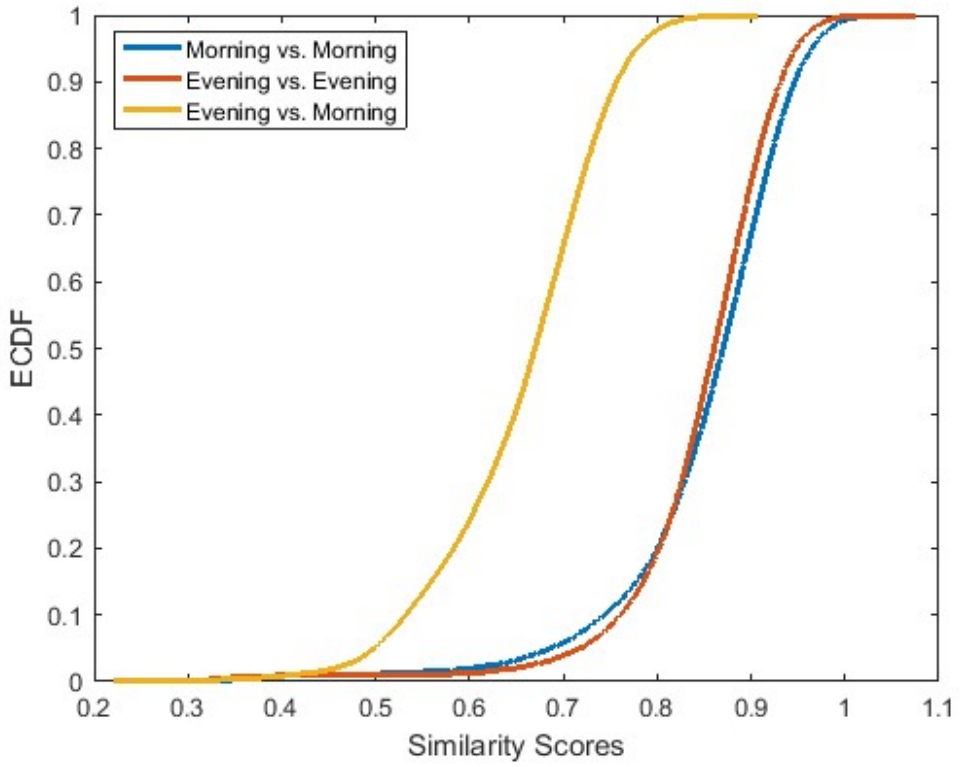


Fig.6-5. ECDF: Morning vs. Morning, Evening vs. Evening and Morning vs. Evening. According to this figure, intrasession results are much better than intersession ones.

Another important conclusion that Figure 6-5 helps draw is that the differences between the morning intrasession and the evening intrasession results are not meaningful. This can assure us about the strength of the chosen speaker recognition system. Hence, the time of the day itself cannot challenge the reliability of the system. As the biometric data changes over the course of a day, the systems' performance will change.

In Figure 6-6, three histograms for the mentioned comparison scenarios are presented. The same conclusion can be made according to the results presented in Figure 6-6.

It is important to note that Figures 6-5 and 6-6 only present the genuine scores.

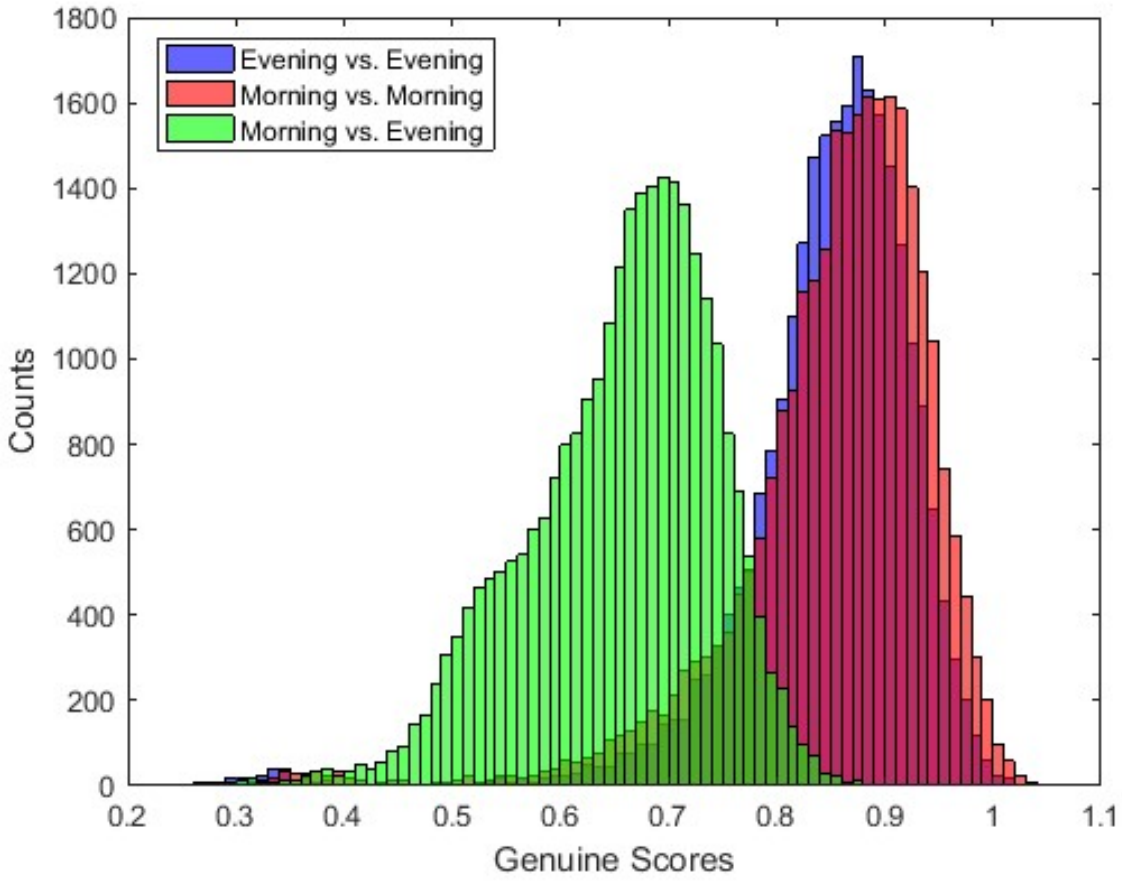


Fig. 6-6. Histograms: Morning vs. Morning, Evening vs. Evening and Morning vs. Evening. It shows that no matter whether it is evening or morning, the intrasession results are almost same.

To have a more comprehensive understanding of the results, Figure 6-7 a-c, present the histograms of both impostor and genuine scores separately. As Figure 6-7 a-c shows, there is no overlap between intraclass genuine scores and intraclass impostor scores, while the system has a harder time dealing with a user who enrolled in the early morning and is being identified in the evening.

Figure 6-8 depicts the ROC curves of all of the possible comparison scenarios. An equal error rate of up to 10 per cent was achieved for the morning vs. evening comparison scenario. According to Figure 6-8, the system performs well when both recognition events (enrollment and use) happen at same time of day. In other words, the comparison scores obtained by choosing

evening vs. evening or morning vs morning comparison scenarios are better than the interclass comparison scores.

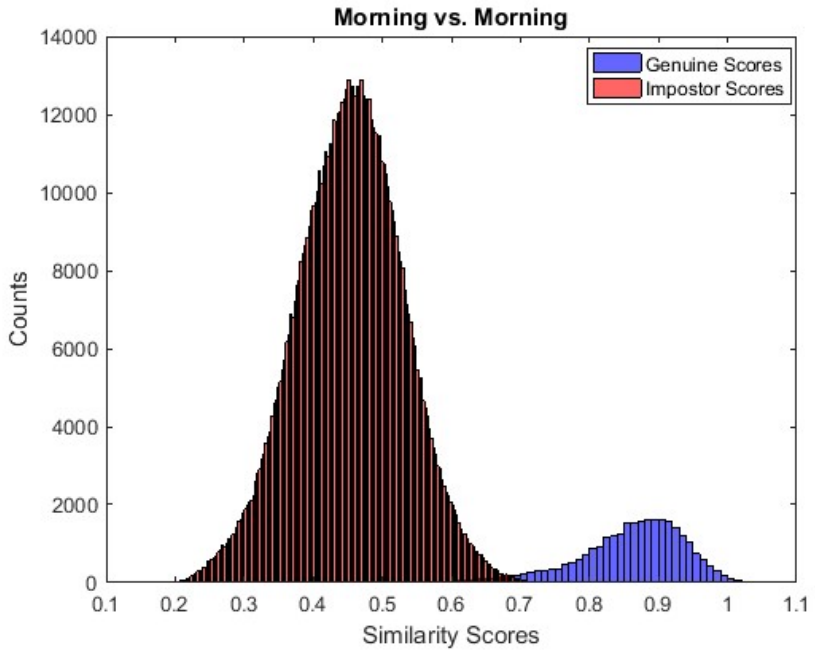


Fig. 6-7.a. Histogram: Morning vs. Morning. The histograms show that by determining the threshold = 0.7, the genuine and impostor users can be recognized.

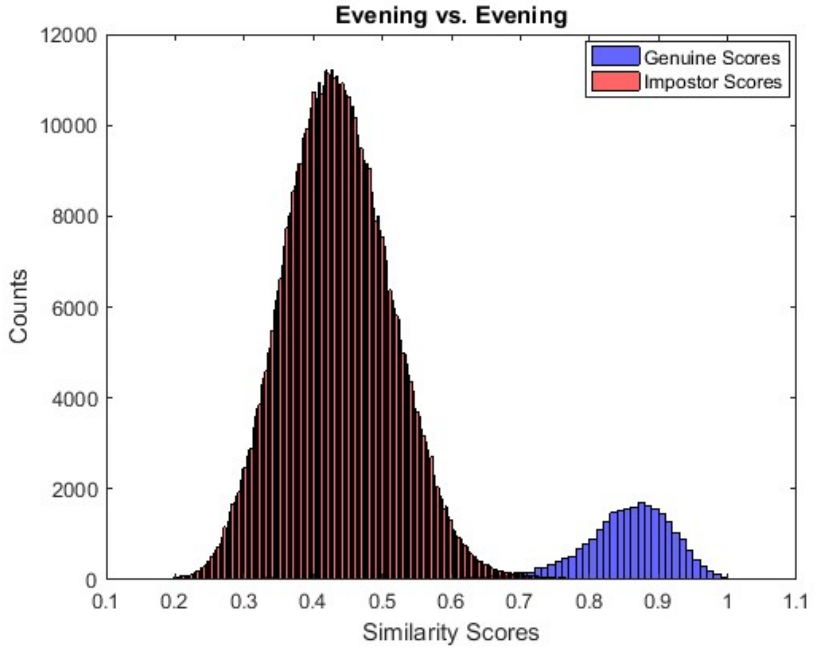


Fig. 6-7.b. Histogram: evening vs. evening. The histograms show that genuine and impostor results are discriminable for the evening vs evening scenario.

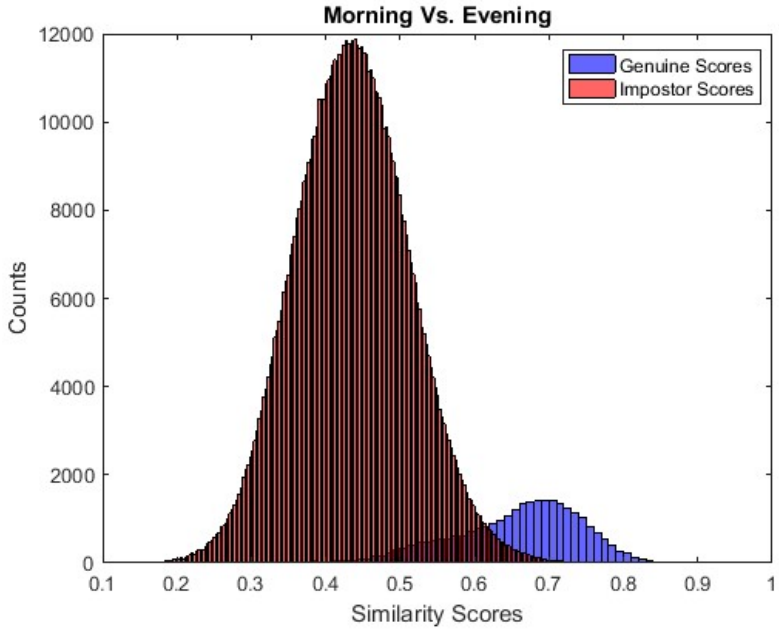


Fig. 6-7.c. Histogram: morning vs. evening. It shows that there is an overlap between genuine and impostor results for the intersession comparison scenario.

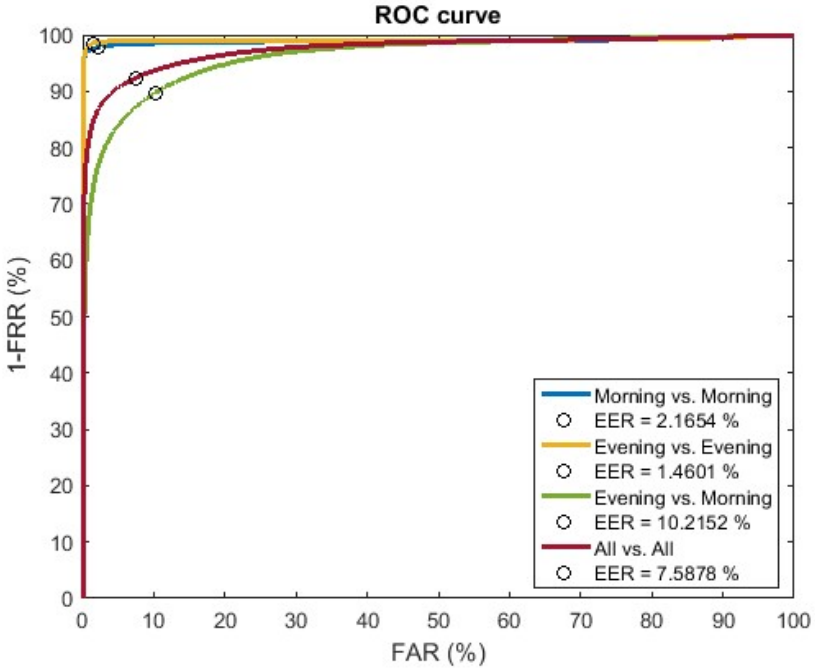


Fig. 6-8. ROC Curves: it presents the performance evaluation results of the system.

To tackle the detected problem, we need to build a sophisticated morning voice detection algorithm that can improve biometric recognition**.**

In order to train our model, we used 1,300 samples and to test the model we used 480 voice samples. Table 6-1 shows the statistics of the achieved results (to obtain the feature vectors, a VGG voice recognition code was implemented).

	precision	Recall	F1-score	support
Morning	0.96	0.91	0.94	263
Evening	0.90	0.96	0.93	217
Avg. / total	0.93	0.94	0.93	480

Table 6-1. Classification Results – Morning Voice

We have proposed a computer vision tool for detecting morning voice with an accuracy of 97%.

To be sure about the high matching accuracy of our chosen speaker recognition system, the Influence of Acted Mood Variation on the reliability of the same system (VGG-Voice model for speaker recognition) was investigated in Appendix B. The system performs well under optimal circumstances.

6.8. Conclusions

In this chapter, we have investigated the effect that the time of day can have on the reliability of a text independent speaker recognition system. For this purpose, a VGG speaker recognition system was used and we have also collected and offered a new database comprising 916 morning voice samples (voice samples were donated immediately after the participants got up in the early morning) and 864 voice samples collected from the same users in the evening. Each sample was six seconds long and was at a bit rate of 705 kbps. Interclass similarity scores were much lower than intraclass similarity scores. Hence, we can enhance the reliability of the speaker recognition system through the use of a support vector machine-based classifier.

7. Joint Influences of Make-up and Mood Variation on the Reliability of the Facial Recognition System

This chapter was previously published in Journal of Computers and Electrical Engineering in 2020, Authors: [M Azimi, A Pacut], Title: "Investigation into the Reliability of Facial Recognition Systems under the Simultaneous Influences of Mood Variation and Makeup"

7.1. Overview

In Chapters 3 to 5, in order to prove the first sub-statement (S1), iris recognition under the sole/combined influences of diabetes, physiological aging and gender (biological factors) was investigated. In Chapter 6, to prove sub-statement No. 2 (S2), the combined influences of behavioral and biological factors (morning voice) on the reliability of another contactless biometric system was investigated. To complete the puzzle, in this chapter we will show that the effects of facial expressions on the accuracy of the facial recognition task varies between subjects with full makeup and without full makeup. We will also prove that the effect of facial expression and full makeup are correlated. This chapter will be presented to prove the last sub-statement (S3).

The main aim of this chapter is to present an investigation into the influences of behavioral factor on the matching accuracy of the third chosen contactless biometric system:

Facial recognition systems are increasingly popular and prevalent in our everyday lives, especially on mobile cell phones. This paper is an attempt to investigate the effects that makeup and facial expressions have on the reliability of such systems. While these factors have been shown not to be significant by themselves, it has not been clearly demonstrated whether a combination of both these factors can affect the matching accuracy of the same system in a statistically meaningful way. In this chapter, we carried out numerical experiments through two databases: the Radboud Faces database and the Psychological Image Collection at Stirling (PICS), using a state of the art algorithm, namely dlib. Then, in order to be able to reliably validate the results, we used two more algorithms (Verilook and VGGFace) to give similarity scores.

The results showed that, while the effects of makeup and varied mood expressions are not significant by themselves, the joint effect is. An equal error rate (EER) of 4.68% was achieved when identifying faces under the joint influences of full makeup and mood variation, while the EER under the effect of each of these factors separately was less than 1%.

7.2. Introduction

Facial recognition is an interesting area in the field of biometrics and can be defined as identifying or verifying human subjects in various scenes from a still image or video source. Human beings can recognise and identify faces learned during their lifetime, even after a break of years. Thus, a major approach to face processing is to recognise faces at a level approaching the average human's capacity. Facial images can be easily recorded and stored using smartphone rear or front cameras, meaning that facial recognition systems have become one of the most popular mobile biometric systems for cell phones. However, the reliability of facial recognition systems can be affected by various social factors.

Within a person, facial recognition variability related to well-known social issues includes: aging, facial expression, makeup, and facial hair, as well as lighting conditions.

Among these social issues, facial expression depending on mood and differences in makeup are the most popular factors that have a substantial ability to change the appearance of a face significantly in different ways. During the daytime, people may experience considerable changes in mood. On the other hand, using makeup for beautification will continue to be an indispensable way of life for most people. In terms of mood, we can confidently assume that in most cases, users will upload their official photos to be enrolled to a biometrics system database. In such facial images, users tend to try to have neutral facial expressions in an attempt to look more serious. In real life, depending on the situation, a person's facial appearance may vary considerably, albeit temporarily. The reliability of a facial recognition system is influenced by the mood of the users. Due to the reasons mentioned above, looking at how the reliability of facial recognition systems can be enhanced to better deal with the simultaneous influences of makeup and mood variation is the major purpose of this study. In this paper, we discuss the simultaneous effects of makeup and facial expression on the reliability of facial recognition

systems in order to document the need to tackle this problem. We investigate the effect of full makeup on the reliability of facial recognition systems under the influence of varied facial expressions. The main contributions of this manuscript are as follows:

a. Calculating similarity scores between samples with no makeup and samples with full makeup in order to investigate the effect of full makeup on expressive faces on the accuracy of recognition systems.
b. Calculating similarity scores between original images and samples with lipstick makeup.
c. Conducting statistical tests in order to find out which moods can make significant changes on facial images with makeup.
d. Finding ways to enhance the accuracy of systems by pre-processing the images (using the dlib code for the automatic application of makeup on the facial images).

Deep learning-based techniques have grown to be very popular, due to the flexible design of layers and the efficiency of results, particularly in the analysis of medical images, which requires precise computations. When the parameters (such as batch size) are chosen carefully, deep learning-based methods provide the most successful results. Therefore, deep neural network-based methods have also been applied in this part of our thesis.

7.3. Related Works

Although there are already a large number of papers looking at the effects of makeup on facial recognition, along with papers considering variations in facial recognition accuracy due to expression, it seems that there has not been any previous investigation into the effects of a combination of both expression and makeup, and there is no published paper to answer the following question:

"Does a combination of full makeup and facial expressions make a statistically significant change in the biometric results?"

At this point, it is important to note the purpose of the systems under consideration. Some systems were created to identify people, while others exist for the purpose of recognising emotions, and in that case sensitivity to emotion is not a disadvantage. In order to make the mood/expression identification more accurate, the Facial Action Units (FAU) set was defined to

score facial muscle movements. The influence of a user's emotional state in facial recognition scores has been the subject of several studies [172, 173].

In order to answer the question as to whether facial makeup affects the accuracy of facial recognition systems, Dancheva et al. [113] collected two different databases from pictures of individuals' faces, before and after applying makeup. In a paper written by Banjeree and Das [174], the SCNN model was designed, using loss functions to deal with the variations due to makeup. According to the reported results, SCNN provides a considerable improvement in GAR (4% at 0.1% FAR) and EER (∼10–42%) values in all cases, even for low FAR values. The advantage of this work is that the transfer of cosmetic variations to a face from that of another individual has been performed with no manual intervention. According to previously published articles [175, 176], the impact due to the application of eye makeup is more acute than from lipstick makeup.

This chapter presents the results of numerical experiments.

We want to find answers to three specific questions:

1- Does the influence of facial expressions on the accuracy of the facial recognition task vary between before and after full makeup?
2- For which mood classes does the application of lipstick makeup combined with facial expressions change the results in a statistically significant way?
3- Which facial mood images are most dissimilar to non-expressive pictures of the same user?

To support the experiment samples were taken from the same users under normal conditions and under the influence of emotions. The trend of genuine scores is studied in detail in this chapter.

7.4. Pre-Existed Databases

In this study we used the Radboud Faces Database (RaFD) [177] and the Psychological Image Collection at Stirling (PICS) [178] (Fig. 7-1). The first database is a set of pictures of 66 models (41 male users and 25 female ones) displaying seven emotional expressions: Anger, disgust, fear, happiness, sadness, surprise, and neutral. The second database contains facial images donated by 13 women and 10 male users displaying same emotional expressions + pain (simulated). In fact, there are many face expression databases, some of which are larger than the two used for this

paper. The two we selected to use are also not the most commonly databases used by previous researchers, but as the samples were provided in highly restricted conditions, it is possible for us to be purely focused on just the simultaneous effects of make-up and facial expressions, and to consider the effects of other potentially influential parameters (illumination, pose, etc.) on the quality of the face verifications, as negligible.

To answer two other questions, and in order to apply virtual light makeup and full makeup on the facial images, we used the standalone 30-day trial version of CyberLink MakeupDirector 2 (the software works alone, without any need to connect to the internet, thereby ensuring that the samples would not be uploaded or transferred). Using this software, makeup must be applied to the facial images individually and manually. However, the quality of makeup is very high and at the first glance it is hard to discriminate the virtual makeup samples from the real ones.

Fig. 7-1 An example of facial images in different moods, (PICS) [178]

It is also possible to use dlib in order to apply virtual makeup. Fig.7-2 shows an original facial image expressing sadness (upper left), the same sample with full makeup made by CyberLink MakeupDirector 2 software (upper right), the same sample with lipstick makeup made by same software (lower left), and the same sample with automatic makeup applied and processed using the automatic dlib code (lower right).

The differences between the CyberLink MakeupDirector 2 software and the dlib automatic program are as follows:

1. The samples obtained by the application of full make-up using the professional software are extremely similar to the real samples.
2. The software works manually (the samples must be processed individually and it is possible for the user to change the style of full makeup), but everything can be automatically done using the dlib code.

The maximum possible number of comparison scores will be achieved between neutral facial images, and expressive facial images (the number of neutral facial images times the number of expressive facial images).

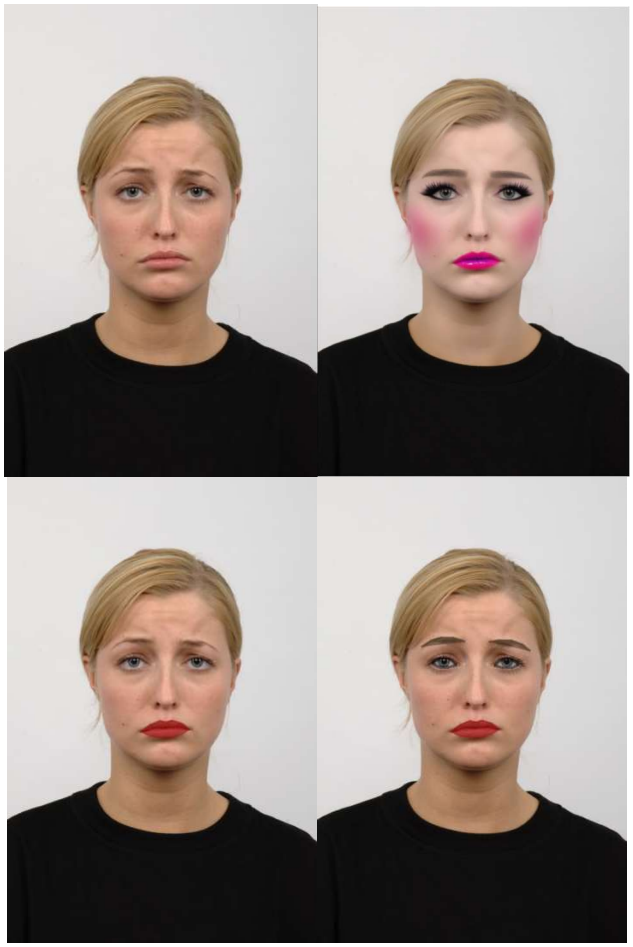


Fig .7-2 The original sample (upper left), the same person with full makeup (upper right), the same person with lipstick (lower left), and the same person with auto makeup (lower right) [177].

When answering the questions concerning makeup, we compared the neutral facial images without makeup with the expressive face images before and after applying virtual lipstick, and then we analysed the results.

As the makeup market is usually targeted towards women, and as we are trying to investigate the social habituation effect on the biometric results, we have only used images donated by female participants. It is also very important to note that, for the purpose of carrying out the numerical experiment regarding the last question (which is the most important question of this study), we have assumed that most women apply full make-up – eyeliner, mascara, eye shadow with lipstick and wearing foundation. Hence, the samples after and before the application of full makeup were used to find the answer to the last question.

7.5. Methodology

To obtain the comparison results, the dlib matcher was used:

- Dlib [179]: Convolutional neural network-based method – the facial recognition dlib gives the facial distance (the Euclidean distance between two 128-dimensional vector spaces) as a result of comparison. The face distance between pictures represents the dissimilarity between samples as a number in the interval [0, 1]. For two identical pictures, the face distance is zero, hence a lower face distance means a higher similarity between pictures. This model is a ResNet network with 29 conv layers [180]. It is a version of the previously published ResNet-34 network [181], with a few layers removed and the number of filters per layer reduced by half. The network was then trained by using a facial image database containing three million facial images of 7,485 subjects. The weighting of the network was initialised to completely random numbers. The network was trained with a structured metric loss that tries to project all the identities into non-overlapping balls with a radius of 0.6 [180].

  Dlib can also distort the chosen image (by flipping, rotating, translating and zooming) randomly in order to take the encoding of each version of the image and return the average. Hence, dlib is capable of making more accurate facial identification.

  The dlib model obtains an accuracy of 99.38% on the standard Labeled Faces in the Wild facial recognition benchmark, while the accuracy of VGG-Face is 99.13% for the same

database. The dlib state-of-the-art facial recognition model in android is also very accurate, but it takes approximately six seconds on a Redmi4 with Snapdragon 435 and Android 7.1 [182].

In order to validate the results to more matches, the following were also used:

- VeriLook neurotechnology Software [183]: VeriLook is a commercial matcher with an unpublished coding methodology. The matching score corresponds to the similarity of images; as a result, a higher comparison score denotes a better match. VeriLook uses a set of robust image processing algorithms, including deep neural networks.
  However, a disadvantage of using Verilook is that, because it is closed-source software, the user will not be able to extract and store the features separately for further classification purposes.
- VGG-Face [184]: The VGG-Face descriptor is generated by a Convolutional Neural Network, and has been proposed by researchers at Oxford University. We used the Keras model of VGG-Face. The VGG-face model was pre-trained with about 2.6 million facial images of 2,622 different individuals. The VGG-face model consists of 13 convolutional layers and five pooling layers in combination with three fully connected layers. The size of the feature map is 224×224×64 in the first convolutional layer.
  To achieve a feature vector, the classifier layer is removed and the output of the penultimate layer (FC7) was used as the extracted feature vector of 4096 elements.

The Frobenius normalisation was applied to vector spaces and the Euclidian distance between these normalised feature vectors was calculated. So the dissimilarity score would be the Euclidean distance between two extracted vectors while the zero is the comparison score between identical pictures and a higher dissimilarity means a lower matching score between samples.

All of the matchers may return symmetrical matching scores, and the score between image A and image B are equal to the matching score between A and B. For resizing and facial detection we used OpenCv (in Python).

In summary, the experiment relies on selecting facial recognition systems (dlib) and two more matchers for the validation of the results (Verilook, and VGG-Face), as well as two databases of faces (Radbound Faces Database and PICS), presenting seven emotional expressions: anger,

disgust, fear, happiness, sadness, surprise and neutral. As the result of the comparison, the systems return matching scores corresponding to the similarity of the input images (VeriLook), or their dissimilarity understood as the distance between feature vectors in a space with a dimension equal to the number of features analysed in the particular method (N=128 dlib, and N=4096 for VGG-Face).

7.6. Statistical Analysis

To judge whether the observed differences in comparison scores across partitions can be considered as samples drawn from the same distribution or not, a two-sample Kolmogorov-Smirnov test was applied, with a significance level $\alpha = 0.05$. The null hypothesis H0 in this test states that the samples originating from the two compared sub-groups are drawn from the same distribution. Alternative hypotheses are (F is the similarity scores distribution for Verilook and the dissimilarity scores distribution for VGG-Face and dlib):

a- For VGG-Face and dlib:

H1: F_withoutmakeup (Neutral vs expressive) < F_withmakeup(Neutral vs expressive)

b- For VeriLook:

H1: F_withoutmakeup(Neutral vs expressive) > F_withmakeup(Neutral vs expressive)

As stated above, in order to answer the second question, we have just analysed the results after the application of light makeup (Lipstick).

In all the cases, two sample t-test and Kolmogorov Simonov test were applied.

When calculating the degree of dissimilarity between the probability distributions, in order to find out which mood can cause more problems for facial recognition systems, we used the Bhattacharyya distance formulation.

To answer the last question, we also analysed the results after “full” makeup, in order to clarify whether the difference between biometric scores before and after full makeup application is statistically significant.

By using the aforementioned methods and materials, we attempt to answer the following questions:

1. Does a combination of full makeup and facial expressions bring a statistically significant change in biometric results? 2. Which emotions can make the genuine scores from the application of lipstick makeup on facial images statistically significant? and 3. Which facial mood images are the most dissimilar to inexpressive pictures of the same user?

7.7. Results and Discussions

In this section we will present and analyse the obtained results.

- First Question (Does a combination of full makeup and facial expressions bring a statistically significant change in biometric results?)

In the results and discussions section, we are trying to find out whether or not the effect of a facial expression and full makeup are correlated.

For this purpose, as was explained in detail in the first part of Section 3 (Materials), we have used the free 30-day trial version of CyberLink MakeupDirector 2 software in order to apply full makeup on the expressive samples.

In order to answer this question, we have used dlib code (as we will use the same software to find a solution that will remedy the detected problem).

We selected three scenarios for finding a proper answer to the first question of this study:

a- Comparing the samples donated in a neutral condition with and without makeup (to investigate the effect of makeup solely),

b- Comparing the expressive facial images without makeup with neutral ones (in order to judge whether or not the application of makeup itself can challenge the reliability of the facial recognition system?), and finally

c- Using all vs. all comparison scenarios by comparing all of the samples with each other (to investigate the simultaneous effects of makeup and mood variation).

Fig. 7-3 presents the ROC curves achieved by choosing the mentioned comparison scenarios.

According to Fig.7-3, an EER of 4.68% will be observed when we want to identify people under the simultaneous influences of full makeup and mood variation. The chosen facial verification system (dlib) is robust against each of the influential factors separately. In fact, while the single effects of each are not significant, the joint effect is.

In Fig.7-4, the ROC curves for scenarios (various moods and with/without makeup) are presented. The effect of full makeup on the various expressive facial images is different and,

according to the presented results in Fig.7-4, **"Disgusted face images are the most sensitive face images to the application of full make-up."**

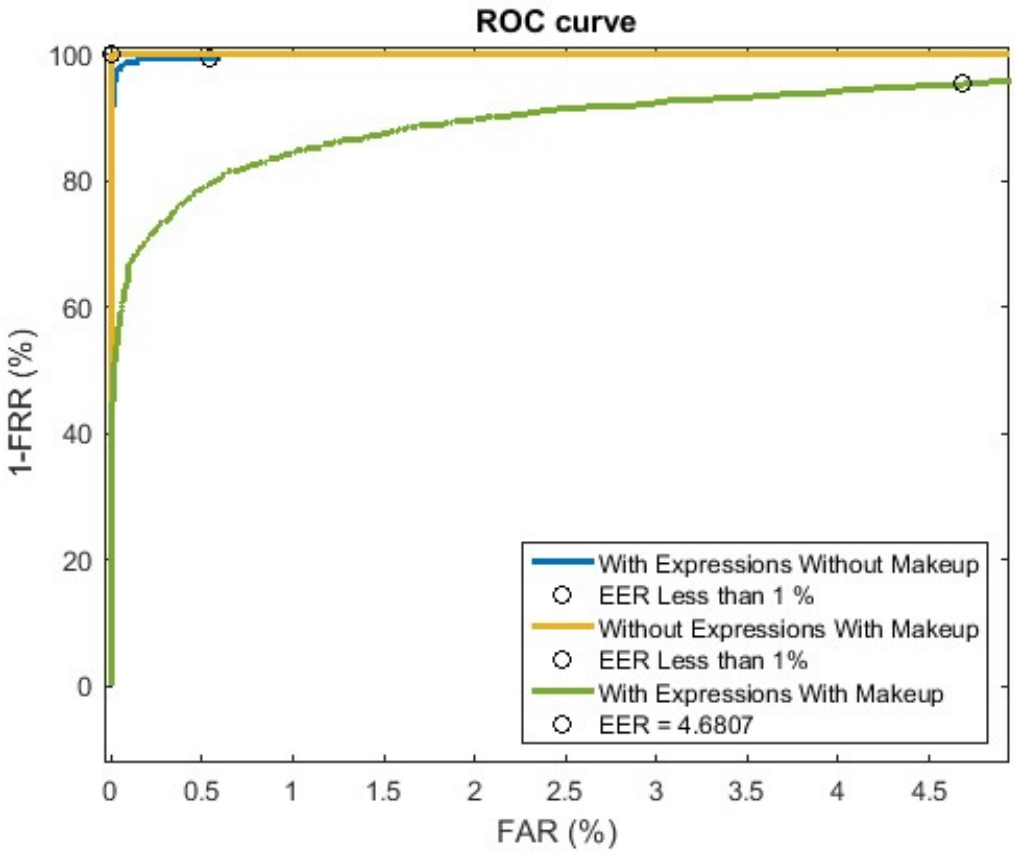


Fig.7-3. ROC curves: with expressions without makeup (Blue), without expressions with full makeup (Orange), and with expressions and full makeup (Green). This figure shows that while the effects of makeup and expression are not meaningful solely but the combined influence is.

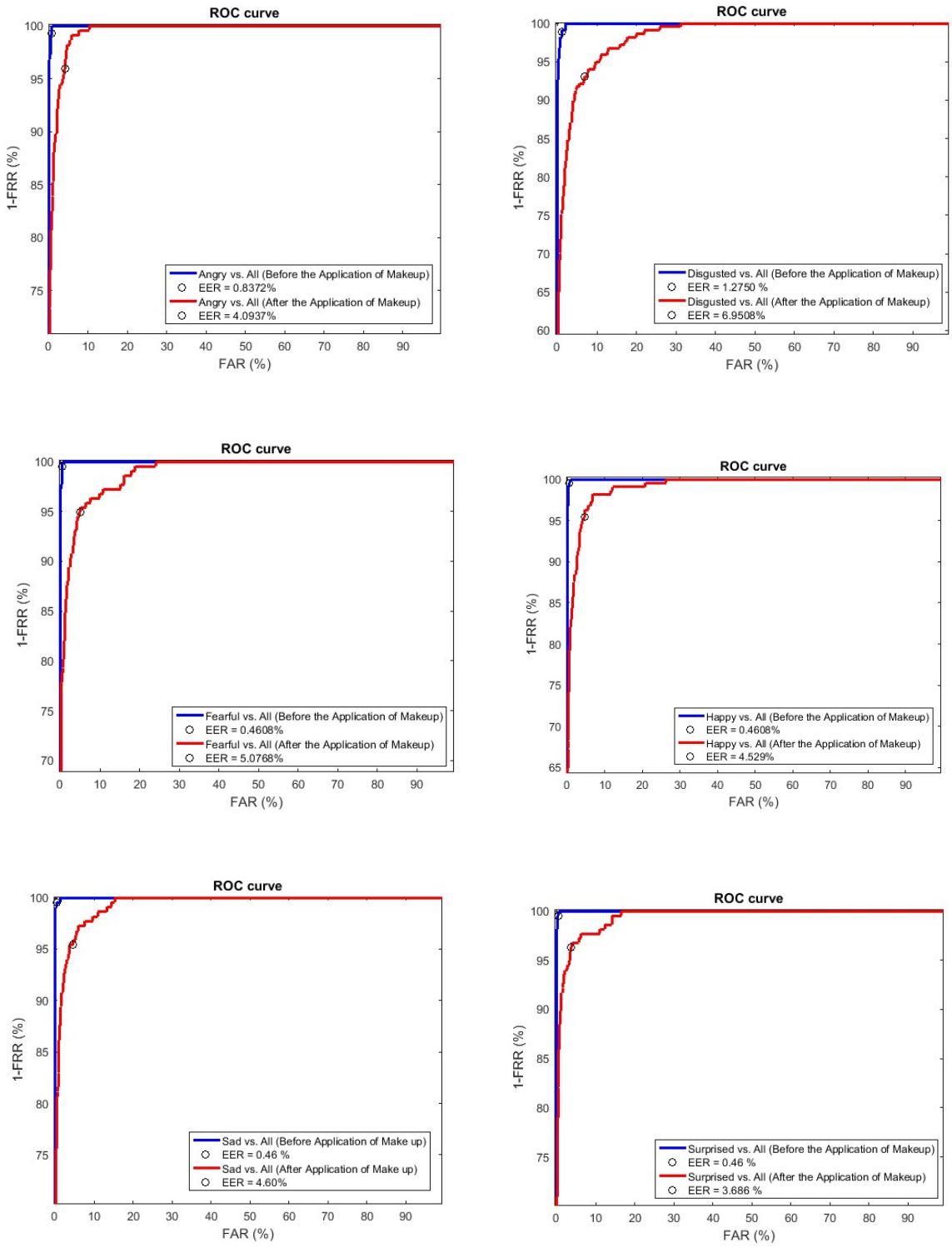


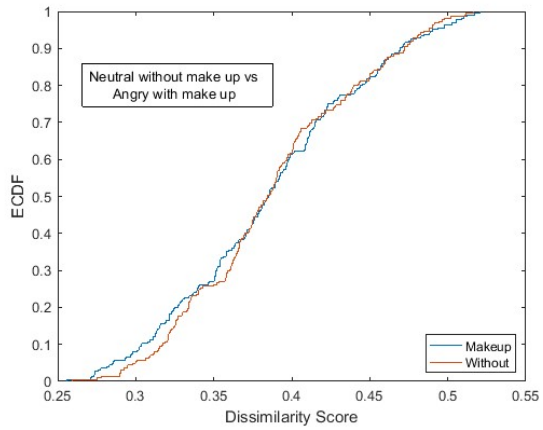
Fig. 7-4 ROC curves obtained by comparing : first row left- Angry, first row right - Disgusted, second row left - Fearful, second row right - Happy, third row left - Sad and third row right - Surprised facial images against the entire database for two cases: 1. With makeup and, 2. Without makeup. These figures show that regardless to the expression type, the joint effects on the performance is significant.

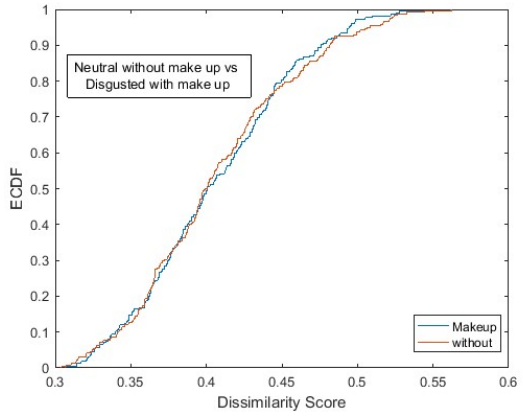
Noise reduction should be performed according to noise types in the images. Therefore, as an extension of this work, an appropriate noise reduction technique after detection of noise type can be added.

- Second Question (Which emotions can make the genuine scores from the application of lipstick makeup on facial images statistically significant?)

In Figs.7-5, the cumulative distributions of genuine scores for the application before and after light makeup (just lipstick) have been presented for various conditions. The results shown in Figs.7-5 were achieved using CNN methodology (dlib). Each of the figures contains two plots. The empirical cumulative distribution functions of comparison scores comparing neutral images without makeup with mood images with lipstick are presented.

We also presented the results achieved by two more matchers, namely the Verilook matcher (upper) and the VGG Face system (lower) in Fig. 7-6. The reason for presenting the results achieved by the other two matchers along with the main results obtained by using dlib code is that we wanted to validate the results. According to results presented in Figure 7-6, there was a strong agreement between the results achieved by dlib and the results obtained by using Verilook and VGG-Face. This gives some reassurance about the validity and high accuracy of our results.





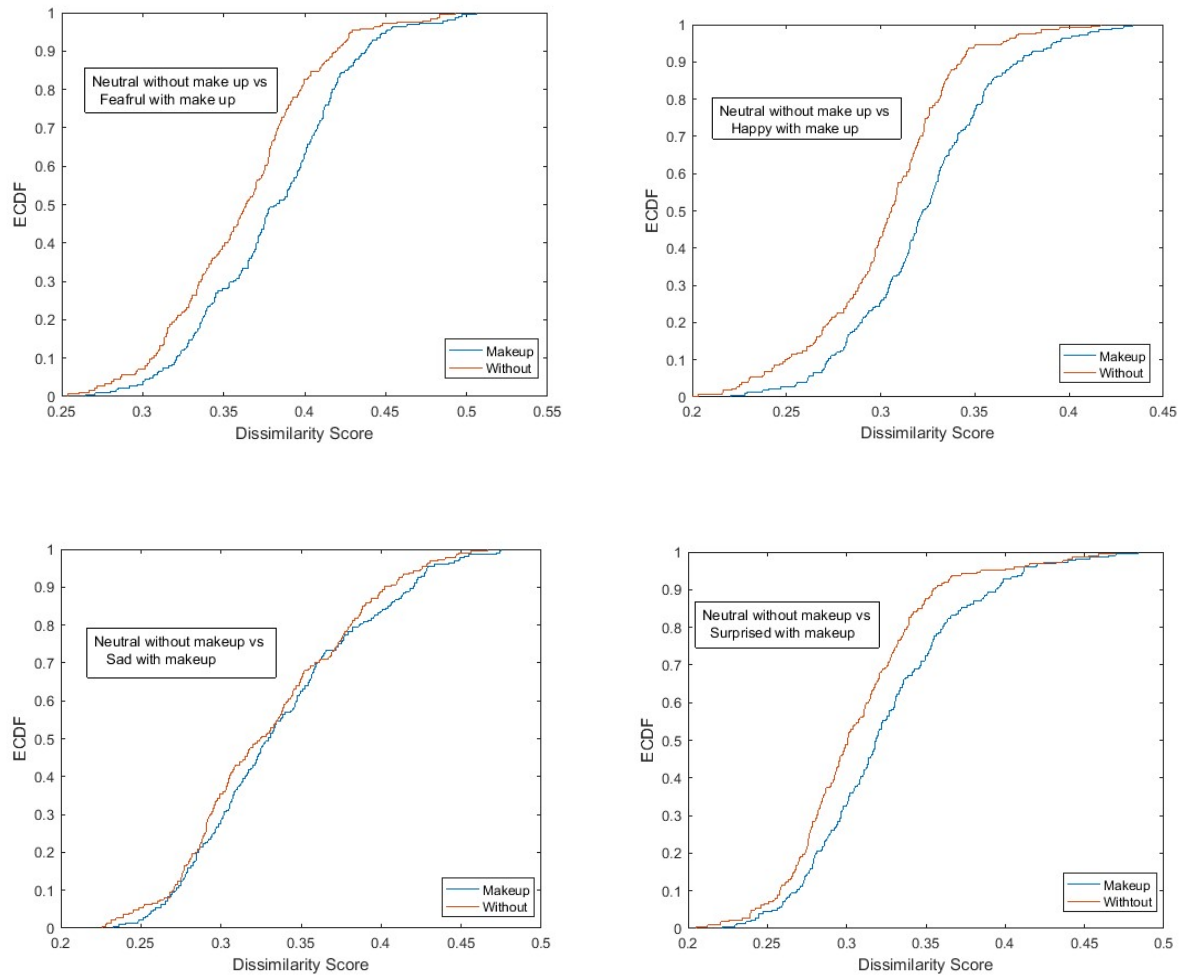


Fig. 7-5 Comparison between neutral and angry face images (no makeup, makeup) obtained by dlib : first row left- Angry, first row right - Disgusted, second row left - Fearful, second row right - Happy, third row left - Sad and third row right - Surprised facial images against the entire database. According to these figures for some of the expressions, the application of make-up has no effect on the similarity scores.

According to the results tabulated in Table.7-1, the application of lipstick cannot change the recognition accuracy rate of users with angry, disgusted and sad faces. For instance, for angry faces the lips would be tightened (Action Unit 23), the application of lipstick would not be significant for the recognition system, as the user would cover her lips. According to the obtained results from all three codes matcher used, it can also be concluded that Happy faces are the only faces that are sensitive to the application of lipstick.

Scenario→ Method↓	Angry vs Neutral	Disgusted vs Neutral	Fearful vs Neutral	Happy vs Neutral	Sad vs Neutral	Surprised vs Neutral
Dlib	K-S test p-value = 0.68	K-S test p-value = 0.86	K-S test p-value = 9.24 e-08	K-S test p-value = 2.25 e-09	K-S test p-value =0.3836	K-S test p-value = 0.0014
	T test p-value = 0.71	T test p-value = 0.89	T test p-value = 3.78 e-07	T test p-value = 1.01 e-11	T test p-value = 0.1972	T test p-value = 7.361e-05
Verilook	K-S test p-value = 0.91	K-S test p-value = 0.96	K-S test p-value = 0.0039	K-S test p-value = 0.01	K-S test p-value = 0.365	K-S test p-value = 0.11
	T test p-value = 0.69	T test p-value = 0.48	T test p-value = 2.13 e-4	T test p-value = 7.9 e-4	T test p-value = 0.2020	T test p-value = 0.38
VGG Face	K-S test p-value = 0.4274	K-S test p-value = 0.6997	K-S test p-value = 0.034	K-S test p-value = 0.0025	K-S test p-value = 0.1968	K-S test p-value = 1.975e-4
	T test p-value = 0.7937	T test p-value = 0.4515	T test p-value =0.072	T test p-value = 1.24 e-4	T test p-value = 0.2883	T test p-value = 1.631 e-6

Table. 7-1. Hypothesis test for the application of lipstick on expressive faces.

- Last question (Which facial mood images are the most dissimilar to unexpressive pictures of the same user?)

In Figs7-7 a-c, the cumulative distributions of genuine scores for neutral vs. all seven mood pictures are presented. We used the convolutional neural network dlib along with VeriLook Skd and VGG-Face to obtain the results presented in Figure 7-7.

Based on the plots, facial images displaying disgust tend to be the most dissimilar pictures to facial images without expression. As can be illustrated Figure 7-7, the same statement can be made for all three matchers.

Table 7-2 shows the Bhattacharyya distance calculated between the obtained distributions, as well as the mean value rate ($\frac{Mean(NN)}{Mean(N-\ Mood)}$ for the similarity rate obtained by VeriLook and $\frac{Mean(N-\ Mood)}{Mean(NN)}$ for the dissimilarity rate achieved by Python dlib and VGG-Face. Based on

the obtained results, the highest distance from the neutral vs neutral distribution belonged to the neutral vs disgusted sample. It can be concluded that **"Disgusted images are most dissimilar face images to non-expressive pictures of same user."**

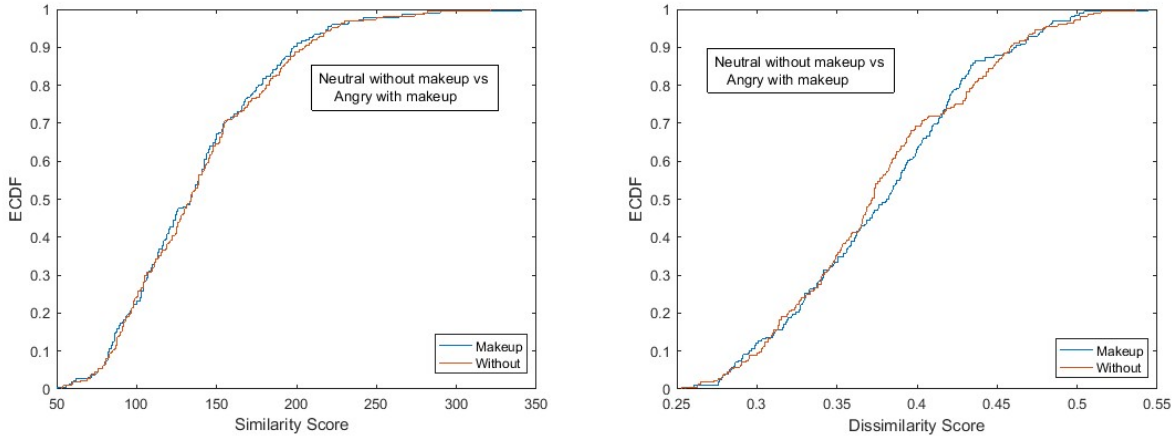
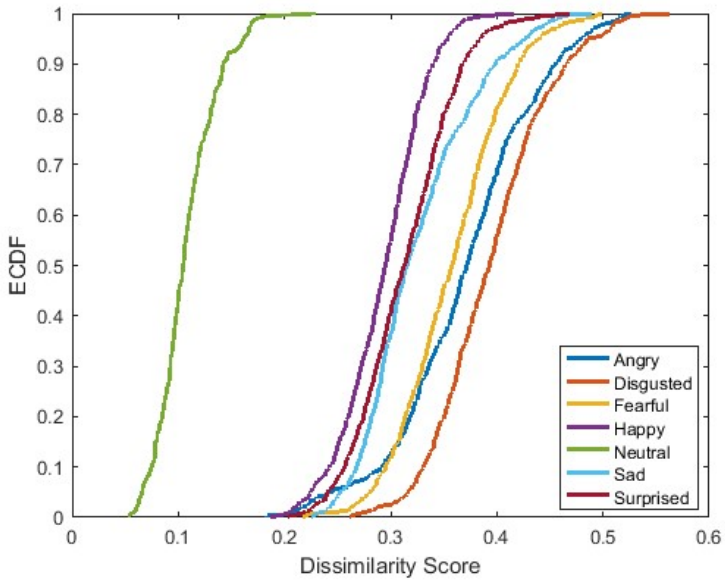


Fig. 7-6 Comparison between neutral and angry face images (no makeup, makeup) obtained by VeriLook (Left) and VGG-Face (Right).



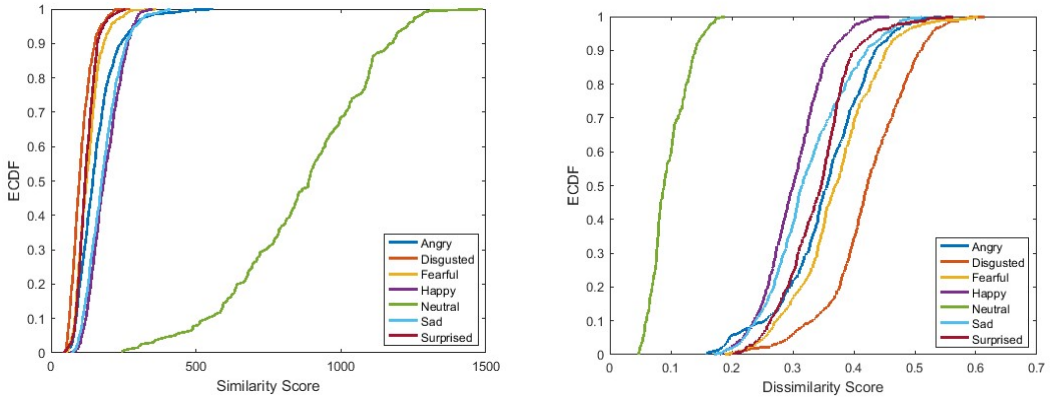


Fig. 7-7. Neutral vs Expressive facial images by Dlib (a: top), Verilook (b: bottom left) and VGG-Face (c: bottom right) – Disgusted images are most dissimilar face images to non-expressive pictures of same user

	NN, NA	NN, ND	NN, NF	NN, NH	NN, NSA	NN, NSU
Python Face recognition dlib	M= 3.4302 BD= 3.439	M= 3.66 BD= 5.851	M= 3.316 BD= 4.935	M= 2.714 BD= 3.811	M= 3.02 BD= 3.501	M= 2.904 BD= 3.769
Verilook	M= 4.6075 BD= 1.809	M= 6.955 BD= 2.57	M= 5.885 BD= 2.305	M= 4.423 BD= 2.002	M= 4.291 BD= 1.874	M= 6.4258 BD= 2.5107
VGG Face	M= 3.6658 BD= 2.8537	M= 4.3808 BD= 4.5512	M= 3.8322 BD= 3.287	M= 3.1157 BD= 2.9792	M= 3.3724 BD= 2.4103	M= 3.5298 BD= 3.4605

Table 7- 2. Mean Value rate (M) and Bhattacharyya distance (BD) between distributions:
(N: Neutral, A: Angry, D: Disgusted, F: Fearful, H: Happy, SA: Sad, SU: Surprised)

As makeup can change the appearance of the face, several handcrafted methods were proposed to achieve a facial recognition system that is invariant to makeup [185].

These methods describe the samples by extracting handcrafted features such as Gabor wavelets. Here, we will investigate if we can use auto makeup solution in order to fix a problem that was detected while seeking the answer to the first question – "Does a combination of full makeup and facial expressions bring a statistically significant change in the biometric results?" As has been mentioned before, our chosen system (dlib) can be considered as a facial recognition system that is makeup-invariable and invariable to mood variations, but dlib does not perform well under the simultaneous influence of makeup and mood variation.

As explained in detail, we have two groups of samples: a- real expressive samples without makeup (original one) and b- the same samples, but after applying full makeup (using CyberLink MakeupDirector 2 software).

In order to enhance the reliability of the system, we built the following system in order to first find landmarks on the samples without makeup, and then put colour on selected parts of the face in order to make them more similar to the samples with makeup.

This gives three groups:

Group A – real expressive facial images (no makeup).

Group B – one by one professionally made up expressive facial images (full makeup).

Group C – automatically made up pictures (Auto makeup).

For the purpose of producing the third group, the samples in group one are used, and the dlib code is implemented (Fig. 7-3).

Fig.7-8 presented the empirical cumulative distribution function of genuine scores for all of the three groups.

The blue line shows the distribution of genuine scores in an intra group comparison scenario (each sample in group A has been compared to all the samples in the same group). The red line belongs to the distribution of genuine scores obtained by using an inter group comparison scenario (each sample in group A was compared to all the samples in group C, but for the same users). Finally, the orange line is the cumulative distribution function of genuine scores obtained by comparing the samples in group B (containing the processed samples) with the samples in group C.

As illustrated in Fig 7-8, by using the proposed trick of using dlib:

I - finding landmarks and painting the chosen parts of the expressive face image automatically, then

II - comparing B with C, rather than comparing A with C,

the biometric results will be obtained with better quality.

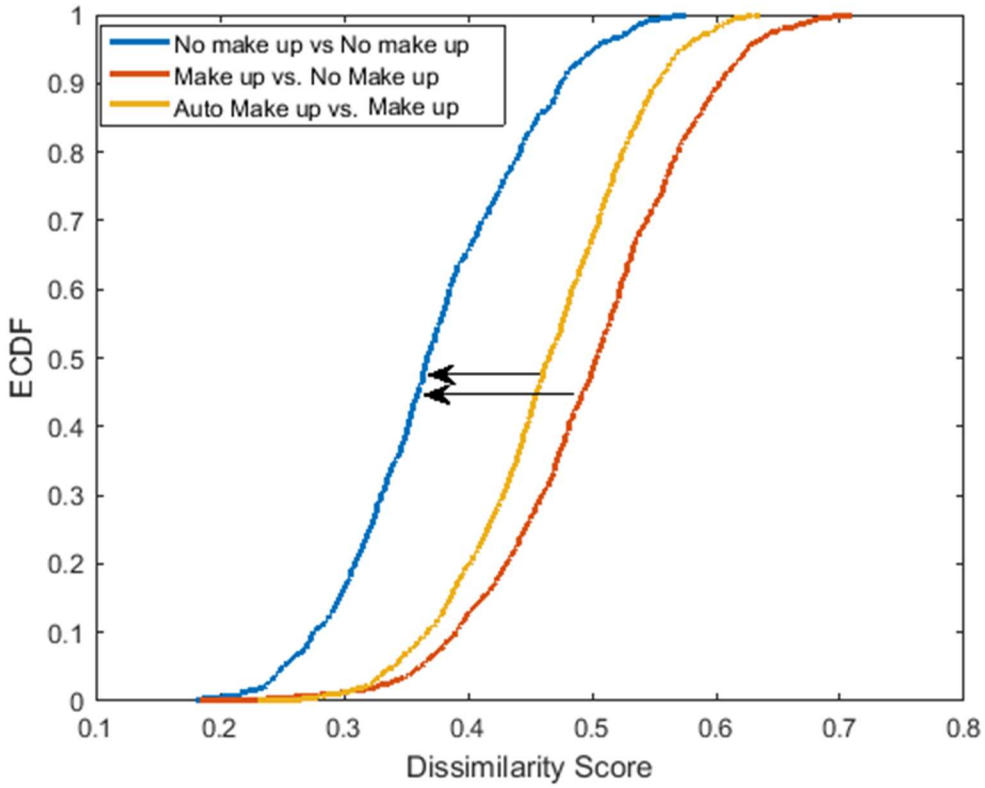


Fig. 7-8 Empirical Distribution functions of genuine scores: no makeup vs no makeup (Blue), with full makeup vs. with full makeup (red), and with automatic makeup vs. with full makeup (orange) – The reliability of system can be enhanced if we can detect the make up worn face images in advance.

Briefly, in this section we used dlib to apply makeup on lips and other facial parts automatically, in order to make the samples with no makeup more similar to the face images with full makeup. This trick can increase the matching accuracy of our systems without changing the chosen algorithm.

To tackle the detected problem, we need to build sophisticated make up detection algorithm which can improve biometric recognition. In order to train our model, we have used 130 samples and for testing the model 304 face images were used.

Table.7-3 shows the statistics of the achieved results. (For obtaining the feature vectors, python face recognition dlib code was implemented.)

	precision	recall	f1-score	Support
Make-up	0.97	0.93	0.95	146
None	0.94	0.97	0.96	158
Avg / total	0.96	0.95	0.95	304

Table.7-3. Classification Results – Make up

We have proposed a SVM-based classifier to detect if the user wears makeup or not with the accuracy of 96%.

For testing the reliability of our dlib-based facial recognition system, the template aging effects on the performance evaluation results were investigated in Appendix.C. According to the presented results in Appendix.C, dlib performs very well (the achieved EER is less than 1%).

7.8. Conclusions

According to the results obtained by using state-of-the-art codes: dlib, as well as two more codes: VeriLook and VGG-Face, and two well-known databases: Radboud and PICS, the following remarks can be reported:

- Facial images displaying disgust are most dissimilar face images to non-expressive pictures of the same user.
- The application of lipstick does not change the recognition accuracy rate of users with angry, disgusted and sad faces.
- According to the results obtained from all three of the used codes and hypotheses tested, it can also be concluded – happy faces are statistically significantly sensitive to the application of lipstick.

- An EER of 4.68% is achieved while identifying people under the simultaneous influences of full makeup and mood variation, while the EER under the effect of each of these factors separately is less than 1%.

This analysis is performed regardless of the features that can be obtained from various facial recognition systems. Here we were interested in the overall system response yielding the measure of similarity or dissimilarity. Instead of changing the verification systems, we used a trick (putting fast makeup on facial images by finding the landmarks and using painting tools) to enhance the reliability of the system.

However, we believe that the emotions shown in the images in the datasets used are prompted or simulated emotions, rather than "real" emotions. And there is no way to be confident that the conclusions would be the same for simulated and real emotions. The makeup used here is not real makeup but virtual makeup, and again we are not confident that the conclusions would be the same for real makeup and virtual makeup.

8. Conclusions

8.1. Achievements

We have investigated the effects of influential parameters on the usage of verification/identification systems, and have attempted to analyze the biometric systems performance variation under various conditions.

Biometric techniques such as face, voice and eye recognition have gained immense popularity (and controversy) in recent years. This thesis reports new findings regarding the influences of certain social factors on biometric recognition. In this thesis, one general statement along with three sub statements (S1, S2 and S3) were proved:

8.1.1. First Sub-Statement (S1):

The iris is the round and colored part of the eye, controlling the diameter of the pupil and the amount of light that reaches the retina. Several studies performed in the last two decades have proven that the pattern of every iris is unique – even in otherwise identical twins. The detailed and complex structure of the iris's front layer therefore makes it an ideal feature for the biometric identification of an individual. Thanks to the unique nature of the human eye, iris recognition is becoming an increasingly popular form of identification for government programs, law enforcement, and access control. Biometric recognition allows a person to be identified on the basis of unique physical characteristics, such as the pattern of the iris.

We examined the relationship between iris recognition and diabetes, finding that iris recognition is less accurate for people with the condition. The study was conducted with a dataset of 1,900 iris images taken from up to 500 eyes, which included healthy irides and the irides of people suffering from diabetes, though it excluded any images with obvious optical impairments. The widespread and well-known disease diabetes very frequently leads to eye conditions such as retinopathy (retinal damage), cataracts (clouding of the lens) and glaucoma (optic nerve damage). Our study, published in Biomedical Engineering, investigated the impact of this medical condition on the accuracy of iris recognition.

Although iris scanning is one of the most reliable methods of personal identification, certain iris abnormalities caused by eye disorders can challenge recognition systems and lower their accuracy.

The biometric accuracy of iris recognition is reduced by the presence of type II diabetes, according to our academic study. In other words, our **research shows that the results of "eye scanning" can be less accurate when a person suffers from diabetes.**

Our third chapter in this thesis, aimed to investigate whether and how exactly diabetes affects the accuracy of iris recognition.

We note that previous studies have indicated the possibility that the false rejection rate in iris recognition may be significantly influenced by factors including ethnicity, gender, and eye color. To account for the results, we also noted that diabetes can affect the eyes in a number of ways that may not be obvious, causing retinal damage, cataracts, and glaucoma.

An IriShield USB MK 2120U device was used for biometric capture, and images compliant with the ISO/IEC 19794-6 standard were used for testing.

From a database of more than 1,900 iris images from 509 eyes (723 iris images from 161 diabetic eyes and 1183 iris images from 348 healthy eyes), we used three different matchers, (open source) and found that accuracy was consistently higher with those who do not have diabetes.

Only irides without visible impairments were used in the study, but we found that iris disorders are often not obvious. We conducted the same test with four different iris recognition systems to make sure they were testing the eyes, and not the quality of the various algorithms. In each case, the results were the same. All three systems had an easier time identifying healthy irides, and were less accurate when scanning diabetic eyes. We compiled a database with more than 1,900 images of healthy and diabetes-affected irides and applied different recognition systems. In all cases, we found that images of healthy eyes allowed for the most successful identification. The analysis of diabetes-affected eyes led to worse results – even though only irides without visible impairments had been considered for the study. The iris disorders are often not obvious, which makes it harder for the systems to recognize diabetic patients.

The study demonstrates that, especially in light of the ever-growing diabetes epidemic, abnormalities in iris patterns need to be taken into account when it comes to the development of new biometric identification methods. We also concluded that differences in pupil dilation

between different age groups should also be considered a possible source of error. In conclusion, our study highlights the limits of this kind of biometric tech, and cautions against the widespread use of methods that only perform well in optimal circumstances and do not account for relatively common conditions like diabetes [186-188].

8.1.2. Second Sub-Statement (S2):

The voice we go to sleep with is significantly different from the one we wake up with. Although having a deeper voice in the morning – a heavier voice after getting up in the early morning – is an abnormal change in the voice, it should not be confused with hoarseness, which is generally caused by an inflamed larynx. The throats tissues collect fluid while we are asleep. During the night, when people are asleep, the lack of use of vocal cord causes mucus to build up. Moreover, most the people breathe through their mouth while they sleep, which causes the vocal cords to dry out slightly, hindering them from moving together without this lubrication during sleep. As a result, we can expect a lower pitch of voice in the early morning. The voice is one the most distinguishable biometric cues that can be used for human identification purpose. Voice samples can be easily recorded and stored using a smartphone microphone, meaning that voice recognition systems have become one of the most popular mobile biometric systems for cell phones. Today's smartphones are equipped with biometric tech such as voice.

The only problem with biometric solutions is when they fail to perform. Here, we presented an investigation on the effect of the time of day on the matching accuracy of a speaker recognition system. A new database was collected and offered. The database contains a dataset of thirty people and we collected 1780 voice samples. There were two different data collection sessions: a- participants were asked to record their voice after getting up in the morning, using their own smartphone devices (916 morning voice samples were recorded), and b- participants were asked to record their voice samples during the day (a further 884 samples were collected from the same users). Each sample is six seconds long and has a bit rate of 705 kbps. All of the users were native speakers of Persian.

In order to conduct the numerical experiments, a pre-trained VGG-Speaker was used. An All vs All comparison scenario was carried out. The intrasession comparison scores are better in comparison with the intersession ones. For the evening versus evening comparison scenario, an

EER of 1.46% was achieved. For the morning versus evening comparison scenario, the EER increased to 10.2%. To tackle this problem, we built a sophisticated morning voice detection algorithm that can improve speaker recognition.

8.1.3. Third Sub-Statement (S3):

Among the influential social issues, facial expression depending on mood and differences in makeup are the most popular factors that have a substantial ability to change the appearance of a face significantly in different ways. During the daytime, people may experience considerable changes in mood. On the other hand, using makeup for beautification will continue to be an indispensable way of life for many people. In terms of mood, we can confidently assume that in most cases, users will upload their official photos to be enrolled to a biometrics system database. In such facial images, users tend to try to have neutral facial expressions in an attempt to look more serious. In real life, depending on the situation, a person's facial appearance may vary considerably, albeit temporarily. The reliability of a facial recognition system is influenced by the mood of the users. Due to the reasons mentioned above, looking at how the reliability of facial recognition systems can be enhanced to better deal with the simultaneous influences of makeup and mood variation is the major purpose of our study. The results showed that, while the effects of makeup and varied mood expressions are not significant by themselves, the joint effect of them taken together is.

An equal error rate (EER) of 4.68% was achieved when identifying faces under the joint influences of full makeup and mood variation, while the EER under the effect of each of these factors separately is less than 1%.

To tackle the detected problems in the mentioned cases, we have built sophisticated makeup and morning voice detection algorithms that can improve face and speaker recognition. The results of this project contribute to the contactless mobile biometric systems, which are important to the biometrics industry. Therefore, it will have a potential economic impact.

8.2. Future Studies

In Chapter 8, we proposed a computer vision tool for age, gender estimation and for diabetes diagnosis using iris texture. We built a support vector machine-based classifier for morning voice

and makeup detection. The feature vectors were obtained by implementing dctc code. While the results were not satisfactory, the use of neural network-based classifiers is encouraged. The number of subjects and number of samples were not sufficient to make general and comprehensive statements. The extended version of other publicly published databases can be used for conducting future numerical experiments.

Appendix. A. Iris recognition for smokers and non-smokers

A1. Overview

Smoking is one of the most common habits around the world and the immediate effect of smoking on the performance of voice recognition systems has been studied in [1]. Smoking is also known as a social issue that can lead to voice deviations, vocal changes, and acoustics complaints, as reported in [1], in which the sub data from the NIST telephone recording database was used.

However, no tests have been conducted to investigate the reliability of iris recognition systems for smokers and non-smokers. The purpose of this section is to understand whether the biometrics results are better for non-smokers in comparison with heavy smokers. According to the obtained results, the iris recognition system performed very well for both smokers and non-smokers groups.

A2. Related works

The negative impact of pupil dilation on iris recognition performance has been reported by several researchers. Karakaya and Celik,[1] quantified the effects of pupil dilation and gaze angle on real frontal and off-angle images at various levels of dilation. The experiments revealed that larger differences in dilation levels and gaze angles between the compared iris images increase the Hamming distance. Temeo-Reyes et al.[2] analyzed the impact of drugs on pupil dilation, (2) proposes the use of a biomechanical nonlinear iris normalization scheme along with a key point-based feature matching for mitigating the impact of drug-induced pupil dilation on iris recognition, and (3) investigates differences between drug-induced and light-induced pupil dilation on iris recognition performance. In a paper written by Erdem et al.,[3] the acute effects of smoking on Intra Ocular Pressure (IOP) and pupil size were presented. In their study, 52 healthy habitual cigarette smokers were recruited and the achieved results showed that the photopic pupil sizes of the subjects before and after smoking were significantly different. The mean intra ocular

[1] Mahmut Karakaya; Elif T. Celik, Effect of pupil dilation on off-angle iris recognition, J. of Electronic Imaging, 28(3), 033022 (2019). https://doi.org/10.1117/1.JEI.28.3.033022

[2] Inmaculada Tomeo-Reyes; Arun Ross; Vinod Chandran, Investigating the impact of drug induced pupil dilation on automated iris recognition, 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), DOI: 10.1109/BTAS.2016.7791178

[3] Uzeyir Erdem; Gokcen Gokce; Fatih Cakir Gundogan; Atilla Bayer; Gungor Sobaci, effect of Smoking on Intra Ocular Pressure Wavefront Aberrations and Pupil Changings, Investigative Ophthalmology & Visual Science April 2011, Vol.52, 5191.

pressure (IOP) after smoking was also significantly higher than the IOP before smoking. For the hypothesis test, a T-test was used. In another study by the same authors,[4] pupil sizes and total ocular aberrations were assessed by an optical path difference scanning system (OPD-Scan II ARK-10000, NIDEK) and only the right eyes were considered for statistical analysis. SPSS version 11.0 statistical software system was used for all statistical analyses and was regarded as statistically significant. They reported that the smokers in the study had an abstinence period for at least 12 hours before the measurements, hence the psychological relaxation after smoking in chronic smokers may also cause the inhibition of sympathetic activation in the abstinence period. In one paper,[5] the authors evaluated the acute changes in objective accommodation and OWA after cigarette smoking.

A3. Database

For the purpose of collecting a new database, regular cigarette smoker volunteers from the staff of Babol University of Medical Sciences were recruited in the study. The participants were required to meet the criteria of having smoked 15 or more cigarettes per day for at least a five-year period. Informed consent was obtained from all of the volunteers. Our new database contains 284 iris samples taken from 41 eyes of smokers and 390 iris samples from 87 eyes of non-smokers (Fig. A-1).

The experiment was specifically designed to investigate whether there is any relation between the accuracy of the iris recognition system and the fact that the users smoked, or not. Personal data were kept separately, in order to guarantee additional security of the personal data. All of the participants were fully aware of the experiment as full detailed information on the study was provided and signed consent forms were also obtained from all of the individuals. The experiment protocol was approved by the Ethics Committee of the Warsaw University of Technology.

---

[4] Uzeyir Erdem, Fatih C. Gundogan, Umut Aslı Dinc, Umit Yolcu, Abdullah Ilhan, and Salih Altun, Acute Effect of Cigarette Smoking on Pupil Size and Ocular Aberrations: A Pre- and Post-smoking Study, Journal of Ophthalmology, Volume 2015, Article ID 625470, 5 pages, http://dx.doi.org/10.1155/2015/625470

[5] Handan Bardak and Murat Gunay and Yavuz Bardak and Yesim Ercalik and Serhat Imamoglu and Elvin Yildiz and Betul Onal Gunay, Evaluation of the acute changes in objective accommodation, pupil size and ocular wavefront aberrations after cigarette smoking, Cutaneous and Ocular Toxicology, Volume 36, 2017 - Issue 1, Pages 25-28.

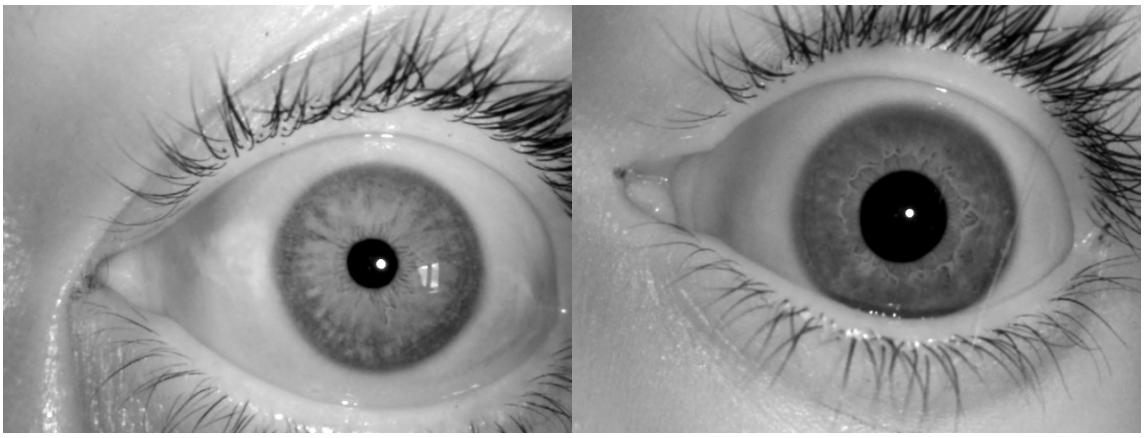


Fig. A-1. Iris samples – a: nonsmoker, b: smoker

A4. Methodology

In order to calculate the comparison score between samples (for an all vs all comparison scenario), we have again used the University of Salzburg Iris Toolkit (USIT). For iris segmentation, the Weighted Adaptive Hough and Ellipsopolar Transform methodology was used.

We have selected a descriptor for feature extraction: the iris coding method based on differences of discrete cosine transform (DCT). In our chosen strategy, all of the images would be compared to the entire database and a score would be obtained for each.

A5. Results and Discussions

The ROC curve was presented in Figure A-2. The methodology used was Discrete Cosine Transform (DCT). As shown in the figure, and according to the results achieved by the matchers, the accuracy of the system is the same when recognizing healthy people and chain smokers using their iris texture images. A very good performance was observed by using the methodology proposed by Monro et al (DCT).

As shown from Figure A-3, the empirical cumulative distribution function of genuine scores obtained by DCT codes does not show any observable differences between those comparison scores obtained by comparing iris samples donated by smokers and those matching results achieved by a comparison of iris

pattern images taken from healthy irides. Iris recognition effectiveness is same for both smokers and non-smokers.

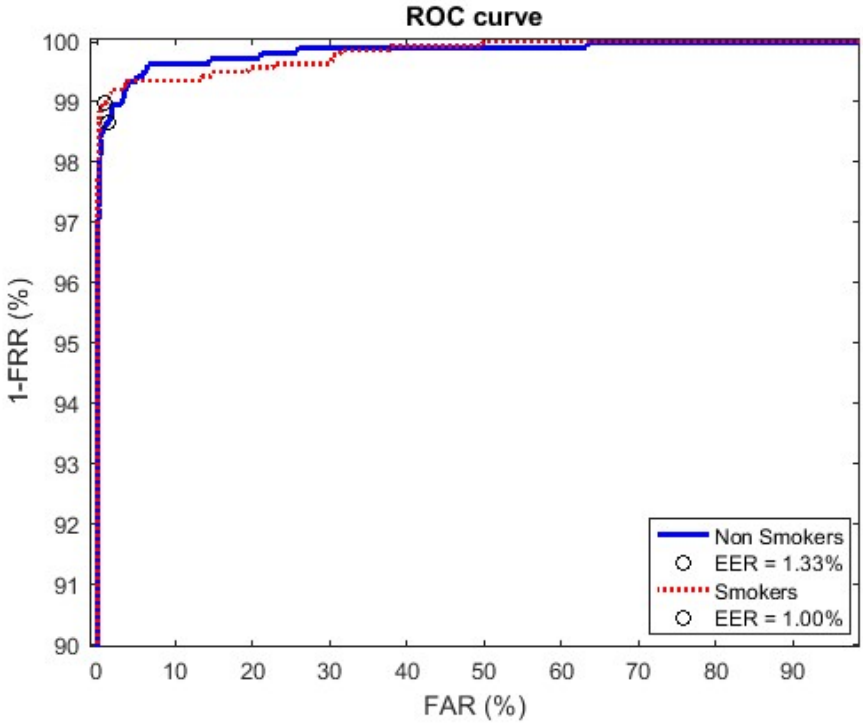


Fig. A-2. Roc curve

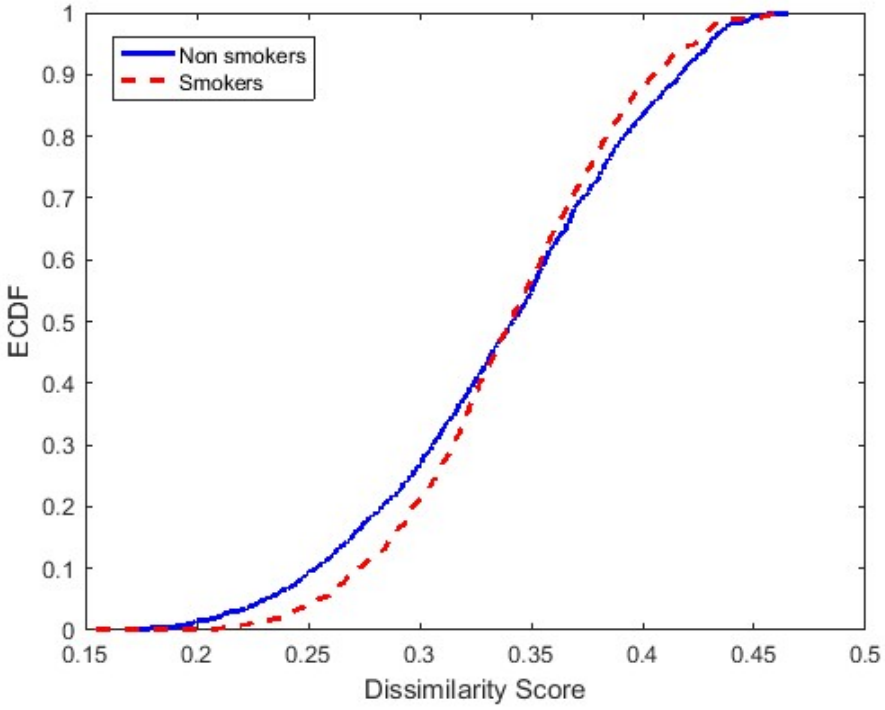


Fig. A-3.ECDF

A6. Conclusions

To address whether cigarette smoking actually has an effect of reducing the matching accuracy of iris recognition systems, our study examined the effect of smoking on the reliability of an iris identification system. Our newly offered database consists of 674 NIR iris images taken from 128 eyes by using a monocular IriShield camera. In order to segment the iris and calculate the comparison score between samples, the University of Salzburg Iris Toolkit (USIT) was successfully used. We conducted the same test with four different iris recognition systems, to make sure we were testing the eyes, and not the quality of the various algorithms. In each case, the results were the same. We can confidently state that **"iris recognition is robust against the smoking condition of the users."**

Appendix. B. The Influence of Acted Mood Variation on Text Independent Speaker Recognition System's Reliability

B1. Overview

To date, there has been very little research on the influence of psychological factors of the use on biometric reliability. Here, we attempt to investigate the performance evaluation results of a chosen text independent speaker recognition system (VGG-Voice model for speaker recognition) under the influence of mood variation. The main goal of this study is to find the best and worst inter mood comparison scenarios. For this purpose we used the Ryson Audio-Visual Database of Emotional Speech and Song (RAVDESS) containing acted expressive voice samples donated by 24 actors (twelve women and twelve men), vocalizing in a neutral North American accent. For extracting the features, and therefore obtaining the comparison scores between samples, the output of penultimate layer of the VGG voice model, as proposed by Xie et al., was used. According to our results, matching calm versus angry samples was the worst comparison scenario (EER = 13.6733 per cent).

B2. Related Works

People can recognize another person's emotions in daily communications, but computers are still barely capable of determining the emotions of people with the same accuracy. The emotional mood variation can affect the reliability of biometrics recognition systems by changing the biometric characteristics of the same individuals. In real life, we may face a range of emotional conditions, (in court, at the hospitals, upon suffering an accident, for example) and the biometric system must be able to recognize and verify the user under the influence of mood variation perfectly. In the context of using voice as a biometric, it is important to assess whether these reported variations among a population affect the performance of a standard system. However, the key challenge in the speaker verification task is to determine whether the performance of voice recognition systems will be significantly degraded in emotional talking environments. In one paper,[6] an emotional speaker recognition system based on a number of feature extraction

[6] Mansour, A. and Lachiri, Z., "Speaker recognition in emotional context," International Journal of Computer Science, Communication and Information Technology (CSCIT), pp. 1-4, Oct. 2015.

methods, focusing on the diversities between simulated and natural emotional speech databases (BERLIN and IEMOCAP) was assessed. Their reported results show that the context of emotional speech also impacts speaker recognition rates significantly. In this section, we are trying to answer the following question: 1. which mood can make the verification task harder over all?, 2. which inter mood comparison scenario will make the identification task harder? and finally, 3. can the intensity of the acted emotions be considered as an influential parameter that can play a key role? To the best of our knowledge, such research analysis has never been reported on.

B3. Database

Here, as we were using a database containing voice samples in eight different moods, we carried out $\frac{8\times7}{2}+8=36$ numerical experiments. We also present the obtained results of two more chosen comparison scenarios: a- Neutral vs. All, and b. All vs. All by a- matching non-expressive and expressive voice samples, and b- matching all of the voice samples, respectively. To answer the predetermined questions, audio files in the Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS)[7] were used. The database contains 1440 audio-only files donated by 24 users (12 actresses and 12 actors) in eight different acted emotional states: neutral, calmness, happiness, sadness, anger, fear, disgust, and surprise). There is no strong intensity for the 'neutral' emotion hence the database provides four non-expressive samples and eight expressive samples for each of the mentioned moods. This means that 60 samples from each of the users, in various moods with varying intensity of expression, are available for the numerical experiments. All of the samples are given as 48 kHz, 16-bit two-channel WAV files.

B4. Methodology

For speaker identification tasks 'in the wild', where utterances may be of variable length and also contain irrelevant signals, Xie et al.[8] proposed a powerful speaker recognition network using a 'thin-ResNet' trunk architecture and dictionary-based NetVLAD and GhostVLAD layers to

[7] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.
[8] W. Xie, A. Nagrani, J. S. Chung and A. Zisserman, "Utterance-level Aggregation for Speaker Recognition in the Wild," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 5791-5795.

aggregate features across time that can be trained end-to-end. The mentioned descriptor is generated by a Convolutional Neural Network (CNN). The output of the layer before the classification layer can be considered as the extracted feature vector containing 512 elements. The Euclidian distances between the extracted feature vectors were calculated. The dissimilarity score is the statistical (Euclidean) distance between the vectors. It is obvious that zero is the comparison score between identical voice samples and a higher dissimilarity means a lower matching score between the voice samples.

The speaker recognition system returns symmetrical matching scores, which means that, as we have the samples in eight different moods, so we need to make 36 numerical experiments in order to answer the main question of this study: what are the best and worst inter mood comparison scenarios for the verification task?

B5. Results and Discussions

In this section the obtained results will be presented. In this study, the following comparison scenarios were chosen to answer the predetermined question:

A. Finding an answer to our main question: which inter mood comparison scenario will make the identification task harder over all. We will consider all of the possible intra and inter mood comparison scenarios to find the best and worst cases (Mood versus Mood).

We have neglected the effects of gender factor on the reliability of system, but the investigation into the effects of expression intensity on the reliability of the system was one of our favorites, hence the last additional numerical experiment is:

B. Making a comparison between the results obtained by matching non-expressive samples with expressive samples with a lower intensity, and those achieved by matching non-expressive samples with expressive ones with a strong intensity.

In Table B-1, the achieved EER for all of the possible cases were tabulated (Scenario A). Figure B-1 shows the RoC curves for matching Scenario B. According the figure, there is a statistically significant different between the two graphs and a higher intensity of emotional expression can cause more serious problems for the speaker recognition system. Although recordings are available for two different sentences: “Kids are talking by the door" and “Dogs are sitting by the

door", as our chosen system is a text independent speaker recognition system, this difference was not taken into the account in our study.

We attempted to observe the structure of the similarity space by using hierarchical clustering to check whether voice samples of a similar mood condition clustered well within each group.
According to our results, a clustering based on the moods of the samples appears evident. This is due to the fact that, although the VGG-speaker pre-trained model performs well in the wild, the algorithm's robustness to vocal mood expression cannot be proven. Another reason could be related to the variation of expression with voice samples, which is very high.
The reliability of a speaker recognition system is influenced by the mood of the users. Due to the reasons mentioned above, looking at how the reliability of voice recognition systems can be enhanced to better deal with the simultaneous influences of mood variation and environmental effects is the major purpose of further studies.

B6. Conclusions

This section looked at a reliability study of a text independent speaker recognition system under the influence of mood variation. We had three questions and we wanted to discover:
1. which mood has highest potential to place the reliability of a system at risk ?,
2. which inter and intra mood comparison scenarios are the most and least similar?
3. whether the intensity of acted emotions can play a key role or not?
The findings are as follows:
1. The best intra mood comparison scenario is Neutral vs. Neutral with an equal error rate of 0.07 per cent; the worst one is Sad vs. Sad with an EER of 9.12 per cent.
2. The best inter mood comparison scenario is Calm vs. Neutral with an equal error rate of 3.23 per cent; the worst one is Calm vs. Angry with an EER of 13.67 per cent.
3. Emotions intensity can change the result in a meaningful way.

Mood	Angry	Calm	Disgusted	Fearful	Happy	Neutral	sad	surprised
Angry	8.35	13.67	10.87	12.11	9.23	11.97	13.42	10.66
Calm	13.67	0.92	9.37	13.56	9.06	3.23	13.28	7.09
Disgusted	10.87	9.37	2.86	11.27	8.73	9.37	10.29	7.68
Fearful	12.11	13.56	11.27	7.84	10.12	11.71	10.67	9.32
Happy	9.23	9.06	8.73	10.12	5.98	7.55	9.81	6.31
Neutral	11.97	3.23	9.37	11.71	7.55	0.07	10.55	7.51

Sad	13.42	13.28	10.29	10.67	9.89	10.55	9.12	10.62
Surprised	10.66	7.09	7.68	7.84	6.31	7.51	10.62	3.12

Table B-1. EER – Different Scenarios

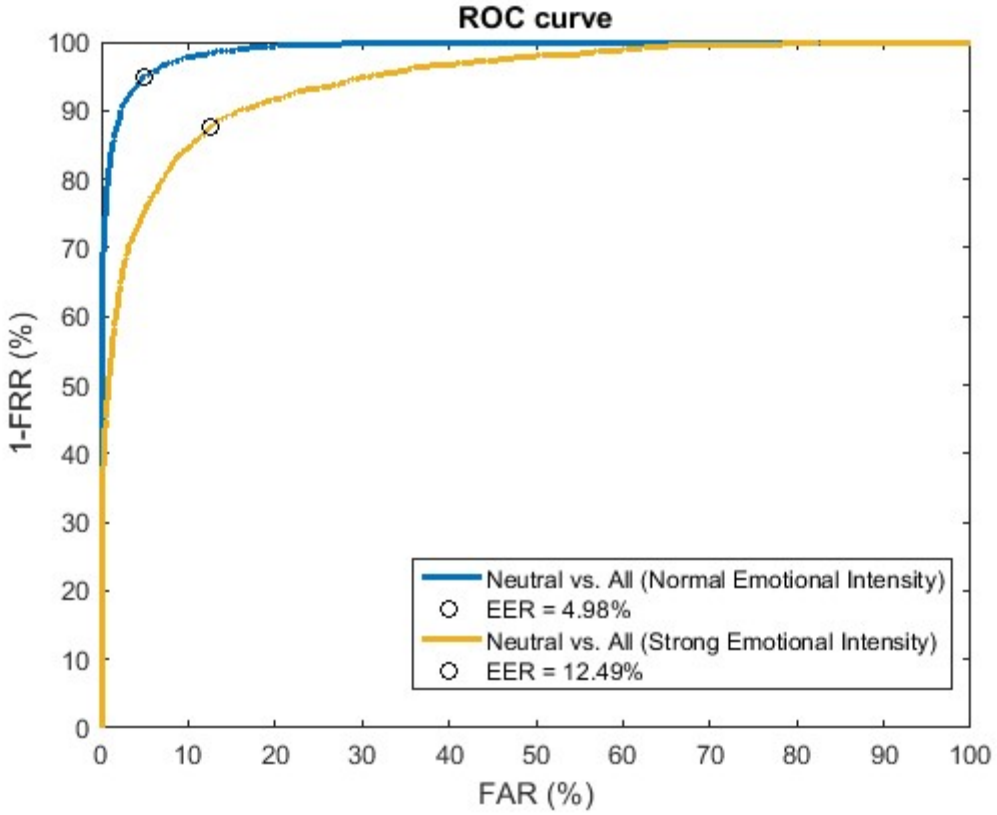


Fig. B-1. Effect of expression intensity on the matching score

Appendix. C. Can we solve facial aging problem through the use of age-progression software?

C1. Overview

In this short report, we will present the results of a numerical experiment in order to test the well-known and popular age-progression software (the FaceApp application). We will also check how identity is preserved on both a synthetic and real dataset? We are trying to find an answer to the question: is it better to change the age of the face in the picture and then match, or whether this trick cannot be considered as an influential strategy in order to enhance the matching accuracy of facial recognition systems?

C2. Database

For this purpose, a database containing facial images from 67 celebrities was collected using open source materials available via Google image search. For each person, two samples were downloaded by searching the name of each celebrities + "young" and "old". After creating the database, we used the FaceApp application in order to produce corresponding virtual aged images of each young face sample. Figure C-1 shows 12 pictures, where the first image of each row is the original sample, the second image is the same picture made to look old through the FaceApp application, and the final image is how each person actually looks now. Hence, for each user there are three face samples (young, virtual and old images). As a result, there are 201 samples from the 67 celebrities in our database. After gathering the samples and collecting the aged images made using FaceApp, we partitioned the database in to three groups: 1. Young (contains original pictures from the individuals when they were young); 2. Virtual (the samples in the first group, but made to look old using the FaceApp application); and finally, 3. Old (this group contains samples showing how each person actually looks now).

C3. Methodology

We compared the samples in the Young group (first group) with those samples included in the Old group (the last group) firstly, to obtain a performance evaluation results of the face recognition system under the influence of the aging factor. Then we made a comparison between the samples in the Virtual group and the Old Group, in order to find an answer to the question

mentioned above, of whether it is possible to solve facial aging problem by the use of age-progression software or not.

Fig. C- 1. Samples from the collected database (The first image of each row: a young image; the second image of each row: the same picture, but aged using the FaceApp application; the third image is how the person actually looked when old)

To obtain the results, the python face recognition dlib code was implemented. Python face recognition dlib is one of the most popular and powerful codes provided for face verification purposes. Dissimilarity is understood as the distance between feature vectors in space with

dimensions equal to the number of features analyzed in a particular method (N=128 for python face recognition).

C4. Results

As presented in Table C-1, for the chosen algorithm, the genuine scores distribution vary significantly for Young-Old and Young-Virtual comparisons (it means that the samples made by FaceApp are more similar to old images than they are to the original young images). However, at same time the mean value of impostor scores distribution is higher for Old-Virtual. According to the presented RoC curve in Figure C-2, the equal error rate for Old-Virtual (EER = 0.048%) is higher than same value for Old-Young (EER = 0.037%).

	GMean	IMean	J-Index	MCC	FMR=0	FNMR0
OV	0.5606371	0.2874024	0.924366	0.8033269	0.7777778	0.3020084
OY	0.5849621	0.3212205	0.9102287	0.8065484	0.9047619	0.5655839

Table C-1. biometrics statistics: OV means Old-Virtual and OY means Old-Young

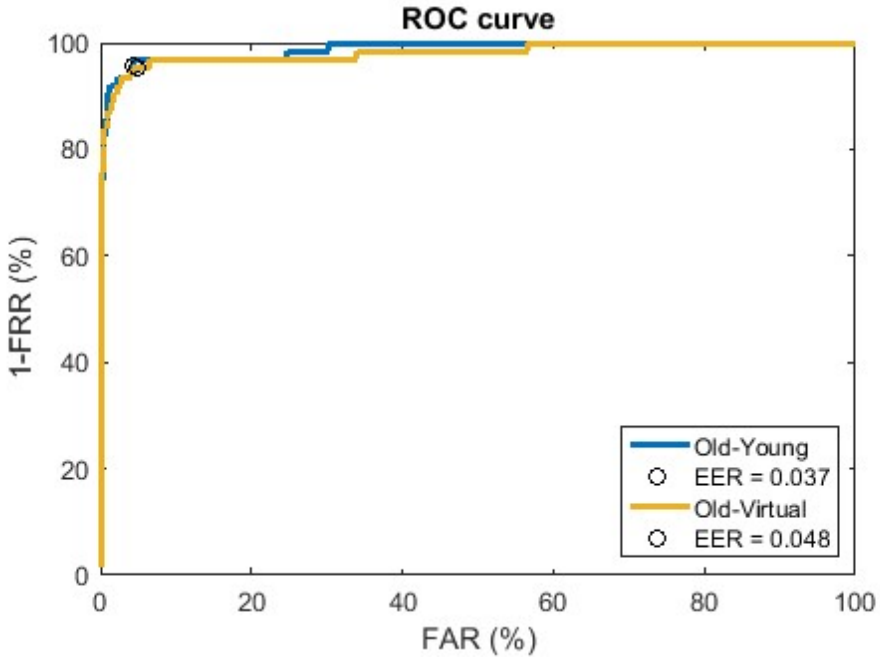


Fig. C-2. ROC curve for two different comparison scenarios.

C5. Conclusions

The observed results were expected, and this is rather obvious because the used age progression algorithm (the FaceApp application) is not an identity-preserving aging algorithm but rather a pleasurable-looking aging algorithm.

As a final remark, it must be reported that the results obtained in the Old-Virtual case are worse than the Old-Young comparison scores, so we cannot make a statement that virtual aging will help us in the process of searching through databases with a large age variation.

References

[1] P. R. Nalla and A. Kumar, "Toward More Accurate Iris Recognition Using Cross-Spectral Matching," in IEEE Transactions on Image Processing, vol. 26, no. 1, pp. 208-221, Jan. 2017.

[2] E. E. Hansley, M. P. Segundo and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," in IET Biometrics, vol. 7, no. 3, pp. 215-223, 5 2018.

[3] X. Yin, Y. Zhu and J. Hu, "Contactless Fingerprint Recognition Based on Global Minutia Topology and Loose Genetic Algorithm," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 28-41, 2020.

[4] Y. Seto, Retina Recognition. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2009.

[5]E. Bartuzi, K. Roszczewska, A. Czajka and A. Pacut, "Unconstrained biometric recognition based on thermal hand images," 2018 International Workshop on Biometrics and Forensics (IWBF), Sassari, 2018, pp. 1-8.

[6] J. Lu, V. E. Liong, X. Zhou and J. Zhou, "Learning Compact Binary Face Descriptor for Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pp. 2041-2056, 1 Oct. 2015.

[7] T. Lin and Y. Zhang, "Speaker Recognition Based on Long-Term Acoustic Features With Analysis Sparse Representation," in IEEE Access, vol. 7, pp. 87439-87447, 2019.

[8] R. Sanchez-Reillo, J. Liu-Jimenez and R. Blanco-Gonzalo, "Forensic Validation of Biometrics Using Dynamic Handwritten Signatures," in IEEE Access, vol. 6, pp. 34149-34157, 2018.

[9] Y. Zhang, Y. Huang, S. Yu and L. Wang, "Cross-View Gait Recognition by Discriminative Feature Learning," in IEEE Transactions on Image Processing, vol. 29, pp. 1001-1015, 2020.

[10] X. Sha, C. Lian, Y. Zhao, J. Yu, S. Wang and W. J. Li, "An Explicable Keystroke Recognition Algorithm for Customizable Ring-Type Keyboards," in IEEE Access, vol. 8, pp. 22933-22944, 2020.

[11] Y. Yan and L. A. Osadciw, Bridging Biometrics and Forensics, EECS, Syracuse University, Syracuse, NY, USA,

[12] A. Jain, A. Ross, K. Nandakumar, Introduction to Biometrics, Springer (2011).

[13] J. Blasco, T.M. Chen, J. Taoiador, P. P. Lopez, A survey of wearable biometric recognition systems, Journal ACM Computing Surveys Volume 49, Issue 3, December 2016, Article No. 43

[14] A. Rattani, R.Derakhshani, Ocular biometrics in the visible spectrum: A survey, Image and Vision Computing, Volume 59, March 2017, Pages 1-16.

[15] P. Sh. Teh, N. Zhang, A. B. J.Teoh, K. Chen, A survey on touch dynamics authentication in mobile devices, Computers & Security Volume 59, June 2016, Pages 210-235.

[16] Q. Tao, R. Veldhuis, Biometric authentication system on mobile personal devices IEEE Trans Instrum Meas, 59 (2010), pp. 763-773

[17] K. Lee, B. Ma, L. H. Speaker verification makes its debut in smartphone IEEE Signal Process. Soc. SLTC Newslett. (2013)

[18] L.M. Mayron, Biometric Authentication on Mobile Devices, IEEE Security & Privacy, Vol.13, Issue 3, May-June 2015, pp. 70-73.

[19] A. Wójtowicz, K. Joachimiak, Model for adaptable context-based biometric authentication for mobile devices, Personal and Ubiquitous Computing, 2016, Volume 20, Issue 2, pp. 195–207.

[20] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, M. Savvides Gait-ID on the move: pace independent human identification using cell-phone accelerometer dynamics 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE (2012), pp. 8-15

[21] S. Farokhi, J. Flusser, U. U. Sheikh, Near infrared face recognition: A literature survey, Computer Science Review, Volume 21, August 2016, Pages 1-17. [22] U. Park, R. Jillela, A. Ross, A.K. Jain Periocular biometrics in the visible spectrum IEEE Trans. Inf. Forensics Security, 6 (1) (2011), pp. 96-106

[23] H. Proença, Iris recognition: on the segmentation of degraded images acquired in the visible wavelength, IEEE Trans. Pattern Anal. Mach. Intell., 32 (8) (2010), pp. 1502-1516.

[24] C. Okoli, K. Schabram, A Guide to Conducting a Systematic Literature Review of Information Systems Research. Working Papers on Information Systems, 2010, 10(26), 1–51.

[25] G. Panis, A. Lanitis, N. Tsapatsoulis, T. F. Cootes, An Overview of Research on Facial Aging using the FG-NET Aging Database IET Biometrics, 2015, DOI: 10.1049/iet-bmt.2014.0053.

[26] M. C. Fairhurst and M. Erbilek, “Analysis of physical ageing effects in iris biometrics,” IET Computer Vision, vol. 5, no. 6, pp. 358–366, Nov. 2011.

[27] J. Galbally, M. Martinez-Diaz, J. Fierrez, Aging in Biometrics: An Experimental Analysis on On-Line Signature, PLoS ONE, Published: July 23, 2013, https://doi.org/10.1371/journal.pone.0069897.

[28] S. K. Modi, S. J. Elliott, J. Whetsone, and H. Kim, “Impact of age groups on fingerprint recognition performance,” in Proc. IEEE Workshop Autom. Identification Adv. Technol., Alghero, Italy, Jun. 2007, pp. 19–23.

[29] R. Merkel, J. Dittmann, and C. Vielhauer, “How contact pressure, contact time, smearing and oil/skin lotion influence the aging of latent fingerprint traces: First results for the binary pixel feature using a CWL sensor,” in Proc. IEEE Int. Workshop Inf. Forensics Security, Nov. 2011, pp. 1–6.

[30] N. C. Sickler and S. J. Elliott, “An evaluation of fingerprint image quality across an elderly population vis-a-vis an 18-25 year old population,” in Proc. 39th Annu. Int. Carnahan Conf. Security Technol., Oct. 2005, pp. 68–73.

[31] A. K. Jain; S. S. Arora; K. Cao; L. Best-Rowden; A. Bhatnagar, Fingerprint Recognition of Young Children, IEEE Transactions on Information Forensics and Security, Volume: 12, Issue: 7, July 2017.

[32] E. Liu, Infant Footprint Recognition, 2017 IEEE International Conference on Computer Vision (ICCV), 10.1109/ICCV.2017.183

[33] M. Madry-Pronobis. Automatic gender recognition based on audiovisual cues. Master Thesis, 2009.

[34] R. Singh, M. Vatsa, A. Noore, and S. K. Singh, "Age transformation for improving face recognition performance," in Pattern Recognition and Machine Intelligence (Lecture Notes in Computer Science), vol. 4815, A. Ghosh, R. De, and S. Pal, Eds. Berlin, Germany: Springer, 2007, pp. 576–583

[35] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in Proc. 3rd IEEE Int. Conf. Biometrics, Theory, Appl. Syst., Piscataway, NJ, USA, 2009, pp. 139–146.

[36] M. Erbilek and M. C. Fairhurst, "A methodological framework for investigating age factors on the performance of biometric systems," in Proc. Int. Conf. Multimedia Security, New York, NY, USA, 2012, pp. 115–122.

[37] R. M. Guest, "Age dependency in handwritten dynamic signature verification systems," Pattern Recog. Lett., vol. 27, no. 10, pp. 1098–1104, 2006.

[38] Faundez-Zanuy, M., Sesa-Nogueras, E., & Roure-Alcobé, J. (2012). On the relevance of age in handwritten biometric recognition. In Security Technology (ICCST), 2012 IEEE International Carnahan Conference on (pp. 105-109). IEEE.

[39] R. Winkler, M. Brückl, W. Sendlmeier, The aging voice: An acoustic, electroglottographic and perceptive analysis of male and female voices Proceedings of the ICPHS, Barcelona (2003).

[40] R.J. Morris, C.R. McCrea, K.D. Herring, Voice onset time differences between adult males and females: Isolated syllables, Journal of Phonetics, 36 (2008), pp. 308-317

[41] M.M. Gorham-Rowan, J. Laures-Gore, Acoustic-perceptual correlates of voice quality in elderly men and women Journal of Communication Disorders, 39 (2006), pp. 171-184

[42] M. Nishio, S. Niimi, Changes in speaking fundamental frequency characteristics with aging Folia Phoniatrica et Logopaedica, 60 (2008), pp. 120-127

[43] P. Torre, J.A. Barlow, Journal of Communication Disorders, Age-related changes in acoustic characteristics of adult speech, Volume 42, Issue 5, September–October 2009, Pages 324-333

[44] F. Zappasodi, L. Marzetti, E. Olejarczyk, F. Tecchio, and V. Pizzella, Age-Related Changes in Electroencephalographic Signal Complexity, PLoS One. 2015; 10.1371/journal.pone.0141995.

[45] H. Akatsu & H. Miki (2004): 'Usability research for the elderly people'. Oki Technical Review (Special Issue on Human Friendly Technologies); 71(3), 54–57

[46] M. Ziefle & S. Bay (2006): 'How to overcome disorientation in mobile phone menus: A comparison of two different types of navigation aids'. Human Computer Interaction, 21(4), 393–432.
[47] M. Obrist, R. Bernhaupt, E. Beck&M. Tscheligi (2007): 'Focusing on elderly: An iTV usability evaluation study with eye-tracking. In: P. Cesar, K. Chorianopoulos, and J.F. Jensen (Eds.), 5th European Conference on Interactive TV – EuroITV 2007 May 24, 25, 2007, Amsterdam, the Netherlands, pp. 66–75
[48] A. Calero-Valdez, M. Ziefle, F. Alagöz & A. Holzinger (2010): 'Mental models of menu structures in diabetes assistants'. In: K. Miesenberger, J. Klaus, W. Zagler & A. Karshmer (Eds.), ICCHP 2010. LNCS, vol. 6180, pp. 584–591.
[49] M. Theofanos, B. Stanton, S. Orandi, R. Micheals & N.F. Zhang (2007): 'Ten-print fingerprint capture: Effect of instructional modes on user performance'. Proceedings of the Human Factors and Ergonomics Society Annual Meeting (vol. 51, no. 10, pp. 597–601). Sage. New Orleans, Louisiana.
[50] A. Sasse, Kat Krol, Usable biometrics for an ageing population, September 2013, DOI: 10.1049/PBSP010E_ch16.
[51] E. Ahmed, B. DeLuca, E. Hirowski, C. Magee, I. Tang, J. F. Coppola, Biometrics: Password Replacement for Elderly?, 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pages: 1 – 6
[51] A. Holzinger, G. Searle, T. Kleinberger, A. Seffah & H. Javahery (2008): 'Investigating usability metrics for the design and development of applications for the elderly, In: K. Miesenberger, J. Klaus, W.L. Zagler, A.I. Karshmer (Eds.), ICCHP 2008, Springer, Berlin Heidelberg.
[52] F. Zouaoui; N. Bouadjenek; H. Nemmour; Y. Chibani, Co-training approach for improving age range prediction from handwritten text, 10.1109/ICEE-B.2017.8192233.
[53] M. Fairhurst, M. Erbilek and M. Da Costa-Abreu, "Selective Review and Analysis of Aging Effects in Biometric System Implementation," in IEEE Transactions on Human-Machine Systems, 45(3), 294-303, 2015.
[54] https://en.wikipedia.org/wiki/Longitudinal_study.
[55] O.V. Komogortsev, et al., "Template aging in eye movement driven biometrics", Proceedings Volume 9075, Biometric and Surveillance Technology for Human and Activity Identification XI; 90750A, 2014.
[56] A. Czajka, "Analysis of diurnal changes in pupil dilation and eyelid aperture", IET Biometrics, 7(2), pp. 136-144, 2018.
[57] T. Scheidat, K. Kummel, and C. Vielhauer, "Short term template aging effects on biometric dynamic handwriting authentication performance," in Proc. 13th IFIP TC 6/TC 11 Int. Conf. Commun. Multimedia Security, 2012, pp. 107–116.

[58] S. E. Linville, “The Aging Voice,” The American Speech Language-Hearing Association (ASHA) Leader, pp. 12–21, 2004.
[59] Y. Samona; C. Pintavirooj; S. Visitsattapongse, Study of ECG variation in daily activity, 2017 10th Biomedical Engineering International Conference (BMEiCON), 10.1109/BMEiCON.2017.8229170.
[60] I. Manjani, et al., “Template aging in 3D and 2D face recognition”, IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016.
[61] E. Maiorana, P. Campisi, "Longitudinal Evaluation of EEG-Based Biometric Recognition", IEEE Transactions on Information Forensics and Security, 13(5), 1123-1138, 2018.
[62] F. Kelly, J. H. L. Hansen, “Score-Aging Calibration for Speaker Verification”, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (12), 2414-2424, 2016.
[63] J. Galbally, et al., “A Study of Age and Ageing in Fingerprint Biometrics”, IEEE Transactions on Information Forensics and Security, 14 (5) 1351-1365, 2019.
[64] H. Liu, "Age-related differences in vocal responses to pitch feedback perturbations: a preliminary study", J Acoust Soc Am., 127(2):1042-6, 2010.
[65] M. Johnson et al., “A longitudinal study of iris recognition in children”, IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), 1-7, 2018.
[66] S. P. Fenker and K. W. Bowyer, “Analysis of template aging in iris biometrics,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops, 2012, pp. 45–51.
[67] G. Guo, G. Mu, and K. Ricanek, “Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits,” in Proc. 20th Int. Conf. Pattern Recog., Istanbul, Turkey, Aug. 2010, pp. 3392–3395.
[68] H. Beigi, “Effects of time lapse on speaker recognition results,” in Proc. 16th Int. Conf. Digit. Signal Process., Piscataway, NJ, USA, 2009, pp. 1260–1265.
[69] F. Alonso-Fernandez, J. Fierrez, A. Gilperez, and J. Ortega-Garcia, “Impact of time variability in off-line writer identification and verification,” in Proc. 6th Int. Symp. Image Signal Process. Anal., Salzburg, Austria, Sept. 2009, pp. 540–545
[70] S. Yoon, A.K. Jain: Longitudinal study of fingerprint recognition. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 112(28), 8555–8560 (2015).
[71] M. Erbilek and M. C. Fairhurst, “Analysis of ageing effects in biometric systems: Difficulties and limitations,” in Age factors in Biometric Processing. Institution of Engineering and Technology, London, U.K., 2013, Ch. 15.
[72] G. Praprotni, N. Pavesic, The impact of template aging on the performance of automatic fingerprint recognition, January 2016Revija za kriminalistiko in kriminologijo 67(4):371-388.

[73] S. Kirchgasser; A. Uhl, Template ageing in non-minutiae fingerprint recognition, 10.1109/IWBF.2017.7935091, 2017 5th International Workshop on Biometrics and Forensics (IWBF).

[74] S. Kirchgasser; A. Uhl, Fingerprint Template Ageing Vs. Template Changes Revisited, 10.23919/BIOSIG.2017.8053507, 2017 International Conference of the Biometrics Special Interest Group (BIOSIG).

[75] L. Best-Rowden; A. K. Jain, A longitudinal study of automatic face recognition, International Conference on Biometrics (ICB), 10.1109/ICB.2015.7139087.

[76] A. P. Rebera and B. Guihen, "Biometrics for an ageing society societal and ethical factors in biometrics and ageing," 2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), Darmstadt, 2012, pp. 1-4.

[77] S. E. Baker, K. W. Bowyer, and P. J. Flynn, “Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches,” in Proc. 3rd Int. Conf. Adv. Biometrics, 2009, pp. 1170–1179

[78] A. Czajka, Influence of Iris Template Aging on Recognition Reliability, Chapter, November 2014, DOI: 10.1007/978-3-662-44485-6.

[79] K. Browing, N. Orlans, Biometric Aging effects of Aging on Iris Recognition, MITRE Innovation Program, 51MSR609-EA.

[80] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, “Face recognition vendor test 2002,” in Proc. IEEE Int. Workshop Anal. Modeling Faces Gestures, Oct. 2003, p. 44

[81] J. Galbally, M. Martinez-Diaz, J. Fierrez, Aging in Biometrics: An Experimental Analysis on On-Line Signature, Plos one, Published: July 23, 2013, https://doi.org/10.1371/journal.pone.0069897

[82] A. Lanitis, N. Tsapatsoulis, A. Maronidis: Review of ageing with respect to biometrics and diverse modalities. Age Factors in Biometric Processing (2013).

[83] A. Uhl, P. Wild, Experimental evidence of ageing in hand biometrics, IEEE, Biometrics Special Interest Group (BIOSIG), 2013 International Conference of the IEEE, 2013.

[84] S. Kirchgasser, A. Uhl, K. Castillo-Rosado, D. Estévez-Bresó, E. Rodríguez-Hernández and J. Hernández-Palancar, "Fingerprint Template Ageing Revisited - It’s the Quality, Stupid!," 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), Redondo Beach, CA, USA, 2018, pp. 1-9.

[85] F. Kelly, R. Saeidi, N. Harte, D. Leeuwen, Effect of long-term ageing on i-vector speaker verification, INTERSPEECH 2014, 14-18 September 2014, Singapore.

[86] Alonso-Fernandez, F., Fierrez, J., Gilperez, A., Ortega-Garcia, J. (Eds.): ‘Impact of time variability in off-line writer identification and verification’. Proc. Sixth Int. Symp. on Image and Signal Processing and Analysis, 2009, ISPA 2009, 16–18 September 2009

[87] F. Kelly, J.H.L. Hansen, The effect of short-term vocal aging on automatic speaker recognition performance.

[88] G. Amayeh, G. Bebis, and M. Nicolescu. Gender classification from hand shape. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008.

[89] T. Alafif; Z. Hailat; M. Aslan; X. Chen, On Classifying Facial Races with Partial Occlusions and Pose Variations, 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, 10.1109/ICMLA.2017.00-82.

[90] G. P. Mabuza-Hocquet; F. Nelwamondo; T. Marwala, Ethnicity Prediction and Classification from Iris Texture Patterns: A Survey on Recent Advances, 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 10.1109/CSCI.2016.0159.

[91] J. J. Howar, D. Etter, The Effect of Ethnicity, Gender, Eye Color and Wavelength on the Biometric Menagerie, 978-1-4799-1535-4/13/$31.00 ©2013, IEEE.

[92] J. Daugman; C. Downing, Searching for doppelgängers: assessing the universality of the IrisCode impostors distribution, IET Biometrics, Volume: 5, Issue: 2, 6, 2016, 10.1049/iet-bmt.2015.0071.

[93] S. Kumar; S. Singh; J. Kumar, A study on face recognition techniques with age and gender classification, 2017 International Conference on Computing, Communication and Automation (ICCCA), 10.1109/CCAA.2017.8229960.

[94] M. Trokielewicz, A. Czajka, P. Maciejewicz, Database of iris images acquired in the presence of ocular pathologies and assessment of iris recognition reliability for disease-affected eyes, Cybernetics (CYBCONF), 2015 IEEE 2nd International Conference on Strony, 495-500, IEEE.

[95] M. Trokielewicz, A. Czajka, P. Maciejewicz, Implications of ocular pathologies for iris recognition reliability, Image and Vision Computing, Volume 58, February 2017, Pages 158-167.

[96] F.L. Darley, A.E. Aronson, J.R. Brown Differential diagnostic patterns of dysarthria, J. Speech Lang. Hear. Res., 12 (2) (1969), p. 246

[97] H. Ackermann, W. Ziegler, Articulatory deficits in parkinsonian dysarthria: an acoustic analysis, J. Neurol. Neurosurg. Psychiatry, 54 (12) (1991), pp. 1093-1098

[98] J. Kegl, H. Cohen, H. Poizner Articulatory consequences of Parkinson's disease: perspectives from two modalities Brain Cogn., 40 (2) (1999), pp. 355-386

[99] Seyeddain O, Kraker H, Redlberger A, Dexl AK, Grabner G, Emesz M., Reliability of automatic biometric iris recognition after phacoemulsification or drug-induced pupil dilation, Eur J Ophthalmol. 2014 Jan-Feb; 24(1):58-62. doi: 10.5301/ejo.5000343. Epub 2013 Jul 17.

[100]K. Hollingsworth, K. Bowyer, and P. Flynn. Pupil dilation degrades iris biometric performance. Computer Vision and Image Understanding, 113(1):150–157, 2009.

[101] L. Dhir, N. E. Habib, D. M. Monro, and S. Rakshit. Effect of cataract surgery and pupil dilation on iris pattern recognition for personal authentication. Eye, 224(6):1006-1010, 2010.
[102] I. Tomeo-Reyes and V. Chandran. Effect of pupil dilation and constriction on the distribution of bit errors within the iris. In IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2014.
[103] I. Tomeo-Reyes, A. Ross, V. Chandran, "Investigating the Impact of Drug Induced Pupil Dilation on Automated Iris Recognition," Proc. of 8th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Buffalo, USA), September 2016.
[104] M. Drahansky, M. Dolezel, J. Urbanek, E. Brezinova,and T. h. Kim, Influence of Skin Diseases on Fingerprint Recognition, Journal of Biomedicine and Biotechnolog Volume 2012 (2012), Article ID 626148, 14 pages.
[105] M. Theofanos, B. Stanton, C. A. Wolfson, Usability & Biometrics, Ensuring Successful Biometric Systems, NIST, JUNE 11, 2008.
[106] M. Brockly, S. Elliott, R. Guest, and R. B. Gonzalo, "Human-Biometric Sensor Interaction," in Encyclopedia of Biometrics, S. Z. Li and A. K. Jain, Eds. Boston, MA: Springer US, 2014, pp. 1–9.
[107] "ISO/IEC 19795-1:2006. Information technology - Biometric performance testing and reporting - Part 1: Principles and framework", ISO/IEC, Geneva, 2006.
[108] O. Miguel-Hurtado, R. Blanco-Gonzalo, R. Guest, C. Lunerti, Interaction evaluation of a mobile voice authentication system, Security Technology (ICCST), 2016 IEEE International Carnahan Conference on 24-27 Oct. 2016, IEEE, 10.1109/CCST.2016.7815697.
[109] V. smejkal, J. Kodl, L. Sieger, The influence of stress on biometric signature stability, 2016, 978-1-5090-1072-1/16/IEEE.
[110] D. Gunetti, C. Picardi, and G. Ruffo, Keystroke Analysis of Different Languages: A Case Study, International Symposium on Intelligent Data Analysis, IDA 2005: Advances in Intelligent Data Analysis VI pp 133-144.
[111] Levi, S. V., & Schwartz, R. G. (2013). The development of language specific and language-independent talker processing. Journal of Speech, Language, and Hearing Research, 56(3), 913–920. https://doi.org/10.1044/1092-4388 (2012/12-0095)
[112] Z. Syed, S. Banerjee, Q. Cheng, B. Cukic, Effects of User Habituation in Keystroke Dynamics on Password Security Policy, High-Assurance Systems Engineering (HASE), 2011 IEEE 13th International Symposium on, 10.1109/HASE.2011.16.
[113] A. Dantcheva, C. Chen, A. Ross, Can Facial Cosmetics Affect the Matching Accuracy of Face Recognition Systems?, Proc. of 5th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2012.

[114] N. Kohli, D. Yadav, M. Vatsa and R. Singh, "Revisiting iris recognition with color cosmetic contact lenses," 2013 International Conference on Biometrics (ICB), Madrid, 2013, pp. 1-7.
[115] A.S. Osman Ali, V. Sagayan, A. Malik, A. Aziz, Proposed face recognition system after plastic surgery, IET Computer Vision, E-First on 25th February 2016, doi: 10.1049/iet-cvi.2014.0263

[116] P. Bours; A. Evensen, The Shakespeare experiment: Preliminary results for the recognition of a person based on the sound of walking, 2017 International Carnahan Conference on Security Technology 10.1109/CCST.2017.8167839.
[117] Poorjam A.H., Hesaraki S., Safavi S., van Hamme H., Bahari M.H. (2017) Automatic Smoker Detection from Telephone Speech Signals. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol. 10458. Springer, Cham.
[118] Satori, H., Zealouk, O., Satori, K. et al. Voice comparison between smokers and non-smokers using HMM speech recognition system. Int J Speech Technol 20, 771–777 (2017). https://doi.org/10.1007/s10772-017-9442-0.
[119] S. S. Arora, M. Vatsa, R. Singh, A. Jain, Iris recognition under alcohol influence: A preliminary study, Biometrics (ICB), 2012 5th, 2012, IEEE, DOI: 10.1109/ICB.2012.6199829
[120] J. Shin, T. Kuyama, Detection of alcohol intoxication via online handwritten signature verification, Pattern Recognition Letters, Volume 35, 1 January 2014, Pages 101-104.
[121] S. Wan, J. Aggarwal, A scalable metric learning-based voting method for expression recognition, in: Automatic Face and Gesture Recognition (FG), 2013, 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
[122] L. Ivanovsky, V. Khryashchev, A. Lebedev, I. Kosterin, Facial Expression Recognition Algorithm Based on Deep Convolution Neural Network, Proceeding of the 21st Conference Fruct Association,
[123] T. Li, J. Zhou, N. Tuya, C. Du, Z. Chen, S. Liu, Recognize Facial Expression Using Active Appearance Model And Neural Network, 978-1-5386-2209-4/17, 2017 IEEE, DOI 10.1109/CyberC.2017.32.
[124] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, vol. 147, pp. 195-197, 1981.
[125] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs.Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, pp. 711-720, 1997.

[126] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba. “Coding facial expressions with Gabor wavelets”, Proceedings of the Third IEEE Conference on Face and Gesture Recognition, pp. 200-205, 1998.

[127] X. Liu, T.Chen and B.V.K. Vijaya Kumar. Face Authentication for Multiple Subjects Using Eigenflow Pattern Recognition, Volume 36, pp. 313-328, 2003.

[128] Mansour, A. and Lachiri, Z., "Speaker recognition in emotional context," International Journal of Computer Science, Communication and Information Technology (CSCIT), pp. 1-4, Oct. 2015.

[129] Nandwana, M. K. and Hansen, J. H., "Analysis and identification of human scream: implications for speaker recognition," in Proc. Interspeech, pp. 2253-2257, Sep. 2014.

[130] A. Revathy, P. Shanmugapriya and V. Mohan, "Performance comparison of speaker and emotion recognition," 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), Chennai, 2015, pp. 1-6.

[131] S. Parthasarathy, C. Zhang, J. H. L. Hansen and C. Busso, "A study of speaker verification performance with expressive speech," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp.
5540-5544.

[132] Krothapalli, S.R., Yadav, J., Sarkar, S. et al. Int J Speech Technol (2012) 15: 335. https://doi.org/10.1007/s10772-012-9148-2

[133] M. V. Ghiurcau, C. Rusu and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011, pp. 4944-4947

[134] C.H. Chang, P.W. Johnson, J.N. Katz, E.A. Eisen, and J.T. Dennerlein, “Typing keystroke duration changed after submaximal isometric finger exercises,” European Journal of Applied Physiology, vol. 105, (no. 1), pp. 93-101, Jan 2009.

[135] S. Komandur, P.W. Johnson, and R.L. Storch, “Relation between mouse button click duration and muscle contraction time,” in Proc. 30th Annual International IEEE EMBS Conference, 2008, pp.

[136] H. Al-Libawy, A. Al-Ataby, W. Al-Nuaimy, M. A. Al-Taee, Q. Al-Jubouri, Fatigue Detection Method Based on Smartphone Text Entry Performance Metrics, 978-1-5090-5487-9/17 $31.00 © 2017 IEEE, DOI 10.1109/DeSE.2016.940, 9th International Conference on Developments in Systems Engineering.

[137] M. A. Haque; K. Nasrollahi; T. B. Moeslund, Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain?, 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), 10.1109/ISBA.2017.7947690.

[138] A. Krupp, C. Rathgeb and C.Busch, Social Acceptance of Biometric Technologies in Germany: A Survey, 2013, IEEE, 978-3-88579-606-0.

[139] S. M. Furnell, P. S. Dowland, Illingworth H. M., and P. L. Reynolds. Authentication and Supervision: A Survey of User Attitudes. Computers & Security, 19(3):529–539, 2000

[140] C. Perakslis and R. Wolk. Social acceptance of RFID as a biometric security method. IEEE Technology and Society Magazine, 25(3):34 –42, 2006

[141] S. Y. Mok and Ajay Kumar. Addressing biometrics security and privacy related challenges in China. In Proc. Int'l Conf. of the Biometrics Special Interest Group (BIOSIG'12), pages 1–8, 2012.

[142] A. S. Rashed; A. Henrique, Biometrics acceptance in Arab culture: An exploratory study, 2013, 10.1109/ICCAT.2013.6521970, IEEE.

[143] F. Al-Harby, R. Qahwaji, M. Kamala, (2001), The feasibility of biometrics authentication in e-commerce: user acceptance, IADIS International Conference, Freiburg, Germany.

[144] T. Alhussain, S. Drew, Towards User Acceptance of Biometric Technology in EGovernment: A Survey Study in the Kingdom of Saudi Arabia. I3E 2009: 26-38.

[145] A. Liljander, Attitudes towards biometric authentication technologies between cultures: acceptance in Finland And Brazil, 2019, PHD thesis, https://jyx.jyu.fi/handle/123456789/66405.

[146] Hollingsworth K, Bowyer KW, Lagree S, Fenker SP, Flynn PJ. Genetically identical irides have texture similarity that is not detected by iris biometrics. Comput Vis Image Und, 2011, 115 (11):1493-1502.

[147] Daugman JG. High confidence visual recognition of persons by a test of statistical independence. Ieee T Pattern Anal, 1993, 15(11):1148–1161.

[148] Rasoulinejad SA, Zarghami A, Hosseini SR, Rajaee N, Rasoulinejad SE, Mikaniki E. Prevalence of age-related macular degeneration among the elderly. Caspian J Intern Med 2015; 6(3): 141–147.

[149] Rasoulinejad SA, Hajian-Tilaki K, Mehdipour E. Associated factors of diabetic retinopathy in patients that referred to teaching hospitals in Babol. Caspian J Intern Med 2015; 6(4): 224–228.

[150] Noor-ul-huda M, Tehsin S, Ahmed S, Niazi FAK, Murtaza Z. Retinal images benchmark for the detection of diabetic retinopathy and clinically significant macular edema (CSME). Biomed Eng-Biomed Tech 2019 27;64(3):297-307.

[151] Samant P, Agarwal R. Machine learning techniques for medical diagnosis of diabetes using iris image. Comput Meth Prog Bio 2018, 157: 121-128.

[152] Heydari M, Teimouri M, Heshmati Z. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 2016, 36(2):167-173.

[153] Chaskar UM, Sutaone MS. On a Methodology for Detecting Diabetic Presence from Iris Image Analysis. Int Conf Power Signals Control Comput 2012: 1–6.

[154] Aslam TM, Tan SZ, Dhillon B. Iris recognition in the presence of ocular disease. J R Soc Interface 2009, 6(34): 489–493.

[155] Borgen H, Bours P, Wolthusen SD. Simulating the Influences of Aging and Ocular Disease on Biometric Recognition Performance. International Conference on Biometrics 2009, LNCS 5558, 857–867.

[156] Nigam I, Vatsa M, Singh R. Ophthalmic Disorder Menagerie and Iris Recognition. In: chapter 22, Handbook of Iris Recognition, Springer, 2016:519-539.

[157] Nigam, I., Keshari, R., Vatsa, M. et al. Phacoemulsification Cataract Surgery Affects the Discriminative Capacity of Iris Pattern Recognition. Sci Rep 9, 11139 (2019). https://doi.org/10.1038/s41598-019-47222-4.

[158] Available from: http://www.iritech.com/products/hardware/irishield%E2%84%A2-series

[159] Uhl A, Wild P. Weighted Adaptive Hough and Ellipsopolar Transforms for Real-time Iris Segmentation. International Conference on Biometrics, 2012: 283 - 290.

[160] Zhang D, Monro DM, Rakshit S. DCT-Based Iris Recognition. Ieee T Pattern Anal 2007, 29: 586-595.

[161] Masek L, Kovesi P. MATLAB Source Code for a Biometric Identification System Based on Iris Patterns. The School of Computer Science and Software Engineering, The University of Western Australia. 2003.

[162] Rathgeb C, Uhl A. Secure Iris Recognition based on Local Intensity Variations. Proceedings of the International Conference on Image Analysis and Recognition (ICIAR'10), Springer, LNCS 6112, 2010: 266-275.

[163] Rathgeb C, Uhl A, Wild P, Hofbauer H. Design Decisions for an Iris Recognition SDK. Bowyer K, Burge MJ, editors, Handbook of iris recognition, second edition, Advances in Computer Vision and Pattern Recognition, Springer, 2016.

[164] F.Mauvais-Jarvis, Gender differences in glucose homeostasis and diabetes, Physiology & Behavior, Volume 187, 1 April 2018, Pages 20-23.

[165] A. Kautzky-Willer, J. Harreiter, Sex and gender differences in therapy of type 2 diabetes, Diabetes Research and Clinical Practice, Volume 131, September 2017, Pages 230-241.

[166] I. Campesi, F. Franconi, G. Seghieri, M. Meloni, Sex-gender-related therapeutic approaches for cardiovascular complications associated with diabetes, Pharmacological Research, Volume 119, May 2017, Pages 195-207.

[167] F. C. Schweppe, On the Bhattacharyya Distance and the Divergence between Gaussian Processes, Information and Control, 1967, 11, 373-395.

[168] Y. Wei, H. G. Hosseini, A. Cameron, M. J. Harrison and A. Al-Jumaily. Voice analysis for detection of hoarseness due to a local anesthetic procedure. 3rd International Conference on Signal Processing and Communication Systems, Omaha, NE, 1-7, 2009.

[169] W. Xie, A. Nagrani, J. S. Chung, A. Zisserman, Utterance-level Aggregation For Speaker Recognition In The Wild, ICASSP, 2019.

[170] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, INTERSPEECH, 2018.

[171] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: A Large-scale Speaker Identification Dataset, INTERSPEECH, 2017.

[172] H. T. Nguyen, "Contributions to Facial Feature Extraction for Face Recognition," Université Grenoble Alpes, thesis, 2014.

[173] J. Ou, "Classification Algorithms Research on Facial Expression Recognition," Physics Procedia, vol. 25, pp. 1241-1244, 2012.

[174] S. Banerjee, S. Das, MakeUpMirror: mirroring make-ups and verifying faces post make-up, IET Biometrics, 2018, Vol. 7 Iss. 6, pp. 598-605 doi: 10.1049/iet-bmt.2017.0265.

[175] Chen, C., Dantcheva, A., Ross, A.: 'Automatic facial makeup detection with application in face recognition'. IEEE Int. Conf. on Biometrics (ICB), Madrid, Spain, 2013, pp. 1–8

[176] Chen, C., Ross, A.: 'Local gradient Gabor pattern (LGGP) with applications in face recognition, cross-spectral matching and soft biometrics'. Proc. SPIE Biometric and Surveillance Technology for Human and Activity Identification X, Washington DC, USA, 2013, pp. 87120R–87120R.

[177] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 24(8), 1377—1388. DOI: 10.1080/02699930903485076

[178] Available: http://pics.stir.ac.uk/2Dfacesets.htm.

[179] E. K. Davis, "Dlib-ml: A Machine Learning Toolkit" J. Mach. Learn. Res., vol. 10, pp. 1755-1758, 2009.

[180] E. K. Davis. High quality face recognition with deep metric learning. 2017. 33.

[181] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[182] Available: https://github.com/gv22ga/dlib-face-recognition-android

[183] Available: http://www.neurotechnology.com/verilook.html

[184] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, British Machine Vision Conference, 2015.

[185] T. Y. Wang and A. Kumar, “Recognizing human faces under disguise and makeup,” in Proceedings of the 2nd IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2016, Japan, March 2016.

[186] https://www.biometricupdate.com/201908/biometrics-researchers-show-iris-recognition-accuracy-reduced-by-type-ii-diabetes

[187] https://sciencediscoveries.degruyter.com/iris-recognition-under-influence-diabetes/

[188] https://findbiometrics.com/study-shows-iris-recognition-less-effective-people-diabetes-082005/